ИКОНОМИЧЕСКИ ИЗСЛЕДВАНИЯ
ECONOMIC STUDIES
ГОДИНА XXVI, 2017, 2

Маргарита Ламбова[1]

ЗА „РЕПРЕЗЕНТАТИВНИТЕ" ИЗВАДКИ И ТЯХНАТА „ОБОСНОВКА"

От много десетилетия в българската специализирана литература се е наложило понятието „репрезентативна извадка", а през последните години все по-често при научни изследвания се използна терминът „обосновка" на „репрезентативната" извадка. Изследването е осъществено в две направления. Първото е свързано с аргументиране на тезата, че понятията „репрезентативна извадка" и „случайна извадка" не са взаимозаменими, тъй като повечето съществуващи концепции за репрезентативност са логически несъвместими със същността на случайния подбор и с възможните резултати от случаен експеримент, а второто – с аргументация против придобилото популярност твърдение, според което всяка „репрезентативна" извадка задължително трябва да бъде с изчислен по определена формула обем, за да се смята за „обоснована".
JEL: C12; C13

Въведение

В българската специализирана литература, свързана със статистическата методология, по-конкретно с теорията на статистическите заключения, от десетилетия са се наложили понятията „репрезентативна извадка" и „репрезентативно изучаване", като с тях се визират предимно случайните извадки и статистическото оценяване. Същевременно все по-често в науката и практиката се прилагат извадкови изследвания, базирани на статистическите заключения, т.е. осъществявани с помощта т.нар. репрезентативни извадки, понятие, което се използва като синоним на случайни извадки. Много често към нас статистиците се обръщат със следния въпрос: Как да обоснова извадката, която използвам, по такъв начин, че тя да предизвика доверието на рецензенти, научно жури, специалисти и т.н., т.е. по коя формула да изчисля обема на извадката? Посоченото дотук съдържа няколко проблемни момента, свързани с понятийния апарат, както и с влагания в понятията смисъл, на които не се обръща достатъчно внимание не само при приложението, но и при преподаването на статистически извадкови способи и методи.

[1] *Маргарита Ламбова е доц. д-р в Икономически университет – Варна, катедра „Статистика", тел: 0882-164714, e-mail: lambowa@yahoo.de.*

Обект на изследване тук е случайната извадка. **Предмет** на изследване са понятията „репрезентативност" и „обосновка", използвани в науката и практиката във връзка със случайната извадка. **Целта** се състои в аргументиране на тезата, че понятията „репрезентативна извадка" и „случайна извадка" не са взаимозаменими, тъй като се основават на различна логика и не са тъждествени по съдържание, а също и да се представи и аргументация против придобилото популярност твърдение, че всяка „репрезентативна" извадка задължително трябва да бъде с изчислена по определена формула обем, за да се смята за „обоснована".

Във връзка с поставената цел се открояват две основни задачи:

1. Да се направи теоретична характеристика на понятието „репрезентативност" и на тази основа обосноваване на логическото различие между т. нар. репрезентативна и случайната извадка.
2. Да се разкрият основните теоретични проблеми, свързани с определянето на обема на случайната извадка при използване на различни статистически извадкови способи и методи, както и да се отговори на въпроса какво може да гарантира обем на извадката, изчислен по формула, изведена от интервал на доверителност за даден параметър на разпределението по определен признак и използван за извадково изследване, включващо множество променливи, при което до заключения се стига освен чрез статистическо оценяване и чрез проверка на статистически хипотези.

1. Репрезентативност и случайна извадка

1.1. Теоретична характеристика на понятието „репрезентативност"

Широко разпространено е твърдението, че случайната извадка задължително е репрезентативна (представителна), както и обратно – че за да бъде представителна, извадката трябва да бъде излъчена чрез случаен подбор.

Според някои автори понятието „репрезентативност", използвано във връзка с извадки, не е еднозначно дефинирано в специализираната статистическа литература (Von der Lippe, 2011), въпреки че е въведено в употреба още през 1895 г. Най-общо то се приема като свойство на емпирични изследвания, чиито резултати могат да се използват за съждения относно генералната съвкупност, като тези изследвания се базират както на случайни, така и на *преднамерени* извадки. Често използваният термин „репрезентативна извадка" (Representative sampling) не е специализирано научно понятие заради разтегливостта на понятието, за което няма точна и ясна дефиниция и липсва обективно мерило за степента на представителност. Във водещите по развитие на статистическата наука държави той се използва основно в практиката, но не и в статистическата наука, във връзка с различни по вид и логика извадкови изследвания.

Първите, които обръщат по-сериозно внимание на вътрешната противоречивост на понятията „статистическа репрезентативност" и „репрезентативна извадка", са

Kruskall и Mosteller (1980). Те изследват развитието на смисъла, който се влага в тях след въвеждането им в употреба, и откриват няколко различаващи се напълно и отчасти несъвместими техни значения. Въпреки че от издаването на публикацията са изминали 35 години, тя още е актуална, тъй като формулираните в нея представи за статистическата репрезентативност все още циркулират в научното и образователното пространство на някои държави и по този начин оказват въздействие върху научните и практическите изследвания. Двамата автори чрез карикатури извеждат следните основни значения на понятието „репрезентативност":

1. отсъствие на селективни влияния при подбора на извадката: правдивост на отражението (Absence of selective forces: Justice balancing the scales);
2. извадката като миниатюра (умалено копие) на съвкупността (Miniature of the population: Model train set);
3. репрезентативност като синоним на типичност (Typical case);
4. обхващане на съвкупността на принципа „Ноев ковчег" при подбора на извадката (Coverage of the population: Noah's Ark);
5. репрезентативна е тази извадка, която институционалната или университетската статистика дефинират като такава на базата на своите собствени правила (Some specific sampling method: The Sampling Departament in action).

Дори без обстоен анализ на различните значения на понятието „репрезентативност", респ. „репрезентативна извадка", се забелязват логическите противоречия. Представата за репрезентативност, според която извадката трябва да бъде правдиво отражение на разпределението в съвкупността, може да влезе в противоречие с принципа „Ноев ковчег", според който в извадката трябва да е застъпено цялото многообразие, което е налице в съвкупността. Същевременно правдивото отражение и принцпът на Ноевия ковчег се конфронтират с представата за типичност, тъй като правдивостта изисква включване и на нетипичното, а принципът „Ноев ковчег" цели максимално обхващане на вариацията, разнообразието в съвкупността.

Според Schnell, Hiller и Esser „многообразието и размитостта на различните дефиниции са причината голата констатация, според която дадена извадка е „репрезентативна" или дадено допитване е „репрезентативно", да не изразява нищо, камо ли да става въпрос за точно дефиниран критерий за надеждност" (вж. Schnell, Hiller and Esser, 1995).

На базата на формулираните от Kruskall и Mosteller представи за „репрезентативност" Von der Lippe (2011) прави задълбочен анализ на понятието, като го противопоставя на стохастичната грешка, която според него е единствената годна концепция за критерий за качеството на дадена извадка. Той посочва, че понятието „репрезентативност", което в неспециализираната езикова практика се употребява във връзка с извадки като синоним на надеждност и сериозност, почти не се среща в учебници по статистика. (Ясно е, че авторът визира най-вече немскоезична и англоезична учебна литература.) Основната и особено разбираема причина според него лесно се забелязва – не съществува измерител за степента на

репрезентативност, чрез който представителността на различни извадки би могла да бъде сравнена.

Von der Lippe извежда няколко концепции за репрезентативност, които отчасти кореспондират с формулираните от Kruskall и Mosteller значения на понятието:

- структурна концепция (RS);
- концепция на миниатюрата (RM);
- концепция на заместника (представителя) (RV);
- концепция на Ноевия ковчег (RA);
- неселективна концепция (RN);
- концепция на стохастичната грешка (SF).

Най-популярна е структурната концепция, според която дадена извадка е репрезентативна, когато структурата ѝ се доближава до тази на генералната съвкупност. Проблемът тук е, че не съществува мярка за степента на припокриване на двете структури, която да позволи разграничаването на „репрезентативните" и „нерепрезентативните" извадки. Ако например в съвкупността е налице равномерно разпределение по пол, т.е. 50% мъже и 50% жени, тогава извадка, в която разпределението е идентично, ще бъде по-представителна, отколкото такава, където е например 40 на 60%, която от своя страна е по-репрезентативна от извадка със съотношение 30:70. Дотук всичко звучи напълно логично, но ако бъде включен и обемът на извадката, тогава става ясно, че концепцията не позволява сравнение на степента на представителност на извадки с различен обем. Авторът дава следния пример: Според структурната концепция извадка, в която попаднат 3 жени и 3 мъже, би трябвало да бъде толкова добра, колкото такава с 30 жени и 30 мъже, но значително по-добра от извадка, в която има 305 мъже и 295 жени. Логичен ли е този извод? При дихотомен признак структурата е проста и все пак позволява еднозначно интерпретиране на различието, но само когато става въпрос за извадки с еднакъв обем. Проблемът се задълбочава при признаци с повече от две значения, при които сравнението на „степента" на представителност се усложнява, както се вижда от приведения елементарен пример (Von der Lippe, 2011):

Таблица 1

Структура по признака „семейно положение" в генерална съвкупност и в две излъчени от нея извадки (%)

Семейно положение	Относителен дял в:		
	Генералната съвкупност	Извадка 1	Извадка 2
Несемеен	36	34	38
Семеен	52	54	51
Други	12	12	11

Коя от двете извадки е по-добра, т.е. по-репрезентативна? Ясно е, че структурата зависи от обема на извадката, като пропорциите, които са налице в съвкупността, могат да се отразят абсолютно идентично само когато броят на единиците позволява това. В посочения пример минималният обем, който позволява точното спазване на пропорциите, е n=25 (9 несемейни, 13 семейни и 3 други). Следващите са n=75, n=100 и т.н. Извадка с междинен обем обективно няма да бъде в състояние да възпроизведе точно структурата на съвкупността и според тази концепция ще се смята за по-малко представителна. Това може да бъде онагледено с помощта на примера в табл. 2, в който разпределението в извадките е съобразено с пропорциите в съвкупността.

Таблица 2

Възможно приближение на структурата на извадки с различен обем до структурата на генералната съвкупност по признака „семейно положение"

Семейно положение	Генерална съвкупност (%)	Извадка 1 n=24		Извадка 2 n=25		Извадка 3 n=26	
		Брой	%	Брой	%	Брой	%
Несемеен	36	9	37.5	9	36	9	34.61
Семеен	52	12	50.0	13	52	14	53.85
Други	12	3	12.5	3	12	3	11.54
Общо	100	24	100.0	25	100	26	100.00

Когато е налице двумерно или многомерно разпределение, сравнението на представителността на различни извадки според стуктурната концепция се усложнява още повече, дори в много случаи става невъзможно, особено когато структурата в съответните едномерни разпределения съвпада.

Концепцията на миниатюрата е още по-неопределена и неясна от структурната концепция. Според нея извадката би трябвало да бъде правдиво умалено копие на съвкупността. Ако правдивостта на отражението се свързва единствено с идентичност на структурата по изследваните признаци, тогава не би имало съществена разлика със структурната концепция, но концепцията на миниатюрата изисква освен това наличието или отсъствието на определени единици в извадката, за да бъде умаленото копие по-добро. Основен проблем е, че няма точни и ясни критерии, които да дават възможност да се прецени дали дадена извадка може, или не може да бъде смятана за миниатюра на съвкупността. Докато при структурната концепция относителните честоти на значенията на признаците могат да бъдат използвани за измерване на различието на структурата на извадката и съвкупността, при концепцията на миниатюрата такъв количествен измерител не съществува. Не става ясно също кои единици задължително трябва да попаднат в извадката, за да бъде тя правдиво умалено копие, и какъв трябва да бъде минимално необходимият обем на извадката, който да позволява миниатюризация на съвкупността. При един признак с малко на брой значения една сравнително малка извадка е в състояние да пресъздаде правдиво съвкупността. Колкото повече „нюанси" на съвкупността трябва да бъдат отразени в извадката, за да бъде тя „репрезентативна" според тази концепция, т.е. колкото повече признаци подлежат на изследване, толкова по-голяма

трябва да бъде миниатюрата, като при много голям брой изследвани признаци тя дотолкова трябва да се доближи до „оригинала", че става излишна. Не е за подценяване и начинът на „създаване" на такава миниатюра. Според Kruskall и Mosteller (1979) идеята за извадка като отражение или миниатюра на съвкупността рядко е подходяща, защото миниатюрата обикновено се конструира преднамерено, а не чрез процес на вероятностен подбор.

При концепцията на заместника (представителя) се изхожда от буквалния превод на понятието „репрезентативен", като се приема, че излъчените единици, формиращи извадката, могат да представляват (заместват) неизлъчените единици на съвкупността. Това означава, че те би трябвало да бъдат идентични или да притежават висока степен на сходство с неизлъчените. Как обаче да се прецени дали единиците в извадката успешно могат да заместят неизлъчените, след като не се познават последните? Дори да се разполага с информация за неизлъчените единици, възниква въпросът какъв измерител на подобието да бъде използван, за да може да покаже дали сходството е достатъчно и единиците от извадката да бъдат наречени представители на неизлъчените единици на съвкупността? Концепцията на заместника се гради върху разбирането, че „репрезентативност" се приема като синоним на „типичност". Дадена единица успешно може да замести всички останали единствено ако значенията на характеризиращите я признаци съвпадат с центъра на съответното разпределение в съвкупността, т.е. за да бъде представителна, единицата трябва да възпроизвежда характерната за съвкупността средна величина във всяко отношение. Според Von der Lippe (2011) това не е възможно поради следните причини:

1. Разпределението по значенията на изследваните признаци в съвкупността, оттам и съответните средни, не са напълно известни, като точно това е основанието за провеждането на извадково изследване.
2. Ако реално съществуваше една единица, която във всяко отношение е средностатистическа, то тогава би било достатъчно излъчването на извадка с обем n=1, включваща точно тази единица, когато трябва да бъде направено заключение относно центъра на разпределение.
3. Ако на базата на извадката трябва да бъде направено заключение относно разсейването на значенията на изследваните признаци, тогава извадка, която според концепцията на заместника съдържа само „типични" единици, би била напълно неподходяща, тъй като тя винаги ще бъде с дисперсия, равна на 0, независимо от разсейването в съвкупността.

Противоречията, които съдържа концепцията на заместника, са очевидни. При непознаване на разпределението в съвкупността няма как да се прецени коя единица може да се смята за типична, но дори да има подобна информация, извадка от „типични" единици не може да допринесе за дефинирането на неизвестното разпределение в съвкупността, защото самата тя е съставена от сходни, типични единици, чието разпределение се характеризира с разсейване, клонящо към 0.

Концепцията на Ноевия ковчег е точната противоположност на тази на заместника, като при нея под „репрезентативност" се разбира коректното възпроизвеждане на многообразието, срещащо се в съвкупността. Von der Lippe (2011) се позовава на Kruskall и Mosteller (1979) , според които при това разбиране за репрезентативност извадката трябва да обхваща поне една единица от всеки клас, така, както в Ноевия ковчег от всеки животински вид е бил наличен поне един екземпляр. Както при Ноевия ковчег, големината на класовете отстъпва по значение на съхраняването на видовото разнообразие. Основният принцип тук е селективността, типична най-вече за преднамерения подбор. Тук за разлика от концепцията на заместника репрезентативността се свързва с относително голям обем на извадката, тъй като многообразието на съвкупността не може да се пресъздаде от малък брой единици.

Неселективната концепция за репрезентативност се базира на отсъствието на селективни влияния при подбора на извадката (Absence of selective forces) (Kruskall and Mosteller, 1979). За разлика от предходните концепции, при които репрезентативността се обвързва с резултата от подбора, неселективната концепция акцентира върху механизма на подбор, без да се интересува от резултата. При нея „пътят е целта" (Der Weg ist das Ziel). „Репрезентативност е качеството на процеса на подбор, а не качеството на излъчената извадка." (пак там). За репрезентативен се смята подборът, при който се изключват факторите, които биха могли да подпомогнат или да възпрепятстват попадането на определени единици в извадката. Като проблем на концепцията авторът посочва това, че никога не можем да сме сигурни, че наистина е изключена всякаква форма на селективност.

Концепцията на стохастичната грешка се базира на случайния подбор, т.е. подобно на неселективната концепция, обвързва репрезентативността с механизма на подбор, а не с качеството на излъчената извадка. Това означава, че не извадката се приема за репрезентативна, а само нейният подбор, който не дава никакви гаранции за подобието на структурата ѝ с тази в съвкупността. Случаен подбор е налице, когато (Bourier, 2002):

1. Всяка единица на съвкупността има шанс, различен от 0 за попадане в извадката.
2. Вероятността за попадане в извадката е изчислима за всяка единица.
3. Не съществува зависимост между вероятността за попадане в извадката на отделните единици и значенията на изучавания статистически признак.

Само при извадки, излъчени чрез случаен подбор, може да бъде изчислена абсолютната и относителната стохастична грешка ($\sigma_{\bar{x}}$ и $\frac{\sigma_{\bar{x}}}{\mu}$). Като измерител на степента на репрезентативност Von der Lippe предлага относителната стохастична грешка $\frac{\sigma_{\bar{x}}}{\mu}$. По-малка величина на грешката се свързва с по-голяма степен на репрезентативност на подбора.

1.2. Логическо различие между „репрезентативна" и случайна извадка

Според Von der Lippe и Kladroba „репрезентативната" извадка се различава съществено от случайната, т.е. между двете не може да се постави знак на равенство. Двамата автори обобщават интуитивната представа за репрезентативност по следния начин: „..Формирането на извадката трябва да бъде осъществено така, че въз основа на резултатите от извадковото изследване да са възможни максимално точни и сигурни заключения относно свойствата на съвкупността, т.е. извадката трябва с голямо приближение да възпроизвежда разпределението по изучаваните признаци в съвкупността. Това е възможно само когато тя е правдиво нейно отражение, т.е. умалено копие на съвкупността... Обобщено, може да се твърди, че според общоприетата терминология представителност на извадка е налице, когато структурата ѝ по определени признаци е подобна на тази в генералната съвкупност. Според редица автори от това следва, че само въз основа на подобна извадка могат да се правят заключения относно съвкупността." (Von der Lippe and Kladroba, 2002). Това обобщение съдържа елементи от представените концепции за репрезентативност и при внимателен прочит могат да се открият взаимно противоречащи си твърдения, които няма как да бъдат съвместени със случайната извадка.

На първо място, при характеризиране на репрезентативността обикновено се поставя структурата на извадката (**структурната концепция за репрезентативност**), която според интуитивните представи за репрезентативност би трябвало да е близка до тази на съвкупността. Възможно ли е да се очаква подобна „репрезентативност" при случаен подбор, след като това би означавало, че всички различаващи се една от друга случайни извадки с обем *n*, които е възможно да бъдат излъчени от дадена съвкупност, трябва да възпроизвеждат нейната структура и особености, т.е те би трябвало да са почти идентични? Подобно изискване противоречи на логиката на случайния експеримент, следователно и на случайния подбор. Ако например бъде излъчена случайна извадка от текущото производство, включваща 100 изделия, и се знае, че делът на некачественост възлиза на 10%, то тогава тя ще бъде репрезентативна, ако възпроизвежда структурата на съвкупността, т.е в нея са налице около 10 некачествени единици. При биномно разпределение с параметри $n=100$ и $\theta=0{,}1$ вероятността за излъчването на извадка, която повтаря структурата на съвкупността, възлиза едва на 13.2%, а вероятността в извадката процентът на некачественост да бъде между 9 и 11 – на 38.2%. Следователно преобладаващата част от извадките, които е възможно да се формират при подобен случаен експеримент, няма да са или ще бъдат в много по-малка степен представителни по отношение на разпределението по изучавания признак в съвкупността. това води до извода, че голяма част от случайните извадки не са представителни, тъй като не възпроизвеждат достатъчно добре структурата на съвкупността, от която са излъчени. Колкото е по-голяма стохастичната грешка при даден обем на извадката, т.е. стандартното отклонение на статистическата оценка като случайна величина, толкова по-висок ще бъде делът на възможните извадки с дадения обем, чиято структура се отклонява съществено от структурата на съвкупността. Ако в посочения пример при постоянен обем на извадката ($n=100$) се променя вероятността за некачественост θ, като се

предполага биномно разпределение на случайната величина „брой некачествени изделия в извадка с обем 100", вероятността за излъчване на извадка, повтаряща структурата в съвкупността се променя по посочения в табл. 3 начин.

Таблица 3

Вероятност за структура, идентична с тази в съвкупността и стохастична грешка при $n = 100$

Вероятност за некачественост θ	Брой некачествени изделия в извадката x	Вероятност за структура, идентична с тази на съвкупността $W(X = x) = f_B(x/100\ ;\theta)$	Стохастична грешка на оценката $\sigma_P = \sqrt{\frac{\theta(1-\theta)}{n}}$
0,05	5	0,1800	0,0212
0,10	10	0,1319	0,0300
0,20	20	0,0993	0,0400
0,30	30	0,0868	0,0458
0,40	40	0,0812	0,0490
0,50	50	0,0796	0,0500

Вероятността за структура на извадката, идентична с тази на съвкупността, зависи до голяма степен и от обема на извадката. Колкото по-голям е той при равни други условия, толкова по-малко вероятно ще бъде чрез случаен подбор да бъде излъчена „репрезентативна" извадка, чиято структура е напълно идентична с тази на съвкупността (вж. табл. 4). Това обаче не е недостатък на случайния подбор, а основно предимство, тъй като при увеличаване на обема на извадката намалява стандартното отклонение на статистическата оценка (в случая – на извадковия относителен дял *P*), т.е. стохастичната грешка на оценката, която е основен критерий за надеждност на заключението.

При случаен подбор са възможни както репрезентативни, така и нерепрезентативни от гледна точка на структурата извадки, като делът на тези, чиято структура се отклонява съществено от структурата на съвкупността, е твърде голям, за да може да се твърди, че по правило случайните извадки са представителни. За горепосочения пример при $n = 100$ и $\theta = 0{,}1$ случайната величина „брой некачествени изделия в извадка с обем 100" може да приеме значения между 0 и 100 ($x = 0, 1, 2, ..., 100$).

Макар и малко вероятни, са възможни екстремни извадки, при които значението на случайната величина попада в краищата на разпределението. Дори при липса на точен критерий за характеризиране на степента на подобие на структурата такива извадки трябва да бъдат квалифицирани като абсолютно непредставителни. Самата същност на случайния подбор предполага всякакви възможни комбинации от значения, наблюдавани при излъчените единици, и няма никаква гаранция за формиране на репрезентативна по отношение на структурата извадка, след като тя е излъчена чрез случаен подбор.

Таблица 4

Вероятност за структура, идентична с тази в съвкупността, и стохастична грешка при $\theta = 0{,}1$

Обем на извадката n	Брой некачествени изделия в извадката x	Вероятност за структура, идентична с тази на съвкупността $W(X = x) = f_B(x/n; 0{,}1)$	Стохастична грешка на оценката $\sigma_P = \sqrt{\frac{\theta(1-\theta)}{n}}$
20	2	0,2852	0,0670
40	4	0,2059	0,0474
60	6	0,1693	0,0387
80	8	0,1471	0,0335
100	10	0,1319	0,0300
200	20	0,0936	0,0212

Обикновено освен подобието на структурата като изискване за репрезентативност се посочва, че извадката трябва да бъде умалено копие на съвкупността, т.е. нейна миниатюра (**концепция на миниатюрата**). Може ли случайният подбор да гарантира формирането на извадка, която да възпроизвежда в умален вид всички нюанси на съвкупността? Отново се сблъскваме с противоречие между същността на случайния експеримент и изискването за миниатюризация на съвкупността. Както вече беше посочено, при случайния подбор е възможно формирането на извадки с всякакви комбинации от значения на изучаваните признаци, срещащи се в съвкупността. Някои от възможните извадки ще могат да се приемат за умалени копия, но друга част, която съвсем не е незначителна, няма да отразява правдиво всички нюанси на съвкупността, т.е. няма да бъде представителна за нея. Идеята за извадка като умалено копие на съвкупността не е съвместима със случайния подбор, тъй като „миниатюрата обикновено се конструира преднамерено" (Kruskal and Mosteller, 1979), т.е. става въпрос за съзнателен, а не за вероятностен подбор.

Представата за репрезентативност, изградена на базата на не толкова популярната **концепция на заместника**, е логически абсолютно несъвместима със случайния подбор. Излъчването само на „типични" единици би могло да се осъществи единствено селективно, т.е. чрез преднамерен подбор, и то само в случаите, когато е налице предварителна информация за центъра на разпределение в съвкупността.

Репрезентативността, разбирана от **концепцията „Ноев ковчег"**, като коректно отражение на многообразието, срещащо се в съвкупността, в излъчената извадка, т.е. включването на поне една единица от всеки клас независимо от големината на наличните класове, получени при групировката на единиците на съвкупността по значенията на изучаваните признаци, също е несъвместима със случайния подбор. Основният принцип при формиране на извадка, която в максимална степен възпроизвежда видовото разнообразие в съвкупността, е селективността, влизаща по правило в противоречие със същността на случайния експеримент. Известна селективност има при районирания случаен подбор, но тя не се изразява в това да бъдат формирани голямо множество различаващи се помежду си и еднородни в себе

си райони, като от всеки да се излъчи извадка с обем $n = 1$, какъвто е смисълът на концепцията „Ноев ковчег".

При не толкова популярната **неселективна концепция за репрезентативност** на пръв поглед изглежда, че е налице пълно съответствие с принципите на случайния подбор, които повеляват безпристрастност по отношение на отделните единици на подбора. Акцентът се поставя не върху качеството на извадката, а върху механизъма на подбор, „който трябва да гарантира определена априорна вероятност на подбора" (Von der Lippe, 2011). Пълна неселективност обаче може да се гарантира единствено от простия случаен подбор, при който (Ламбова и др., 2012):

първо, единиците на подбор са тъждествени с единиците на наблюдението;

второ, вероятността за излъчване на всяка възможна за дадена съвкупност извадка с обем *n* е еднаква и константна.

При този способ на случайния подбор всички единици на съвкупността са с равен шанс за попадане в извадката, следователно отсъства всякаква форма на селективност. В практиката обаче той се използва по-рядко, като се предпочитат способи, които са комбинация от преднамерен и случаен подбор, например районираният подбор. При него районирането се осъществява преднамерено с цел включването в извадката на единици от всички райони, които би трябвало да са „максимално хомогенни в себе си и хетерогенни помежду си по отношение на значенията на изследваните признаци" (Ламбова и др., 2012). Следователно още на етапа на райониране се осъществява целенасочена селекция, която да гарантира излъчването на определени „типове" единици, получени при групировката по значенията на райониращия признак. Неселективност при районирания подбор е налице само при излъчване на извадките от отделните райони. Общата извадка обаче се формира чрез селективен механизъм.

Концепцията на стохастичната грешка е единствената, която обвързва репрезентативността директно със случайния подбор, като представата за репрезентативност е свързана само с процеса на подбор, а не с качеството на излъчената извадка. Следователно при тази концепция може да се говори за репрезентативен подбор, но не за репрезентативна извадка. От случайността зависи дали конкретна извадка, излъчена чрез случаен подбор, ще бъде представителна по отношение на структурата на съвкупността, или миниатюрно нейно копие. Характерно за случайния подбор е, че вероятността за попадане в извадката за всяка единица на съвкупността може да бъде изчислена преди излъчването на извадката. Тази вероятност е свързана с разпределението на статистическата оценка, което поне с приближение е известно предварително и служи като теоретичен модел, на базата на който се стига до статистическо заключение. Теоретичният модел ще съответства с приближение на емпиричното разпределение на статистическата оценка само когато подборът на единиците е случаен, т.е. когато са налице условията за осъществяване на случаен експеримент. Само случайният подбор прави възможно определянето на стохастичната грешка, която измерва разсейването на статистическата оценка и представлява нейното стандартно отклонение. Качеството на случайната извадка, дефинирано като „степен" на репрезентативност по

отношение на структурата на съвкупността или друга концепция за представителност, не оказва влияние върху надеждността и сигурността на статистическото заключение, следователно не съществува изискване случайната извадка да бъде репрезентативна. Подобно изискване би влязло в противоречие с принципите на случайния подбор, които осигуряват спазването на изискванията за провеждане на случаен експеримент.

Казаното дотук позволява да се твърди, че използването на понятието „репрезентативна извадка" като синоним на „случайна извадка" не само е неподходящо, но и неправилно заради несъвместимостта на концепциите за репрезентативност с логическата същност на случайната извадка.

2. Обемът на случайната извадка като инструмент за „обосновката" ѝ?

2.1. Теоретични проблеми при определянето на минимално необходим обем на случайна извадка

Преди да бъде зададен въпросът за обема на сучайната извадка, която трябва да бъде излъчена, е необходимо да е осъществен избор на конкретни статистически извадкови способи и методи, с чиято помощ информацията ще бъде обработена и ще се стигне до статистическо заключение относно изследваната съвкупност.

Според заложения в основата им логическия подход способите на статистиката на заключенията формират два основни класа:

1. Статистическо оценяване на параметри на генералната съвкупност въз основа на характеристиките на конкретна случайна извадка с обем n.
2. Проверка на статистически хипотези относно параметрите или относно разпределението на генерални съвкупности.

Способите на статистическото оценяване, които в българската специализирана литература все още се срещат под некоректното наименование „репрезентативно изучаване", се базират на индуктивен логически подход, а тези за проверка на статистически хипотези – на дедуктивен логически подход.

Под статистически извадков способ тук ще се разбира логическият подход, придаващ форма на изучаването на статистически съвкупности с помощта на случайна извадка. Статистическите извадкови способи не са обвързани с конкретен модел на вероятностно разпределение, следователно не показват конкретната процедура, с чиято помощ се стига до заключение относно състоянието на съвкупността. Те намират приложение чрез редица алтернативни статистически методи, включващи правила, формули и процедури, чрез които се осъществява изучаването на статистическата съвкупност с помощта на случайна извадка. Алтернативните методи, чрез които може да бъде приложен даден статистически извадков способ, се различават най-вече по заложения в основата им теоретичен модел на вероятностно разпределение. Надеждността на даден статистически извадков способ в условията

на конкретна ситуация зависи от избора на статистически метод за неговото приложение. В случай на съответствие между теоретичния модел на разпределение, на който се основава избраният метод, и разпределението на статистическата оценка на изследвания параметър, е гарантирана максимална надеждност на статистическия извадков способ.

Направените разсъждения показват, че преди да бъде определен обемът на случайната извадка, трябва да се направи предварителен качествен анализ на ситуацията, който ще позволи да бъде избран най-подходящият статистически извадков метод. Алтернативните методи са с различни изисквания, свързани с обема на извадката, които пряко кореспондират с правилата за апроксимация на теоретични разпределения. В случаите, когато се избира метод, основаващ се на стандартното нормално разпределение, което апроксимира изходно дискретно разпределение (например биномно или хипергеометрично) на статистическата оценка, обикновено се изискват много по-големи обеми на извадката, които да гарантират спазването на предварително зададената сигурност на заключението или на рисковете за допускане на невярно заключение, отколкото при използването на методи, които се базират на действителното разпределение на статистическата оценка.

Дотук става въпрос само за минималните изисквания относно обема на извадката, които гарантират надеждност на заключенията, направени с помощта на конкретни статистически извадкови методи. Често точно тези изисквания се пренебрегват при практически изследвания. Загубата на надеждност поради неспазването на изискванията не може да бъде компенсирана чрез приложението на сложни формули за определяне на минимално необходим за дадена сигурност и точност обем на извадката, които са изведени от доверителен интервал, основаващ се на теоретично разпределение, което апроксимира друго теоретично разпределение, на което съответства разпределението на статистическата оценка, като не са спазени правилата за апроксимация. Твърдението може да бъде онагледено с помощта на следната примерна ситуация:

За проучването на степента на известност на определен продукт на пазара е необходимо определянето на такъв обем на извадката, който ще гарантира ширина на интервала, възлизаща на 10 процентни пункта, и сигурност 95%. По предположение, изведено от предходни изследвания за сходни продукти, относителният дял на потребителите, познаващи продукта, е 5%.

За определяне на минималния обем на извадката, необходим за оценяването на относителния дял на познаващите продукта при дадените изисквания за точност и сигурност, обикновено се използва формула, изведена от доверителния интервал за относителен дял, основаващ се на стандартното нормално разпределение, което в случая ($\frac{n}{N} \to 0$) апроксимира биномното разпределение на извадковия относителен дял P. Следователно, ако $2e = 0{,}1$ е желаната ширина на интервала, $e = 0{,}05$, $\hat{p} = 0{,}05$, $1 - \alpha = 0{,}95$ и $z_{1-\frac{\alpha}{2}} = z_{0{,}975} = 1{,}96$ е квантил от порядък 0,975 на

стандартното нормално разпределение, тогава минимално необходимият обем на извадката се получава по следния начин:

$n \geq \frac{z^2_{1-\frac{\alpha}{2}}.\hat{p}.(1-\hat{p})}{e^2} = \frac{1{,}96^2.0{,}05.0{,}95}{0{,}05^2} = 72{,}9904$, т.е. той би следвало да бъде $n = 73$.

Ако се използва този обем и в излъчената извадка броят на познаващите продукта възлиза на 4, т.е. $p = 0{,}0548$, тогава съставянето на доверителен интервал на база стандартното нормално разпределение няма да доведе до надеждно заключение, тъй като няма да е изпълнено изискването за допустимост на апроксимацията на биномно от нормално разпределение (Bleymüller, Gehlert and Gülicher, 1992; Mosler and Schmid, 2006; Rüger, 2002), според което $np(1-p) \geq 9$. В случая $np(1-p) = 3{,}7812 < 9$. Ако се бяхме съобразили с правилото за допустимост на апроксимацията, паралелно с определянето на минимално необходим за дадената сигурност и точност обем на извадката с помощта на доверителния интервал на апроксимиращото разпределение, трябваше да определим минимално необходим за допустимост на апроксимацията обем на извадката, т.е. такъв, изчислен по формулата: $n \geq \frac{9}{\hat{p}(1-\hat{p})}$. Този обем в случая възлиза на $n = 190$ и е значително по-голям от $n = 73$.

Какви ще бъдат последствията, ако не се съобразим с правилото за апроксимация и разчитаме единствено на определения с помощта на апроксимиращото разпределение минимално необходим за дадените точност и сигурност обем на извадката? Съставеният въз основа на данните от извадка с изчисления обем доверителен интервал ще се измести спрямо интервала, който съответства на изходното разпределение на статистическата оценка, и ще бъде реализирана сигурност, по-малка от изискуемата, т.е. ще има загуба на сигурност на заключението. Следователно заключение, получено по подобен начин, няма да е надеждно.

За приведения пример доверителният интервал, съставен въз основа на традиционния метод с помощта на стандартното нормално разпределение, е:

$p - z_{1-\frac{\alpha}{2}} s_P \leq \theta \leq p + z_{1-\frac{\alpha}{2}} s_P$, където $s_P = \sqrt{\frac{p(1-p)}{n-1}} = \sqrt{\frac{0{,}0548.0{,}9452}{72}} = 0{,}0268$,

$0{,}05 - 1{,}96 \,.0{,}0268 \leq \theta \leq 0{,}05 + 1{,}96 \,.0{,}0268$

$0 \leq \theta \leq 0{,}1$

Желаната точност, изразяваща се в ширина на интервала, която възлиза на 10 процентни пункта, е спазена, но каква е действително реализираната сигурност на заключението и доколко е надежден полученият резултат? Доверителният интервал, който се основава на изходното биномно разпределение на статистическата оценка, в случая ще бъде (Wissenschaftliche Tabellen Geigy, 1980):

$0,0151 \le \theta \le 0,1344$, като вероятността параметърът θ да попада в този интервал възлиза на:

$$W\left(0,0151 \le \theta \le 0,1344\right) = 1 - \alpha = 0,95$$

Този доверителен интервал може да бъде установен и с помощта на границите на Пирсън-Клопър, основаващи се на връзката между биномно и *F*-разпределение (Hartung, Elpert and Klösner, 2005):

- Долна граница:

$g_д = \dfrac{x.F_{\frac{\alpha}{2};2x;2(n-x+1)}}{n-x+1+x.F_{\frac{\alpha}{2};2x;2(n-x+1)}}$, където x е броят на единиците с въпросното значение на признака в извадката, $F_{\frac{\alpha}{2};2x;2(n-x+1)}$ е квантил от порядък $\dfrac{\alpha}{2}$ на F-разпределение със степени на свобода $\nu_1 = 2.x$ и $\nu_2 = 2.(n-x+1)$.

- Горна граница:

$g_г = \dfrac{(x+1).F_{1-\frac{\alpha}{2},2.(x+1),2.(n-x)}}{n-x+(x+1).F_{1-\frac{\alpha}{2},2.(x+1),2.(n-x)}}$, където $F_{1-\frac{\alpha}{2},2(x+1),2(n-x)}$ е квантил от порядък $1-\dfrac{\alpha}{2}$ на F-разпределение със степени на свобода $\nu_1 = 2.(x+1)$ и $\nu_2 = 2.(n-x)$.

Полученият на базата на апроксимиращото стандартно нормално разпределение доверителен интервал е изместен наляво спрямо действителния, т.е. подценява относителния дял в съвкупността. Действително реализираната сигурност $(1-\alpha)^*$ може да бъде изведена от следните уравнения (Ламбова, 2003):

$$W\left(X \le x-1\right) = F_B\left(x-1/n, g_д\right) = \sum_{k=1}^{x-1}\binom{n}{k} g_д{}^k \left(1-g_д\right)^k = 1-\frac{\alpha}{2}$$

$$W\left(X \le x\right) = F_B\left(x/n, g_г\right) = \sum_{k=1}^{x}\binom{n}{k} g_г{}^k \left(1-g_г\right)^k = \frac{\alpha}{2}$$

След като бъде заместено с получените на базата на стандартното нормално разпределение граници на доверителност, т.е. с $g_д = 0$, $g_г = 0,1$, се получават следните величини:

$W\left(X \le 3\right) = F_B\left(3/73;0\right) = \sum_{k=1}^{3}\binom{73}{k} 0^k \left(1-0\right)^k = 1$, т.е. $\left(1-\dfrac{\alpha}{2}\right)^*.100 = 100\%$

$W(X \le 4) = F_B(4/73,0,1) = \sum_{k=1}^{4} \binom{73}{k} 0,1(1-0,1)^k = 0,1337$, т.е. $\left(\frac{\alpha}{2}\right)^* .100 = 13,37\%$

Двете граници на доверителност се разглеждат независимо една от друга, като за всяка от тях се изисква сигурност, не по-малка от $\left(1-\frac{\alpha}{2}\right)100\%$, т.е. в случая 97.5%. За конкретния пример долната граница, която е 0, е 100% сигурна, докато сигурността за горната възлиза на $(1-0,1337)100 = 86,63\%$, т.е. за нея е реализирана загуба на сигурност, равна на $97,5-86,63 = 10,87$ процентни пункта. Общата действително реализирана сигурност на заключението, основаващо се на интервала, получен с помощта на апроксимиращото стандартно нормално разпределение, в случая е 86.63%, което означава загуба на сигурност в размер на 8.37 процентни пункта.

Следователно незачитането на правилото за допустимост на апроксимацията и акцентирането единствено върху формулата за определяне на минимално необходим обем на извадката, изведена от доверителен интервал, основаващ се на апроксимиращо разпределение, може да доведе до ненадеждни заключения.

Друг съществен проблем при определянето на обема на случайната извадка е свързан със степента на съответствие между емпиричното разпределение на статистическата оценка и теоретичното разпределение, към което то асимптотично се доближава при увеличаване на обема на извадката и поради тази причина се използва за неговото апроксимиране, като лежи в основата на даден параметричен извадков метод. Традиционният подход, използван от класическите извадкови методи, включва предположения за вероятностното разпределение на статистическата оценка. От коректността на направеното предположение зависи надеждността на заключението относно величината на параметъра на съвкупността. При оценяване на средна аритметична величина например се изхожда от следствие на Централната гранична теорема, според което при произволно разпределение в генералната съвкупност извадковата средна $\overline{X}$, изчислена от n на брой независими идентично разпределени случайни величини X_i $(i = 1,2,...,n)$ при $n \to \infty$ е асимптотично разпределена случайна величина с математическо очакване μ и дисперсия $\frac{\sigma^2}{n}$.

Статистическото оценяване на средната аритметична се базира на това разпределение и неговите параметри. При определянето на минимално необходимия за реализиране на дадена точност обем на извадката също се предполага, че извадковата средна $\overline{X}$ е случайна величина с нормално разпределение, при което $E(\overline{X}) = \mu$ и $\sigma_{\overline{X}}^2 = \frac{\sigma^2}{n}$.

Скоростта на конвергенция на разпределението на извадковата средна към нормалното разпределение зависи от степента на асиметрия на разпределението в генералната съвкупност. Колкото е по-силна асиметрията на изходното разпределение в съвкупността, толкова по-бавно разпределението на извадковата средна се доближава до нормално разпределение при увеличаване на обема на извадката. Традиционните правила за допустимост на апроксимацията на

разпределението на статистическата оценка от нормално разпределение предполагат разпределение в съвкупността, което е близко до нормалното.

Следователно основен проблем при използване на традиционния подход при статистическото оценяване е установяването на обем на извадката, гарантиращ достатъчно приближение между извадковото и нормалното разпределение в случай на неизвестно, но вероятно асиметрично разпределение в съвкупността. Според Cochran (1972) определянето на достатъчния за нормалност на разпределението на извадковата средна $\overline{X}$ обем на извадката в зависимост от степента на асиметрия на разпределението в съвкупността може да се осъществи въз основа на следното условие:

$n > 25G_1^2$, където G_1 е моментният коефициент на асиметрията на Фишер: $G_1 = \frac{\sum_{i=1}^{N}(x_i - \mu)^3}{N\sigma^3}$

При силно изразена асиметрия на разпределението в съвкупността достатъчният за нормалност на разпределението на $\overline{X}$ според посоченото условие обем на извадката може да се окаже толкова голям, че да обезсмисли извадковото изучаване. Същевременно могат да се правят само предположения относно асиметрията на неизвестното разпределение в съвкупността, за които не може да се гарантира, че съответстват на действителното състояние.

Съществува още един проблем, свързан с изходното разпределение в съвкупността. Симетрията на това разпределение е необходимо, но не и достатъчно условие за бързото приближение на извадковото разпределение към нормалното при увеличаване на обема на извадката. Възможно е въпреки камбановидната форма на емпиричното разпределение в съвкупността то да се отклонява значимо от нормалното поради наличието на т.нар. heavi tails (Pöhlmann, 1987). Става въпрос за голяма вероятностна маса в краищата на разпределението, т.е. относителният дял на единиците с екстремни значения е по-голям от очаквания при нормално разпределение. Това може да доведе до съществено разминаване между теоретични и действителни вероятности, което ще бъде свързано със загуба на сигурност и недостатъчна надеждност на заключението.

Посоченото дотук позволява да се твърди, че при определени условия традиционният параметричен подход при статистическото оценяване не е в състояние да гарантира приемливо качество на резултатите поради несъответствие между извадковото и теоретичното разпределение, към което то се стреми при увеличаване на обема на извадката, но само при положение, че неизвестното разпределение в съвкупността е приблизително нормално. Заключение, направено въз основа на случайна извадка с обем, изчислен по формулата, изведена от доверителния интервал, който се базира на нормално разпределение, при посочените условия, несъответстващо на разпределението на статистическата оценка, няма да да бъде надеждно, като загубата на сигурност не подлежи на контрол.

Проблем при определяне на оптималния обем на случайната извадка възниква и при използването на извадкови методи, при които статистическата оценка не е обвързана с конкретно теоретично разпределение. Все по-често в практиката и в научните изследвания се прилагат рисемплинг методи, сред които например е методът бутстреп. Те са алтернатива на класическите параметрични методи и позволяват заключения в случаите, когато се очаква последните да дадат ненадеждни резултати. Основно предимство на метода бутреп е отчитането на разпределението в реалната извадка, като по този начин се цели приблизително пресъздаване на разпределението в съвкупността. Следователно при него се взема под внимание и асиметрията, което не е възможно при традиционните извадкови методи, предполагащи приблизително нормално разпределение на статистическата оценка независимо от вида на разпределението в съвкупността. Това не означава, че бутстреп е универсално по-надежден. Недостатък на метода е, че той се опира изцяло на разпределението в една-единствена реална случайна извадка, която може да бъде с различна степен на представителност или дори непредставителна по отношение на съвкупността. При значително несъответствие между разпределението в съвкупността и това в извадката, което, както вече беше посочено, при случаен подбор е възможно, методът няма да даде надеждни резултати. Друг основен проблем, на който трябва да се обърне внимание, е невъзможността за определяне на минимално необходим обем на извадката, който да гарантира зададена предварително точност. За разлика от традиционните методи при бутстреп не е възможно използването на информация за зависимостта между изискуемите сигурност и точност на оценката, от една страна, и обема на извадката, от друга. Квантилите на бутстреп-разпределението се установяват след излъчване на извадката, което прави невъзможно задаването на точността на оценката при определен обем на извадката преди нейното излъчване.

Дотук разсъжденията относно необходимия обем на извадката засягат основно способите и методите за статистическо оценяване, но подобни проблеми възникват и при дедуктивния подход на статистическите заключения, т.е. при проверката на статистически хипотези, където възможностите за предварително определяне на необходим обем на извадката, който да гарантира нещо, са още по-ограничени. По правило при непараметричните тестове не съществува възможност за изчисляване на минимално необходим обем на извадката. Множество непараметрични тестове са съпроводени с изисквания относно обема на извадката, свързани с разпределението на използвания статистически критерий.

Логиката на определянето на необходим обем на извадката при параметрични статистическите тестове се различава коренно от тази при статистическото оценяване освен в случаите, когато обемът трябва да гарантира допустимостта на апроксимацията на разпределението на статистическата оценка чрез разпределението на статистическия критерий на теста, например при използването на Z-критерий за проверка на предположение относно биномно или хипергеометрично разпределен извадков относителен дял P. Докато при статистическото оценяване целта е гарантирането на предварително зададените точност и сигурност на заключението, при проверката на статистически хипотези изчисляването на необходим обем на случайната извадка се налага, когато вероятността за допускане на грешка от втори

род β е зададена предварително заедно с равнището на значимост α. Целта е спазването на предварително зададената величина на риска за допускане на грешка от втори род , т.е. на вероятността β за приемането на невярна нулева хипотеза. До каква степен определеният обем на извадката ще гарантира спазването на β, зависи от начина на формулиране на нулева и алтернативна хипотеза. Единствено когато двете хипотези са точкови (конкретизирани), е възможно изчисляването на фиксиран обем на извадката. В останалите случаи той е променлива, която зависи от размера на разликата между предполагаемата и действителната величина на тествания параметър. Ако например се проверява предположение относно средната величина на съвкупност, тогава фиксиран обем на извадката може да бъде определен само тогава, когато хипотезите са формулирани по следния начин:

$H_0 : \mu = \mu_0$

$H_1 : \mu = \mu_1,$ където $\mu_0 \neq \mu_1$

Разликата $|\mu_0 - \mu_1|$ е фиксирана, като необходимият за спазването на предварително зададените вероятности за допускане на грешка от първи и втори род зависи от нея и от разсейването в съвкупността (Polasek, 1997):

$$n \geq \left[\sigma \frac{z_{1-\alpha} + z_{1-\beta}}{\mu_0 - \mu_1} \right]^2$$

В повечето случаи обаче се работи с интервални алтернативни хипотези и разликата $|\mu_0 - \mu_1|$ няма да бъде фиксирана, тъй като μ_1 тогава е променлива. В случай, че хипотезите са формулирани по следния начин:

$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0,$

необходимият обем на извадката ще се променя в зависимост от абсолютния размер на разликата между предполагаемата и действителната средна, т.е. може да бъде изчислен такъв за различни величини $\mu_1 \neq \mu_0$ по следната формула (Harung, Elpert, Klösner, 2005):

$$n \geq \left[\sigma \frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{\mu_0 - \mu_1} \right]^2$$

Той ще осигури спазването на β единствено н точката $\mu = \mu_1 \neq \mu_0$.

За онагледяване на казаното може да послужи следната примерна ситуация:

Трябва да бъде определен необходим обем на извадката за проверка на предположение относно средната възраст на студентите от даден университет при $\alpha = 0{,}05,\ \beta = 0{,}05$*, като хипотезите са формулирани по следния начин:*

$H_0 : \mu = 22$ *години*

$H_1 : \mu \neq 22$ *години*

Предполага се дисперсия на разпределението по възраст в съвкупността, възлизаща на 9 години^2.

При така формулирани хипотези действителната средна може да бъде $\mu = 22$ години или $\mu = ...19,20,21,23,24,25...$ години при положение, че се задава чрез целочислени значения. За определянето на необходим обем на извадката трябва да се фиксира конкретна алтернативна величина, например $\mu_1 = 25$ години. Тогава се получава следният обем на извадката:

$$n \geq \left[\sigma \frac{z_{0,975} + z_{0,95}}{\mu_0 - \mu_1} \right]^2 = \left[\sqrt{9} \frac{1{,}96 + 1{,}645}{22 - 25} \right]^2 = 12{,}996,$$ т.е. $n = 13$.

Следователно извадката трябва да бъде с обем минимум 13 единици, ако:

1) нулевата хипотеза се допуска да бъде отхвърлена погрешно в максимум 5% от случаите на подобен експеримент;

2) алтернативната хипотеза се допуска да бъде приета погрешно в 5% от случаите, ако действителната средна е 25 години.

Ако действителната средна се намира между 22 и 25 години, при този обем на извадката не могат да се правят съждения относно величината на риска за допускане на грешка от втори род, т.е. той не може да гарантира спазването на β. Ако действителната средна е 24 години, тогава извадката би трябвало да е с обем 30 единици, а при $\mu_1 = 23$ години в нея би трябвало да се включат 117 единици, за да може да се гарантира спазването на зададената величина на риска за допускане на грешка от втори род.

Колкото е по-малка абсолютната разлика между предполагаемата и действителната величина на параметъра, толкова по-голям обем на извадката ще бъде необходим, за да се гарантира спазването на β. Проблемът е в това, че действителната величина на параметъра е неизвестна, в противен случай не би било необходимо да се проверяват предположения за нея. Следователно определеният по посочения начин необходим обем на извадката в крайна сметка не може да гарантира, че рискът от приемането на невярна нулева хипотеза няма да надвиши предварително зададената величина β.

Направените разсъждения относно проблемите, свързани с определянето на необходим обем на случайната извадка при използването на различни извадкови способи и методи, не претендират за изчерпателност, но смятаме, че очертават

основните моменти, на които трябва да се обръща внимание, когато се планира обемът на случайните извадки. Преди да се пристъпи към определяне на обем на извадката, би трябвало да са налице отговори на следните въпроси:

- Какви статистически извадкови способи и методи ще бъдат използвани в изследването?
- Съществуват ли изисквания относно минималния обем на извадката при избраните извадкови методи?
- Какво е разпределението на статистическата оценка и при какви условия то съответства на или се доближава до теоретичното разпределение, на което се основава избраният метод?
- Необходимо ли е да се съобразяваме с правила за апроксимация на теоретични разпределения при определянето на необходимия обем на извадката?
- Какво искаме да гарантира планираният обем на извадката?

В случай, че няма отговор на последния въпрос, изчисляването на обем на извадката става безпредметно. В подобна ситуация до надеждни заключения ще се стигне, ако се съобразим с теоретичните изисквания относно обема на извадката на избрания извадков метод и при необходимост – с правилата за апроксимация.

2.2. Обем и „обосновка" на случайната извадка

Както в научните изследвания, така и в практиката извадките често се „обосновават" с помощта на обем, изчислен по определена формула. Според нас терминът и вложеното в него съдържание са некоректни от гледна точка на статистическата методология поради следните причини:

- Изчисленият по формула, изведена от даден доверителен интервал, обем на извадката не е аргумент за нейната годност при приложението на статистически извадкови способи и методи.
- Обемът на извадката по никакъв начин не гарантира случайния характер на нейния подбор.
- Определеният по дадена формула обем на извадката не е универсален за всички променливи и методи, включени в изследването.
- Изчисленият минимално необходим обем не обосновава извадката, а гарантира, че при определени условия заключенията ще бъдат с качество, което съответства на предварително зададени параметри.

При използването на способи и методи от теорията на статистическите заключения при извадковите изследвания съществува едно единствено общо изискване към извадките – те трябва да са излъчени чрез случаен подбор. Следователно годността на дадена извадка за изследване с помощта на статистически извадкови методи не зависи на първо място от нейния обем, а от начина на подбор на единиците, т.е. ако

използваме термина „обосновка на извадката", то той би трябвало да е свързан със способа и техниката на случаен подбор, които се използват за нейното излъчване.

След избора на конкретни статистически извадкови методи, с чиято помощ да се стигне до заключение за изследваната съвкупност, се поставя въпросът за необходимия обем на случайната извадка. Той не я обосновава, а е свързан с надеждността на заключенията, до които се стига при приложението на избраните извадкови методи. Необходимият обем на извадката зависи от следните фактори:

- изисквания относно минимален обем на извадката на статистическия извадков метод;
- правила за апроксимация на теоретични разпределения, когато заложеният в основата на статистическия извадков метод модел на разпределение апроксимира разпределението на статистическата оценка;
- изискуеми точност и сигурност на заключението, направено под формата на доверителен интервал;
- максимално допустими рискове за допускане на грешка от първи и от втори род при заключение, формирано на базата на параметричен статистически тест.

Първите два фактора гарантират надеждност на резултатите, свързана със спазването на предварително зададените рискове за неправилно заключение – съответно риска за построяването на доверителен интервал, който не съдържа оценявания параметър на съвкупността, респ. за отхвърляне на вярна нулева хипотеза, т.е. равнището на значимост при проверка на статистически хипотези.

Третият и четвъртият фактор имат значение само в случаите, когато при планирането на изследването са включени параметрите точност на заключението при статистическо оценяване или конкретизиран риск за допускане на грешка от втори род (β) при проверка на статистически хипотези.

Посоченото дотук води до твърдението, че първите два фактора задължително трябва да се вземат под внимание при всяко изследване с помощта на статистически извадкови методи. Следователно, ако все пак се използва терминът „обосновка" на извадката при планирането на нейния обем като синоним на гарантираната надеждност на бъдещите заключения, то той би трябвало да бъде свързан с тях, а не с третия и четвъртия фактор, които играят роля само в случаите, когато трябва да бъде гарантирано спазването на предварително планирана точност или максимален риск за приемане на невярна нулева хипотеза.

На практика обаче точно третият фактор много често се издига до ранг „задължителен" независимо от предварително зададените параметри на изследването, от вида и броя на променливите и от избраните статистически извадкови способи и методи, без да се задава въпросът какво гарантира той в конкретната ситуация. Изчисляването на необходим обем на случайната извадка по формула, изведена от доверителен интервал, използван при конкретен метод за статистическо оценяване на даден параметър на съвкупността, би имало смисъл само

тогава, когато заключението, до което трябва да се стигне, се отнася за същия параметър и ще бъде резултат от анализ с помощта на същия метод, като предварително е планирана точността и сигурността на доверителния интервал. Единствено в подобна ситуация изчисленият по формулата обем на извадката ще гарантира определено качество на резултата – в случая спазването на предварително зададената точност на заключението относно неизвестния параметър на съвкупността. Не съществува универсална формула, по която може да бъде определен необходим обем на случайната извадка, който да гарантира надеждност на заключенията относно различни параметри на съвкупността, получени като резултат от приложението на различни статистически извадкови методи.

В повечето случаи не се работи само с един статистически признак, т.е. информацията за извадката се използва за анализ на различни променливи. Това означава, че се оценяват параметрите на разпределението по различаващи се според скалирането, разсейването, начина на измерване признаци, като за всеки от тях би трябвало да се установи минимално необходим обем на извадката, който да гарантира предварително зададените точност и сигурност. Възможностите са няколко:

- да се използва най-големият получен по формулите за отделните променливи и параметри минимално необходим обем;
- да не се задава предварително изискуема точност на интервалната оценка, като обемът бъде съобразен с изискванията за приложение на отделните методи за оценяване или правилата за апроксимация, когато доверителният интервал се основава на разпределение, което е само приближение на разпределението на статистическата оценка като случайна величина;
- да се направи компромис по отношение на точността и/или сигурността на оценяването на параметри, за които се изисква прекалено голям обем на извадката, и да се използва по-малък, получен като минимално необходим за останалите променливи;
- да се търсят алтернативни методи за оценяване, при които параметрите на променливи, за които традиционните методи по правило изискват прекалено голям обем на извадката, да могат да бъдат оценени без загуба на сигурност и точност с помощта на значително по-малък обем на извадката. Това се отнася най-вече за оценяването на относителната честота на единиците от даден клас, когато тя клони към 0 или 1.

В множество изследвания извадката предварително се „обосновава" с помощта на обем, гарантиращ спазването на определена точност при статистическо оценяване на даден параметър на съвкупността, но впоследствие се стига до заключения чрез проверка на статистически хипотези. Безсмислието на подобна „обосновка" е очевидно – изчисляването на „необходим" обем на извадката в случая е самоцелно, тъй като този обем по никакъв начин не е обвързан с надеждността на заключението, т.е. той не гарантира нищо. Както вече беше посочено, определянето на необходим обем на извадката при проверка на статистически хипотези е свързано единствено с

гарантиране на спазването на предварително зададения риск за допускане на грешка от втори род (β) при параметрични статистически тестове с конкретизирана алтернативна хипотеза. Във всички останали случаи обемът на извадката би трябвало да бъде съобразен с изискванията за приложение на конкретните статистически тестове, както и с правилата за апроксимация на теоретични разпределения, когато статистическият критерий е с разпределение, апроксимиращо разпределението на статистическата оценка (извадковата характеристика като случайна величина).

Заключение

В съответствие с поставената цел и произтичащите от нея задачи изследването включва две направления:

1. Аргументация на логическото различие между т. нар. репрезентативна и случайната извадка;
2. Аргументация против придобилото популярност твърдение, според което всяка „репрезентативна" извадка задължително трябва да бъде с изчислен по определена формула обем, за да се счита за „обоснована".

Разсъжденията и аргументите, свързани с първото направление, позволяват следните основни изводи:

- Не съществува критерий за измерване на степента на представителност на т. нар. репрезентативни извадки.
- Няма единна дефиниция за понятията „статистическа репрезентативност" и „репрезентативна извадка".
- Различните концепции за репрезентативност са частично или цялостно несъвместими, като са налице редица противоречия между тях.
- Повечето концепции за репрезентативност са логически несъвместими със същността на случайния подбор и с възможните резултати от случаен експеримент.
- Качеството на случайната извадка, дефинирано като „степен" на репрезентативност по отношение на структурата на съвкупността или друга концепция за представителност, не оказва влияние върху надеждността и сигурността на статистическото заключение, следователно не съществува изискване случайната извадка да бъде репрезентативна.
- Използването на понятието „репрезентативна извадка" като синоним на „случайна извадка" не само е неподходящо, но и неправилно заради несъвместимостта на концепциите за репрезентативност с логическата същност на случайната извадка.

Въз основа на разсъжденията и аргументите, свързани с второто направление на изследването, могат да бъдат направени следните обобщаващи изводи:

- Изчисленият по формула обем на извадката не е критерий за нейната годност при приложението на статистически извадкови способи и методи, за което съществува едно-единствено общо изискване към извадките – те трябва да бъдат излъчени чрез случаен подбор.
- Обемът на извадката по никакъв начин не гарантира случайния характер на нейния подбор.
- Определеният по дадена формула обем на извадката не е универсален за всички променливи и методи, включени в изследването. Той е обвързан с конкретен параметър на съвкупността и с конкретен статистически метод, с помощта на който се стига до статистическо заключение.
- Изчисленият обем не обосновава извадката, той гарантира, че при определени условия и при използването на конкретни статистически извадкови методи заключенията ще бъдат с качество, което съответства на предварително зададени параметри, т.е. предварително планирана точност и сигурност при статистическо оценяване или максимално допустим риск за приемане на невярна нулева хипотеза при параметрични статистически тестове.
- Когато точността на заключението не се планира предварително, респ. допустимият риск за приемане на невярна нулева хипотеза при използване на параметрични тестове или заключенията трябва да бъдат направени с помощта на непараметрични тестове, определянето на необходим обем на извадката е свързано единствено с изискванията на конкретния статистически извадков метод и/или с правилата за апроксимация на теоретични разпределения, при положение че заложеният в основата на метода модел на разпределение апроксимира разпределението на статистическата оценка.

Използвана литература

Ламбова, М., Русев, Ч., Косева, Д., Стоянова, В. (2012). Въведение в статистиката. Варна: ИК „СТЕНО".

Ламбова, М. (2003). Надеждност на статистическото оценяване на честотата на единиците от даден клас при безвъзвратен подбор на извадката. Варна: ИК „СТЕНО".

Bleymüller, J., Gehlert, G., Gülicher, H. (1992). Statistik für Wirtschaftswissenschaftler. 8. Auflage. München, Verlag Vahlen.

Bourier, G. (2002). Wahrscheinlichkeitsrechnung und schließende Statistik. 3. Auflage.Wiesbaden, Gabler Verlag.

Cochran,W. G. (1972). Stichprobenverfahren. Berlin, New York: Walter de Gruyter Verlag.

Hartung, J., Elpert, B., Klösner, K.-H. (2005). Statistik. Lehr- und Handbuch der angewandten Statistik. München, Oldenbourg Verlag.

Kruskal, W., Mosteller, Fr. (1979). Representative Sampling. Part I: Non-Scientific Literature. – International Statistical Review, Vol. 47, N 1, pp. 13-24.

Kruskal, W., Mosteller, Fr. (1979). Representative Sampling. Part II: Scientific Literature Excluding Statistics. – International Statistical Review, Vol. 47, N 2, pp. 111-122.

Kruskal, W., Mosteller, Fr. (1980). Representative Sampling. Part IV: the History of the Concept in Statistics, 1895-1939. – International Statistical Review, Vol. 48, N 2, pp. 169-195.

Mosler, K., Schmid, Fr. (2006). Wahrscheinlichkeitsrechnung und schließende Statistik. 2. Auflage. Berlin, Heidelberg, New York: Springer Verlag.

Pöhlmann, H. (1987). Jahresabschlussprüfung auf Stichprobenbasis. Pfaffenweiler, Centaurus-Verlagsgesellschaft.

Polasek, W. (1997). Schließende Statistik. Berlin, Heidelberg, New York, Springer Verlag.

Ross, Sh. M. (2006). Statistik für Ingenieure und Naturwissenschaftler. 3. Auflage. München, Spektrum Akademischer Verlag.

Rüger, B. (2002). Test- und Schätztheorie. Band II: Statistische Tests. München, Wien: Oldenbourg Verlag.

Sachs, L. (2004). Angewandte Statistik. Anwendung statistischer Methoden. 11. überarbeitete und aktualisierte Auflage. Berlin, Heidelberg, New York, Springer Verlag.

Schnell, R., Hill, P. B., Esser, E. (1995). Methoden der empirischen Sozialforschung. München, Wien, Oldenbourg Verlag.

Von der Lippe P., Kladroba, A. (2002). Repräsentativität von Stichproben. – Marketing. ZFP 24, pp. 227-238.

Von der Lippe, P. (2011). Wie groß muss meine Stichprobe sein, damit sie repräsentativ ist? Diskussionsbeitrag aus der Fakultät Wirtschaftswissenschaften derUniversität Duisburg-Essen, No. 187, Campus Essen.

Wissenschaftliche Tabellen Geigy. (1980). Teilband Statistik. 8. Auflage. Basel, CIBA-GEIGY AG.