Глава 4

РЕПРЕЗЕНТАТИВНИ ПРОУЧВАНИЯ

1. Същност на репрезентативните проучвания

Както бе подчертано в предходния раздел, двете основни понятия в статистиката са *„популация“ и „извадка“.* Когато правим заключение за характеристиките на дадена популация на базата на данни, получени при изучаване на определена част от популацията, ние всъщност използваме информация от извадки, за да направим индуктивни изводи за популацията. Такава информация има известни ограничения по отношение на надеждността, точността и валидността. Но независимо от това, проучванията, опиращи се на наблюдения на извадки, представляват най-честата практика в научните изследвания.

Основните доводи за предпочитане на репрезентативните пред цялостните проучвания се свеждат най-често до следното:

– ограничените финансови и материални ресурси;

– ограничения във времето;

– липсата на достъп до цялата популация (тъй като понякога тя е безкрайна) и наблюдението на извадка може да е единственият възможен метод за събиране на информация.

Всичко това важи особено за медицината и здравеопазването, където поради масовия характер на явленията и ограничените ресурси много рядко могат да се обхванат всички единици на дадена популация. Затова медицинското познание, което имаме днес за човешките популации, се опира преди всичко на ***репрезентативни проучвания,*** на основата на които се правят изводи и заключения за съответните популации, от които са извлечени наблюдаваните извадки.

Репрезентативните проучвания имат редица важни ***предимства*** пред цялостните проучвания:

– намаляват потребностите от финансови, човешки и материални ресурси, т.е. те са по-икономични;

– резултатите се извличат по-бързо;

– постига се по-голяма точност на данните, тъй като помалките извадки позволяват да се вложат повече усилия за намаляване на грешките, несвързани с подбора и намаляване на систематичните грешки поради неотзоваване на изследвани лица;

– по-точни са от цялостните проучвания, колкото и учудващо да е това на пръв поглед. При тях се допуска ***репрезентативна грешка*** (наричана още ***случайна или стохастична***), но тя може да се определи сравнително точно (да се изчисли математически) и да се предвиди в крайните изводи за популацията. Тя зависи от броя на наблюдаваните случаи и начина на подбор на извадката и може да бъде сведена до минимум чрез ***наблюдение на достатъчен брой непреднамерено подбрани случаи***, докато при изчерпателните проучвания се допуска грешка на регистрацията, която може да нараства с броя на наблюдаваните случаи и трудно се контролира.

Недостатъци на репрезентативните проучвания:

– винаги се наблюдава ***репрезентативна грешка;***

– работата с извадки може да създаде ***чувство за дискриминация*** в рамките на популацията;

– при някои ***редки явления*** малките извадки може да не предоставят достатъчно случаи за изучаване.

При репрезентативните проучвания се постигат ***три главни цели:***

– да се направят изводи за цялата популация въз основа на анализа на данни от съответни извадки;

– да се направи оценка на степента на сигурност, т. е. да се определи гаранционната вероятност на изводите за популацията;

– да се определи начинът на сформиране на извадката и нейния размер, така че да се осигурят точни и надеждни изводи и заключения.

2. Основни понятия и принципи при подбор на извадки

Извадка (Sample) – подмножество, подклас от дадена популация, чиито свойства предстои да бъдат изучени и генерализирани за популацията.

Подбор на извадка (Sampling) – това е самата процедура, самият метод за избиране на индивиди от една или повече популации.

Извадкова единица (Sampling unit) – това е най-малката единица, използвана в процеса на подбор (лице, домакинство, район).

Единица на изследване или наблюдение (Unit of inquiry) – най-малката единица, върху която се събират данните (извършва се измерване или запис) при едно извадково проучване.

Рамка на извадката (Sampling frame) – това е наборът от извадкови единици, от които се подбира дадена извадка (напр., списък с имена, места или други неща, които се използват като единици на извадката).

Извадкова фракция (Sampling fraction) – това е пропорцията от извадкови единици, които трябва да се подберат от определена извадкова рамка за включване в извадката.

Съществуват два основни подхода за подбор:

- ***непреднамерен подбор***
- ***преднамерен подбор***

Една ***добра извадка*** трябва да бъде:

– подбрана случайно с цел да се намали систематичната грешка;

– репрезентативна по отношение на популацията с цел да се подобри нейната валидност;

– достатъчно голяма с цел да се повиши нейната точност.

Явно е, че на тези условия отговаря единствено ***непреднамереният случаен подбор*** (probality sampling scheme – схема за вероятностен подбор), при който ***всеки индивид (елемент) от популацията има еднакъв шанс да бъде включен в извадката.***

Най-важното изискване, на което трябва да отговаря всяка извадка, е тя да бъде ***репрезентативна (представителна)*** по отношение на популацията. Само при наличието на такова съответствие

между данните от извадката и цялата популация е възможно обобщаване на данните за популацията.

Мярка за репрезентативността на всяка извадка е разликата между средните и относителни величини в извадката и цялата популация, която се измерва с ***репрезентативната грешка.*** Най-голямото предимство на извадка, подбрана чрез непреднамерен подбор в сравнение с преднамерен подбор е в това, че може да се оцени размера на репрезентативната грешка и да се вземе предвид при обобщаване на данните. В зависимост от характера на изучаваните признаци (качествени или количествени), величината на репрезентативната грешка се определя по различен начин.

Репрезентативността зависи от:

- ***числеността на изучаваната извадка;***
- ***от начина на сформиране на извадката***.

Съгласно ***закона за големите числа,*** с увеличаване на броя на наблюдаваните случаи намалява влиянието на случайността и все повече се проявява закономерността. С други думи, с увеличаване на броя на наблюдаваните случаи намалява величината на репрезентативната грешка и резултатите от наблюдението на извадката все повече се приближават до тези за цялата популация. Има специални статистически методи, които позволяват да се определи ***необходимият и достатъчен брой случаи за наблюдение***, така че величината на репрезентативната грешка да бъде сведена до минимум и достоверността на изводите за популацията да бъде достатъчно висока.

3. Видове извадки

Проста случайна извадка (simple random sample)

Това е най-простият случай на подбор на извадки, подходящ за ограничени по размер проучвания. Характерните особености на този подход са:

– всяка единица в извадковата рамка има равен шанс да бъде избрана;

– случайният подбор от извадковата рамка може да бъде направен чрез теглене на жребий, използване на таблица за случайните числа или чрез компютърни програми, генериращи случайни числа;

– необходим е списък на единиците в извадковата рамка.

Напр., за да подберем проста извадка от 50 случая от масив от 800 инвалидизирани лица в гр. А (определен чрез серийни номера от 1 до 800 в документацията на Териториалните експертни лекарски комисии), бихме могли да проследим последователно трицифрени числа от таблицата за случайните числа. Така например, ако табличните означения са 12454, 45730, 07944, 73506, 81149,....., то тогава избираме числата 124, 544, 573, 007, 944, 735, 068 и т. н. докато наберем 50 случая. При това числата над 800 се пренебрегват (тъй като цялата популация е максимум от 800 случая и няма случай с пореден номер 944 напр.), а също така трябва да се внимава с нулите. Ако те не се вземат предвид, има опасност да не включим случаи с поредни номера под 100, което ще бъде съвсем погрешно.

Предимства: Тъй като всяка единица в популацията има равен шанс за включване в извадката, репрезентативността е гарантирана и извадката е изложена само на репрезентативна грешка. Оценъчните характеристики на извадката се изчисляват лесно.

Недостатъци: Ако извадковата рамка е голяма, този метод може да се окаже непрактичен, поради трудността и разходите по съставяне и актуализиране на списъка на извадковите единици.

Систематична извадка (systematic sample)

Нарича се още ***механична (пропорционална)*** извадка и се характеризира се със следните особености:

– включва избор на всяка n-та единица в популацията или извадковата рамка, където 1/n представлява извадковата фракция (пропорцията от извадкови единици, които трябва да се подберат);

– първата единица (стартовото число) се избира случайно сред първите единици в зависимост от извадковата фракция (напр. сред първите 10, ако извадковата фракция е 10%);

– след определяне на стартовото число започва подбор със стъпка равна на извадковата фракция;

– необходим е списък на единиците в извадковата рамка.

Например, за да подберем напр. 10% извадка от популация с численост N, първо избираме случайно началната точка между числата 1 и 10 и по-нататък в извадката включваме всеки 10-ти случай от популацията. Ако сме избрали случайно за старт числото 6, по-нататък включваме в извадката 6-я, 16-я, 26-я, 36-я и т. н.

Този подход е особено полезен, когато случаите се подреждат автоматично с течение на времето, както е при постъпване и изписване на хоспитализираните болни.

Сериозен проблем при систематичния подбор е възможността за допускане на систематична грешка (bias), ако има някаква определена тенденция в популацията, от която се сформира извадката. Например, при изучаване на характеристиките на пациенти, приети в отделенията за спешна помощ, неразумно е да подбираме само приетите в събота през нощта, тъй като пациентите, постъпили в началото и в средата на седмицата могат да се различават доста съществено по редица основни характеристики.

Предимства: извадката се подбира лесно; лесно може да се определи подходяща извадкова рамка; извадката е равномерно разпределена върху цялата референтна популация.

Недостатъци: извадката може да бъде повлияна от систематична грешка, ако някаква скрита периодичност съвпадне със стартовото число и извадковата фракция; трудно е да се определи точността на оценката от едно проучване.

Стратифицирана извадка (stratified sample)

Нарича се още ***послойна извадка*** (от strata – слой). Характеризира се със следните особености:

– популацията първо се разделя на групи или слоеве според характеристиките, от които се интересуваме (напр. пол, възраст);

– след това се подбира проста случайна извадка от всеки слой, използвайки една и съща извадкова фракция, освен ако не се препоръчва нещо друго по специални причини.

Например, данните от преброяването на населението показват, че в даден регион възрастовото разпределение е: 0-14 г. – 15%; 15-34 г.– 20%; 35-49 г. – 27%; 50-64 г. – 22%; 65 г. и + – 16%.

Ако се нуждаем от извадка с численост 200 души, то тогава умножаваме процентите за всеки слой по 200 и ги разделяме на 100 като по такъв начин определяме по колко случая трябва да се включат в извадката от всеки слой (страта) на популацията. В нашия пример това означава, че от различните възрастови категории трябва да включим съответно 30, 40, 54, 44 и 32 случая. Отделните случаи във всеки слой се подбират по метода на случайните числа или систематично (напр. всяко 30-то лице на възраст до 14 г.).

Предимства: Всяка единица в даден слой има равен шанс да бъде избрана; използването на една и съща извадкова фракция за всички слоеве гарантира пропорционална представителност в извадката на характеристиките, според които популацията е стратифицирана.

Недостатъци: Извадковата рамка от цялата популация трябва да бъде подготвена отделно за всеки слой.

Гнездова (кластерна) извадка

Характеризира се със следните особености:

– популацията първо се разделя на кластери (гнезда) от хомогенни единици, обикновено опиращи се на географска близост;

– подбира се извадка от всички гнезда;

– вътре в избраните гнезда се изследват или изучават всички единици.

Предимства: намалява разходите за изготвяне на извадкова рамка и за пътуване между избраните единици.

Недостатъци: репрезентативната грешка е обикновено по-висока, отколкото при проста случайна извадка от същия размер.

Комбинирана или многостепенна извадка (multistage sample)

Този подход се прилага, когато желаем да сформираме извадка за широкомащабно проучване с голям географски обхват. Както означава името й, подборът се осъществява на етапи докато не се достигне до крайните единици на извадката:

– в първия етап се изготвя списък на големи извадкови единици (градове, села, училища и др.);

– от този списък се подбира случайна извадка с вероятност на подбора пропорционална на размера;

– за всяка от избраните в първия етап единици се съставя списък на по-малки извадкови единици (напр., ако в първия етап единиците са били градове, тогава във втория етап единиците могат да бъдат къщи или домакинства);

– след това се подбира случайна извадка от единиците на втория етап, които се изучават;

– процедурата може да съдържа три или повече етапа.

*Предимства: н*амалява разходите за изготвяне на извадкова рамка.

Недостатъци: репрезентативната грешка нараства в сравнение с проста случайна извадка от същия размер.

Съществуват и други методи, но те са по-малко надеждни.

Извадки по удобство (convenience samples) – включват избиране на всички лица, които желаят да участват в дадено проучване, при условие, че тези лица отговарят на установени критерии. С други думи, извадките включват индивиди, които най-лесно могат да бъдат разпитани или изследвани, което е предимство от една страна, но те не са достатъчно представителни за цялата популация.

Самоформирали се извадки (self-selected samples) – например, при пощенските анкети лицата, които отговарят на изпратения по пощата или публикуван във вестници или списания въпросник, вероятно се различават от неотговорилите лица. Обикновено тези, които са по-удовлетворени или по-неудовлетворени, отговарят почесто от останалите. Неотговорилите, от друга страна, могат да окажат съществено влияние върху достоверността на резултатите, което трудно може да се прецени.

Въпреки това, посочените два подхода понякога са необходими – напр., при провеждане на пилотни проучвания за тестуване на въпросниците може да се използва извадка по удобство.

4. Систематични грешки при репрезентативни проучвания

Систематична грешка (bias) е всяка тенденция на дадена извадка да се отклонява извън случайното вариране от съответната популация. Тя се причинява от странични, объркващи фактори (confounding effects).

Целта на сформиране на представителна извадка е именно намаляване на систематичната грешка или в идеалния случай – свеждане на нейното влияние до незначително.

Ако е налице сериозна систематична грешка, то увеличаването на обема на извадката може да задълбочи проблема.

Следователно, още на етапа на планиране на проучването трябва да бъдем много внимателни по отношение на възможните източници на систематични грешки и да се опитваме да ги отстраним, когато това е възможно.

Някои източници на такива грешки са посочени по-горе при разглеждане на проблема за подбора на извадката:

– ***систематични грешки при самоформирали се извадки*** (при пощенски анкети);

– ***систематични грешки поради неотзоваване*** или необхващане на всички предвидени лица в извадката;

***– небрежност при прилагане на методите на систематичен подбор**;*

***– включване на резултати от пилотно проучване върху извадка по удобство при анализа на данните от основното проучване**.*

Следователно, ***систематичните грешки, свързани с подбора на извадката***, са най-сериозния проблем.

Други видове систематични грешки са свързани с:

– ***грешки на припомнянето*** *(recall bias)* – те са свързани със селективната способност на паметта и са характерни за ретроспективните проучвания, когато се събират данни за минали събития, *история на заболяването и др.*

– ***систематични грешки, свързани с изследователя*** (*interviewer bias*) – ентусиазъм, свенливост, страх, чарът на противоположния пол и много други фактори се комбинират, понякога подсъзнателно, в съзнанието на интервюиращия и могат да доведат до преднамереност в подбора на изследваните лица.

– ***систематични грешки, свързани с оценяващия*** *(assessor bias)* – оценката на изследователя може да бъде повлияна от предварителната информация, с която той разполага за изследваното лице (или лечение).

– ***систематични грешки поради липсващи данни*** *(missing data bias*) – това се наблюдава често, когато се проучват документи за минал период, съставени по друг повод, а не за целите на дадено проучване.

– ***систематични грешки, свързани с въпросника*** (*questionnaire bias*) – какви въпроси се задават, как са формулирани, какъв е редът на задаването им и т. н.

5. Групови свойства на статистическите съвкупности

Статистическата съвкупност представлява ***множество относително еднородни елементи, взети заедно в известни граници на времето и пространството*** – например, населението

на даден район през дадена година, групата на живородените или мъртвородените през дадена година в дадена територия и т.н.

Първичните елементи, от които се изгражда всяка една статистическа съвкупност, се наричат ***единици (случаи) на наблюдение***. Всеки отделен случай на наблюдение притежава ***признаци на сходство,*** които позволяват включването му в дадена статистическа съвкупност, но едновременно с това има и ***признаци на различие***, които го правят неповторим и индивидуален.

Обединяването на множество единици на наблюдение, с присъщите им елементи на сходство и различие, в една статистическа съвкупност довежда до формиране на нови характеристики, които липсват при отделните случаи и се наричат ***групови свойства на статистическата съвкупност.***

Прилагането на разнообразни статистически методи и показатели (критерии) е насочено именно към измерване и оценка на груповите свойства на изучаваните статистически съвкупности.

Основните групови свойства на всяка статистическа съвкупност са:

– разпределение на изучаваните признаци (променливи величини)

– средно ниво (централна тенденция)

– разнообразие (вариране)

– репрезентативност на признаците

– взаимовръзка между признаците

Разпределение на признаците

Това е едно от най-важните свойства на статистическата съвкупност. Отделните елементи на съвкупността се разпределят нееднакво според стойностите на изучаваните променливи величини и по такъв начин се образува определена вътрешна структура на съвкупността по отношение на всяка променлива. Например, при наблюдение на репрезентативна извадка от 100 живородени момичета са регистрирани стойности на ръста, които варират в границите от 46 см до 54 см, но отделните числени значения на промен-

ливата „ръст“ се проявяват с различна честота в извадката. Както се вижда от ***табл. 4.1,*** тридесет от наблюдаваните 100 живородени момичета имат ръст 50 см, по двадесет са съответно с ръст 49 и 51 см, а останалите 30 случая се разполагат почти симетрично около тези стойности и силно намаляват към крайните стойности.

Табл. 4.1. Стойности на ръста при 100 новородени момичета

Ръст в см – x	Честота f
46	2
47	6
48	7
49	20
50	30
51	20
52	8
53	5
54	2
	$\Sigma f = N = 100$

Видът на разпределението характеризира вътрешната структура на съвкупността по отношение на дадена променлива и очертава в най-общ план закономерностите на изучаваното явление. Разпределението може да се представи също така графично.

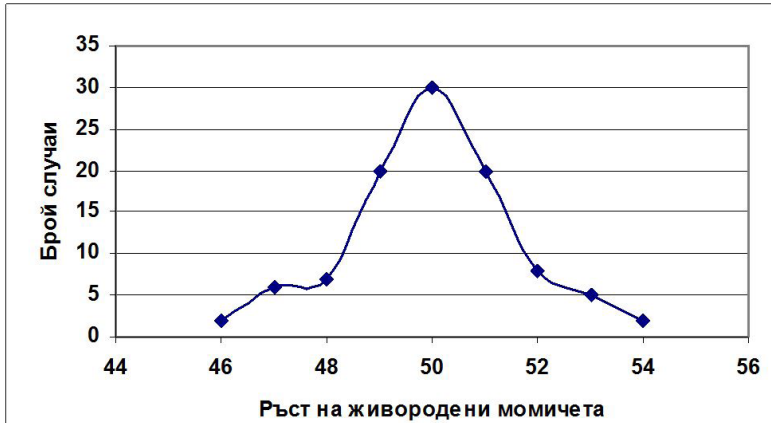
Най-често срещаните форми на разпределение са:

– ***алтернативно разпределение*** – то се наблюдава при качествени признаци с две разновидности;

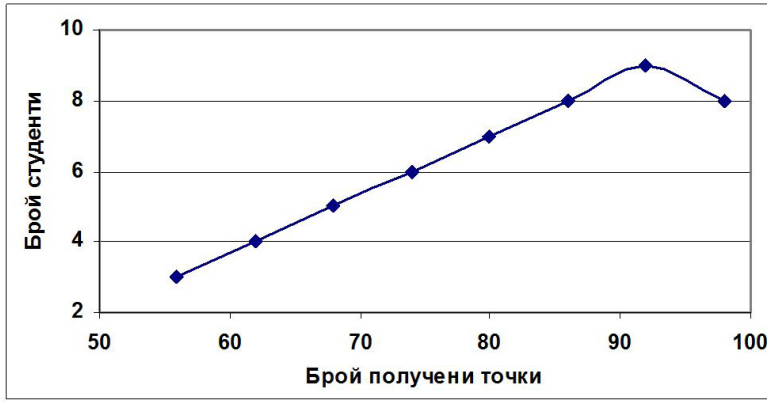
– ***нормално или симетрично разпределение*** (Гаус-Лапласово) – най-честата форма на разпределение на количествени непрекъснати променливи величини. При него случаите с различна стойност на признака се разполагат симетрично и най-голям брой случаи се струпват около средното ниво ***(фиг.4.1)***.

– ***асиметрично разпределение*** – при него най-голям брой случаи се струпват не в средата на реда, а се изместват към по-малките или към по-големите значения на даден изучаван признак. Съответно говорим за дясноизтеглено или лявоизтеглено разпределение. Такива форми на разпределение

са по-чести при дискретните променливи. Например, разпределението на семействата у нас по брой деца е асиметрично – най-голям брой семейства имат 1-2 деца, а броят на семействата с повече от 2 деца рязко намалява. Пример за несиметрично лявоизтеглено (отрицателно) разпределение са резултатите от тест при 50 студенти (***фиг.4.2***).



Фиг. 4.1. Крива на нормално разпределение



Фиг. 4.2. Несиметрично отрицателно изтеглено разпределение

В други ситуации може да се наблюдава струпване на повече случаи в двата края или в други две точки на измерителната скала. Такива разпределения са ***двувърхови или бимодални***. Възможни са и ***полимодални разпределения*** – с повече от 2 върха. Подобни случаи са сериозен сигнал за нееднородност на изучаваната съвкупност, което може да доведе до неверни изводи.

Формата на разпределението е свързана с подбора на най-подходящите статистически методи за обработка на данните. Например, ***параметричните методи се прилагат само при нормално разпределение, а непараметричните методи се използват при всякакви форми на разпределение.***

Средно ниво (централна тенденция)

При болшинството количествени променливи величини, независимо от варирането на стойностите им при отделните наблюдавани случаи, е налице тенденция към формиране на определено средно ниво, т. е. налице е определена ***централна тенденция***. Средното ниво характеризира типичното проявление на количествените променливи, което се формира под влияние на ***определящи, закономерни фактори и причини.*** Това са такива фактори и причини, които се проявяват при всички единици на наблюдение, а не са присъщи само на отделни случаи. Например, средното ниво на показателите за физическо развитие (тегло, ръст и др.) се формира под влияние на такива закономерни фактори като пол, възраст и др. Затова средните нива на тези показатели се различават съществено при лица на различна възраст или от различен пол.

За измерване и характеристика на средното ниво се използват различни видове ***средни величини (средна аритметична, мода, медиана).***

Разнообразие (вариране, разсейване)

Въпреки стремежа на всеки изследовател да постигне еднородност при формирането на изучаваната съвкупност, стойностите на променливите величини при отделните случаи се различават

в по-голяма или по-малка степен. Това се дължи на стремежа на количествените променливи към разнообразие (вариране), което се проявява при всички живи организми и е резултат от влиянието на ***случайни, неопределящи фактори и причини.*** Различните стойности на ръста и теглото на децата, например, се свързват с влиянието на такива фактори като наследственост, недохранване, хронични заболявания и др. Тези фактори не са еднакви при всички деца – при някои те са налице и предизвикват отклонения от типичното ниво на ръста и теглото за съответния пол и възраст, докато при други деца тези фактори отсъстват и стойностите на променливите са много по-близки до средното ниво.

За измерване на варирането се използват ***лимит на вариационния ред, интерквартилен обхват, дисперсия, стандартно отклонение, коефициент на вариация и др.***

Репрезентативност на признаците

Това е едно от най-важните групови свойства и означава ***способността на извадката да отразява свойствата на генералната съвкупност (популацията).*** За обезпечаване на репрезентативността е необходимо да се спазват строго изискванията за непреднамерен подбор на случаите в извадката и обезпечаване на достатъчен обем на извадката. Репрезентативността се измерва т. нар. ***репрезентативна (стандартна) грешка,*** която показва доколко резултатите от наблюдение на извадката се отличават от тези, които бихме получили при наблюдение на популацията, от която е извлечена съответната извадка. Определянето на стандартната грешка е задължителен елемент при формулирането на заключения за параметрите на популацията на базата на данните от наблюдения на извадки.

Взаимовръзка между признаците

Степента на зависимост между променливите, характеризиращи изучаваните явления, се измерва чрез различни показатели, от които най-широко се използват *коефициентите на корелация.*