Глава 5

ОРГАНИЗАЦИЯ И ПРЕДСТАВЯНЕ НА ДАННИ ОТ НАУЧНИ ПРОУЧВАНИЯ

В предишните глави разгледахме въпросите, свързани с планирането на научните проучвания и процедурите за събиране на данни. Обаче, възприемането и осмислянето на суровите данни, състоящи се от голям брой вариращи измервания, може да се окаже извънредно трудно. Ето защо, преди да се пристъпи към обсъждане на резултатите от дадено проучване, необработените данни трябва да бъдат организирани и представени по ясен и интелигентен начин, обезпечаващ стегнат и компактен вид на масива от данни и на неговите основни характеристики. Редуцирането на данните до управляема и лесна за разбиране и осмисляне форма представлява една от най-важните цели на статистиката.

Съществуват различни подходи (ръчни или чрез използване на компютри) за обединяване в едно цяло на данните от различните рутинни или специални проучвания. Начинът за организиране и представяне на данните зависи от техния вид (номинални или ординални, дискретни или непрекъснати) и използваната измерителна скала.

Особено важни са правилната ***групировка и обобщаване*** на данните.

Групировката има за цел да разпредели единиците на изучаваното масово явление в еднородни статистически групи и спомага да се изучат типичните закономерности. Уточняването на разновидностите за извършване на групировката на данните трябва да стане още на етапа на планиране на проучването и зависи от вида на изучаваните променливи и задачите на изследването. Разпределянето на единиците на изучаваното явление според разновидно-

стите на променливите и преброяването на случаите в отделните разновидности и с конкретни стойности се извършва в етапа на обработка на данните.

Обобщаването на данните представлява многоетапна операция, свързана с изчисляване на групови и крайни резултати, като програмата за това се разработва заедно с плана и програмата на проучването. Трябва да се знае какви резултати се очакват на изхода на проучването, за да се съберат необходимите за това данни и то по подходящ начин. Само така може да се обезпечи добър баланс между целта и задачите, програмата на проучването, събирането на данните, методите на тяхната обработка и крайните резултати от проучването.

Обобщаването на данните може да се извърши в:

* ***статистически таблици***
* ***честотни разпределения***

1. Таблично представяне на данните

По същество ***таблицата*** представлява разнообразна по съдържание мрежа от пресичащи се редове и колони, които са озаглавени и запълнени с числа. Изготвянето на всяка таблица трябва да отговаря на редица ***общи изисквания.***

Всяка една правилно изработена таблица следва да съдържа следните формални елементи:

Номер на таблицата – изисква се когато данните са представени в множество таблици. Най-често номерът на таблицата е непосредствено пред заглавието или се разполага на горния ред вдясно или вляво. При анализ на данните авторите цитират номера на таблицата в текста, откъдето се тълкуват данните и се правят съответни изводи.

Заглавие на таблицата – то трябва да отразява в ясна и сбита форма съдържанието на таблицата, т.е. да отразява обекта на изследване, разглежданите характеристики, времето и мястото, измерителните единици (брой, %).

Челна колона – това е първата колона на таблицата, в която обикновено се разполага т. нар. ***статистически подлог***, т. е. това, което ще бъде описано в таблицата (например, видове заболявания, основни причини за умирания и др.).

Челен ред – това е най-горният ред на таблицата, в който обикновено се отразява т.нар. ***статистическо сказуемо***, т.е. признаците, които характеризират статистическия подлог (например, пол и възраст на заболелите или умрелите лица).

Глава на таблицата – това е най-горната лява клетка, в която се пресичат челния ред и челната колона.

Клетки на таблицата – те се получават от пресичането на колоните и редовете. Техният брой зависи от разновидностите или интервалите, чрез които се описват изучаваните признаци.

Сумарен ред – това е най-долният ред на таблицата, в който се отразяват сумите за всяка колона.

Сумарна колона – последната дясна колона на таблицата, в която се отразяват сумите за всеки ред.

Контролна клетка на таблицата – най-долната клетка вдясно, в която се пресичат сумарният ред и сумарната колона. Тя отразява общия брой на изследваните лица, проби и др. и служи за контрол относно правилното попълване на таблицата.

В клетките на таблицата, освен числа, могат да се поставят различни ***условни знаци:***

(.) – липсват данни и не е възможно да се получат въобще
(...) – не са търсени такива данни
(x) – в клетката не е логично да има число поради естеството на данните
(-) – липсват съответни абсолютни числа, въпреки че са търсени
(?) – посочената цифра е под съмнение
(*) – числото в клетката подлежи на уточняване или се отнася за различен период спрямо посочения за останалите клетки
(0) – съответният показател е под ½ от възприетата мерна единица

Видове таблици

Проста (едномерна) таблица – резултатите са групирани само по един признак (напр. разпределение на изследваните лица по професия) или пък представя честотите на една променлива величина (напр. разпределение на група лица по ниво на систолното налягане).

Групова таблица – съдържа повече от един признак, но всеки признак се разглежда поотделно, без да се съчетават помежду си (например, заболяванията могат да се разглеждат отделно по пол и по възраст).

Комбинационна таблица (крос-табулация) – тук е налице съчетание на разглежданите променливи величини.

В зависимост *от броя на крос-табулираните променливи сложните таблици биват:* ***двумерни и многомерни.***

В зависимост от *броя на разновидностите на комбинираните променливи величини* сложните таблици биват:

– четирикратна (четириклетъчна) или таблица **2x2** – когато се съчетават биноминални променливи величини, всяка от които има само по две разновидности (наричани още ***алтернативни признаци***). Тогава са възможни само 4 съчетания на разновидностите на променливите.

– многократна (многоклетъчна) – когато поне един от изучаваните признаци има повече от 2 разновидности и възможните съчетания между разновидностите на признаците са повече от четири.

Всеки вид таблица има своето значение. Простите таблици дават начална представа за основните характеристики на изучаваното явление и подсказват в каква насока би могъл да се провежда последващия анализ.

Най-съвършена е комбинационната таблица, която дава представа за взаимоотношенията между признаците на изучаваното явление, а оттам – и за неговата обусловеност от различни

фактори. Много важно, обаче, е да се избягва прекомерното усложняване и опити за съчетаване на повече от 3-4 признака, тъй като това утежнява анализа. По-добре е да се съставят няколко комбинационни таблици с умело и логично подбрани признаци, което ще позволи да се разкрият по-бързо съществени закономерности.

Начинът на обобщаване на данните е важно изискване за приложението на едни или други статистически методи за обработка на данните (напр., има различни методики за изчисляване на коефициенти на корелация, на критерия χ^2 и др. в зависимост от това дали таблицата е четирикратна или многократна).

2. Честотни разпределения

Честотните разпределения показват начина, по който са разпределени резултатите в дадена извадка или популация.

Пример: На изпит 50 студенти отговарят на тест, съдържащ 50 въпроси от типа Multiple Choice (с множество отговори, от които само един верен) като за всеки верен отговор са присъждани по 2 точки. Резултатите варират от 54 до 100 точки и са оценявани по скала:

под 60 точки – слаб
60-64 точки – среден 3.00
66-70 точки – добър 3.50
72-76 точки – добър 4.00
78-82 точки – много добър 4.50
84-88 точки – много добър 5.00
90-94 точки – отличен 5.50
96-100 – отличен 6.00

Ако срещу поредния номер на всеки изпитан студент посочим неговия резултат (***x***), ще получим ред от 50 числа, които варират в широки граници и не са групирани по някакъв системен начин. Такъв ред се нарича ***прост или непретеглен вариационен ред.*** От него едва ли бихме могли да направим извод доколко добре са се справили студентите с теста.

Сравнително лесен и полезен начин за внасяне на по-голям ред и яснота в представянето на количествени данни, измерени при голям брой случаи (над 20-30) е подреждането на стойностите на измерванията (**x**) във възходящ или низходящ ред, като срещу всяка стойност се посочи колко лица имат такъв резултат, т.е. каква е ***честотата (f – frequency)*** на поява на конкретна стойност на променливата в изучаваната съвкупност (***табл.5.1***).

Табл. 5.1. Честотно разпределение на резултати от тест при 50 студента

x	*f*	x	*f*	x	*f*
54	1	*70*	2	*86*	3
56	1	*72*	2	*88*	3
58	1	*74*	2	*90*	3
60	1	*76*	2	*92*	3
62	1	*78*	2	*94*	3
64	2	*80*	2	*96*	3
66	1	*82*	3	*98*	3
68	2	*84*	2	*100*	2

Този метод за групиране на данните се нарича ***честотно разпределение или претеглен вариационен ред*** (тъй като за всяка стойност на променливата е посочено нейното тегло, т.е. колко пъти се среща).

След групирането на данните се вижда, че само 3 студенти не са покрили минималните изисквания, а 16 студента са се справили отлично. Вместо в 50 реда данните са групирани в 24 реда, което е доста по-лесно за анализ.

Когато срещу всяка измерена стойност на променливата се посочва честотата й, разпределението е ***степенно (степенен вариационен ред).***

По-нататъшната организация на данните може да доведе до обединяване в интервали и превръщане на разпределението в ***интервално (интервален вариационен ред)***, както това е показано в ***табл. 5.2***.

Табл. 5.2. Честотно разпределение на резултати от тест при 50 студента

x (брой точки)	*f*	*x* (брой точки)	*f*
54-58	3	*54-58*	3
60-64	4	*60-64*	4
66-70	5	*66-76*	11
72-76	6	*78-88*	15
78-82	7	*90-100*	17
84-88	8		***N* = 50**
90-94	9		
96-100	8		
	$\Sigma f = N = 50$		

При съставянето на интервално честотно разпределение трябва да се отчитат следните моменти:

1. ***Оптимален брой интервали***. Прието е той да не надвишава 9, за да не се затруднява анализът на данните, но не бива да е твърде малък, за да не се прикрият важни тенденции в данните.

2. Предпочитат се ***интервали с еднаква ширина***, която се избира според изучаваната количествена променлива и опита на изследователя. Когато има затруднения се препоръчва прилагане на ***формулата на Стърджес:***

$$e = (x_{мах} - x_{мин}) : (1 + 3.322.lg\ n),$$

където: ***e*** – ширина на интервала

$x_{мах}$ – максимална стойност на променливата

$x_{мин}$ – минимална стойност на променливата

3.322 – константна величина

n – брой на наблюдаваните случаи в извадката

Ако броят на групите е предварително фиксиран, то ширината на интервала се определя по формулата:

$$e = (x_{мах} - x_{мин}) : к,$$

където ***к*** е броят на групите. Ако се получи дробно число, то се закръглява на цяло число.

3. За предпочитане е ***ширината на интервала да е нечетно число***, което улеснява изчисляването на средните аритметични

и графичното представяне на честотните разпределения, тъй като интервалът се представя чрез неговата среда (при нечетни числа тя е цяло число). Понякога това е невъзможно.

4. ***Границите на интервалите*** трябва да бъдат ***взаимно изключващи се и изчерпателни***, т.е. всеки резултат да попадне само в един интервал. Напр., за възрастта при ширина 5 г. – 0-4, 5-9, 10-14 и т.н.; при ширина 10 г. – 0-9, 10-19, 20-29 и т.н.

5. ***Да се избягват интервали с отворено начало или отворен край,*** но понякога това се налага. Напр., класификацията на СЗО за възраст: млади – до 44 г.; средна възраст – 45-59 г.; възрастни – 60-74 г.; стари хора – 75-89 г.; 90 г. и + – дълголетници.

Честотните разпределения могат да бъдат:

* ***абсолютни***
* ***относителни***
* ***кумулативни***

Абсолютното честотно разпределение е представено с действителния брой случаи, които имат конкретна стойност на променливата и попадат в конкретен интервал (***табл. 5.1 и 5.2***).

При ***относителното разпределение*** действителната честота се превръща в относителна (общия брой случаи се приема за 100 или за 1 (***табл. 5.3***).

Табл. 5.3. Абсолютно, относително и кумулативно разпределение на резултатите от тест при 50 студенти

x	Абс.*f*	Отн.*f* (в %)	Кум. *f* (в %)	*Интервал*	Абс. *f*	Отн. *f* (в %)	Кум. *f* (в %)
54	1	2	-	*54-58*	3	6	-
56	1	2	4	*60-64*	4	8	14
58	1	2	6	*66-70*	5	10	24
60	1	2	8	*72-76*	6	12	36
62	1	2	10	*78-82*	7	14	50
64	2	4	14	*84-88*	8	16	66
66	1	2	16	*90-94*	9	18	84
68	2	4	20	*96-100*	8	16	100
70	2	4	24				
72	2	4	28				

x	Абс.*f*	Отн.*f* (в %)	Кум. *f* (в %)	*Интервал*	Абс. *f*	Отн. *f* (в %)	Кум. *f* (в %)
74	2	4	32				
76	2	4	36	*Интервал*	Абсол. *f*	Относ. *f*	Кумул. *f*
78	2	4	40	*54-60*	3	6	-
80	2	4	44	*60-64*	4	8	14
82	3	6	50	*66-76*	11	22	36
84	2	4	54	*78-88*	15	30	66
86	3	6	60	*90-100*	17	34	100
88	3	6	66				
90	3	6	72				
92	3	6	78				
94	3	6	84				
96	3	6	90				
98	3	6	96				
100	2	4	100				

Кумулативното честотно разпределение се получава чрез прибавяне на относителната честота за даден интервал към тази от предходните интервали (***табл. 5.5***). Това позволява да направим заключение какъв % от всички случаи попадат до съответния интервал.

3. Графично представяне на таблични данни и честотни разпределения

Графичното представяне на данните е неделима част от изследователския процес и с него се постигат ***две основни цели***:

– по-лесно възприемане на резултатите, дори и за неспециалисти;

– по-добро очертаване на закономерностите на изучаваното явление.

Ако използваните средства за онагледяване не отговарят на тези цели, те стават безсмислени, дори може да затруднят интерпретацията на резултатите. Ако се онагледяват данни за няколко групи, между които няма съществено различие, онагледяването

също има по-малка стойност и бихме могли да се задоволим само с текстово описание.

Графичните изображения имат редица ***предимства*** пред таблиците:

1. ***Нагледност*** – добре оформените графически изображения се възприемат зрително по-леко.

2. ***Изразителност*** – статистическите данни, представени графически, се запомнят много по-бързо.

3. ***Атрактивност*** – вниманието на читателя или слушателя може да бъде ангажирано много по-пълно от графически представени данни.

4. ***Логичност*** – вътрешната логика, която е налице при причинно обвързаните явления, проличава изключително добре при тези изображения.

5. ***Експресивност*** – те дават богати възможности за изразяване на известна идея, за защита на определена концепция и за изява на определена теза.

Основните изисквания към графичните изображения включват:

– за всяко изучавано явление и вид данни да бъде подбран ***най-подходящият вид графично изображение***;

– да имат ***ясно формулирани заглавия и легенди;***

– да ***представят предимно основните данни и зависимости***, без излишни детайли;

– да са ***добре изпълнени технически*** с подбор на подходящи щриховки и означения;

– да са ***достатъчно ясни*** и да позволяват бързо възприемане и осмисляне на информацията без допълнителни изчисления.

Видове графични изображения:

♦ ***По измерение***

* плоскостни (едномерни)
* пространствени (двумерни или тримерни)

♦ ***По характер***

* линейни диаграми
* стълбови диаграми (вертикални и хоризонтални)
* кръгово-секторни диаграми
* цилиндрови, конусовидни, пирамидални, радиални
* площи
* ХУ диаграми на разсейването
* фигурни
* картограми и картодиаграми и т.н.

Компютърните технологии днес предоставят изключително богато разнообразие от графични изображения. При ***подбора на вида им*** трябва да се изхожда преди всичко от характера на данните, т.е. дали те са категорийни (номинални, биноминални, ординални) или количествени (прекъснати и непрекъснати) и от начина на тяхното групиране.

Графично представяне на качествени променливи величини

Качествените променливи величини (номинални, биноминални и ординални) се представят чрез два основни вида изображения: ***стълбови диаграми (вертикални и хоризонтални) и кръгово-секторни диаграми.***

Стълбовите диаграми представят абсолютната или относителна честота за всяка разновидност на променливата (***фиг. 5.1 и 5.2***). При използването им следва да се спазват следните общоприети ***правила***:

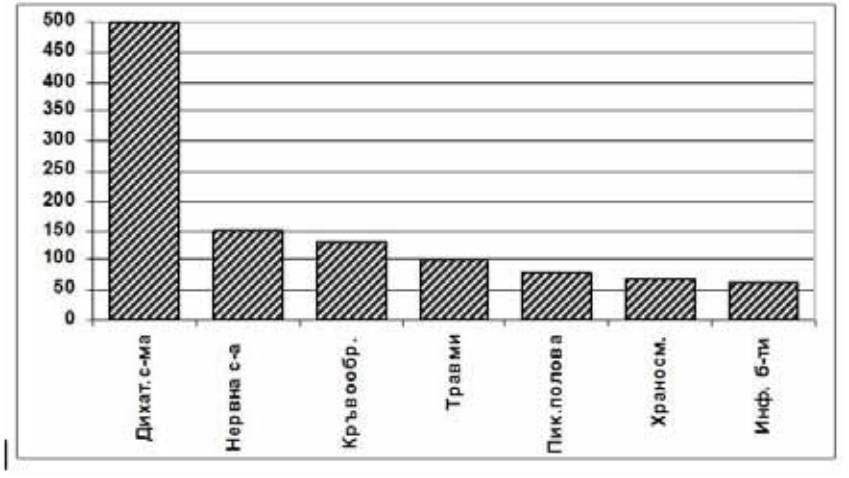
– Оста ***x*** отразява разновидностите на категорийната променлива, а оста ***y*** – честотата на отделните разновидности на променливата.

– Ширината на стълбовете е еднаква, а височината (дължината) им отразява честотата в абсолютни числа или в относителни величини (%, ‰ и др.). Стълбовете или лентите са отделени за разлика от хистограмата.

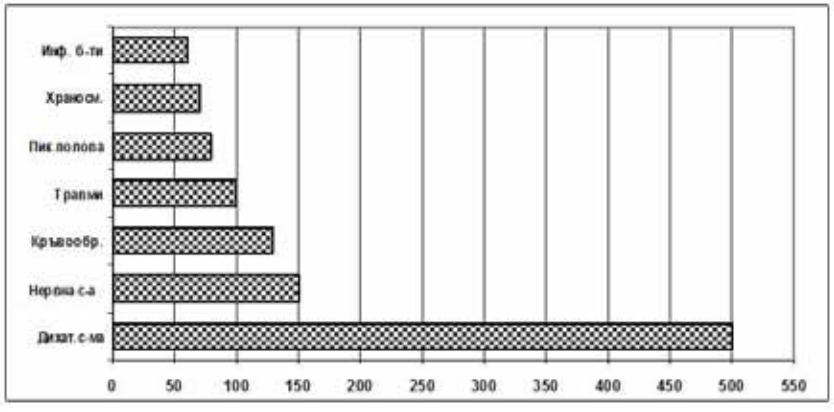
– Оста **y** започва от нула, така че височината на стълбовете (или на лентите при хоризонталните диаграми) да е пропорционална на честотите.

– Цветовете или щрихованието трябва да подчертават закономерностите, да облекчават възприемането и да стимулират естетическо удовлетворение.

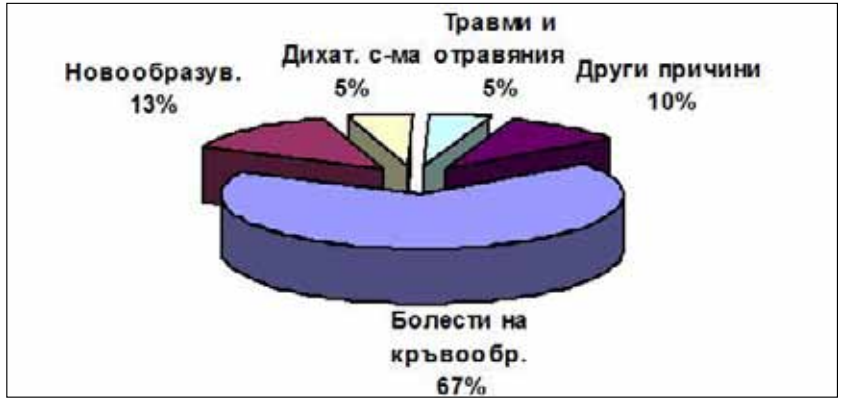
Кръгово-секторните диаграми представляват кръгове, разделени на сектори, пропорционално на абсолютните или относителните честоти на разновидностите на променливата величина (***фиг.5.3***). Всички случаи са 100% (1% = 3.6^0). Понякога за 100% може да се приеме полукръг (1% = 1.8^0).



Фиг. 5.1 Регистрирани заболявания в район А. – 2012 г.



Фиг. 5.2. Регистрирани заболявания в район А. – 2012 г.

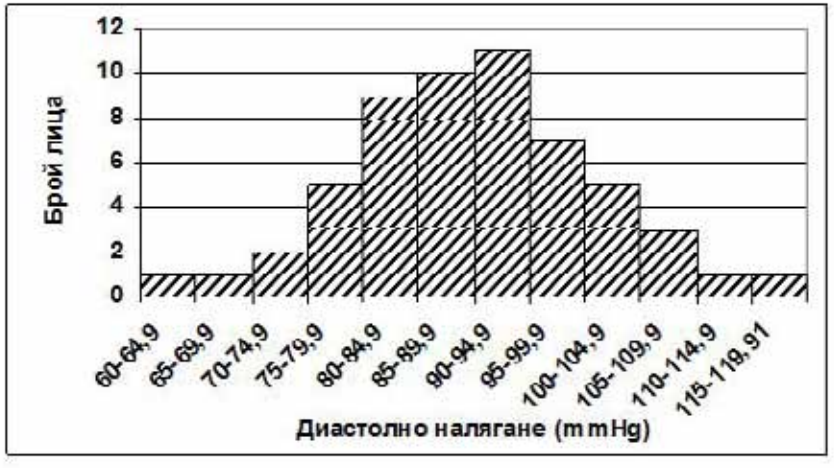


Фиг. 5.3. Водещи причини за умирания в район А. – 2012 г.

Графично представяне на количествени променливи величини

Количествените данни се измерват най-често чрез интервални или пропорционални скали и се представят чрез абсолютни, относителни и кумулативни честотни разпределения, графичният образ на които е най-често *хистограми и честотни полигони* (обикновени и кумулативни).

Хистограмата наподобява стълбова диаграма, но стълбовете са плътно прилепени, за да подчертаят непрекъснатостта на данните. Височината на стълбовете по оста y съответства на честотата в съответните класове и интервали, които са разположени на оста x ***(фиг. 5.4).***



Фиг. 5.4 Хистограма на непрекъсната променлива величина

Пример: При измерване на диастолното налягане в извадка от 56 мъже на възраст 50-59 г., които са силни пушачи (пушат по 2 и повече кутии цигари на ден) са получени следните резултати ***(табл. 5.4):***

Табл. 5.4. Стойности на диастолно налягане при мъже пушачи на възраст 50-59 г.

Диастолно налягане (mmHg)	Абсолютна честота (*f*)	Относителна честота (в %)	Кумулативна честота (в %)
60-64,9	1	1.8	-
65-69,9	1	1.8	3.6
70-74,9	2	3.6	7.2
75-79,9	5	8.9	16.1

Диастолно налягане (mmHg)	Абсолютна честота (*f*)	Относителна честота (в %)	Кумулативна честота (в %)
80-84,9	9	16.1	32.2
85-89,9	10	17.8	50.0
90-94,9	11	19.6	69.6
95-99,9	7	12.5	82.1
100-104,9	5	8.9	91.0
105-109,9	3	5.4	96.4
110-114,9	1	1.8	98.2
115-119,9	1	1.8	100.0
	N = 56	**100.0**	

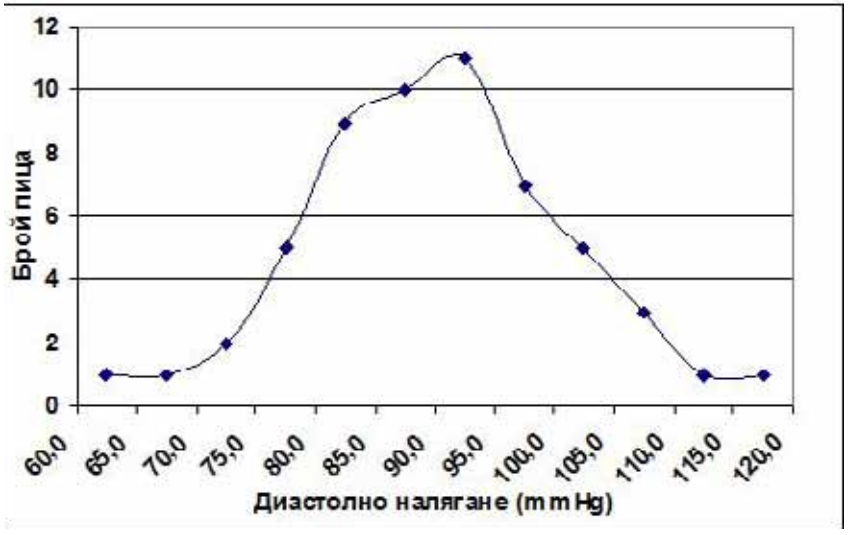
Честотният полигон (линейна диаграма) представлява линия, която свързва срединните точки в горната част на всеки стълб на хистограмата, т.е. отделните точки на линията съответстват на абсолютните честоти за всеки интервал. Честотните полигони имат редица предимства:

– облекчават визуалното сравнение между две или повече разпределения, изобразени на една и съща диаграма;

– позволяват да се интерполират непрекъснати променливи величини;

– позволяват да се оцени честотата на стойностите между отделни точки на полигона.

Честотните полигони могат да приемат различна форма в зависимост от начина на разпределение на стойностите в проучваната съвкупност. Най-честа е симетричната камбановидна крива линия, която характеризира т.нар. *нормално (Гаус-Лапласово) разпределение* и показва, че болшинството от измерените стойности се разполагат около средното ниво и относително малко попадат в двете „опашки“ ***(фиг. 5.5).***



Фиг. 5.5. Честотен полигон на непрекъсната променлива величина

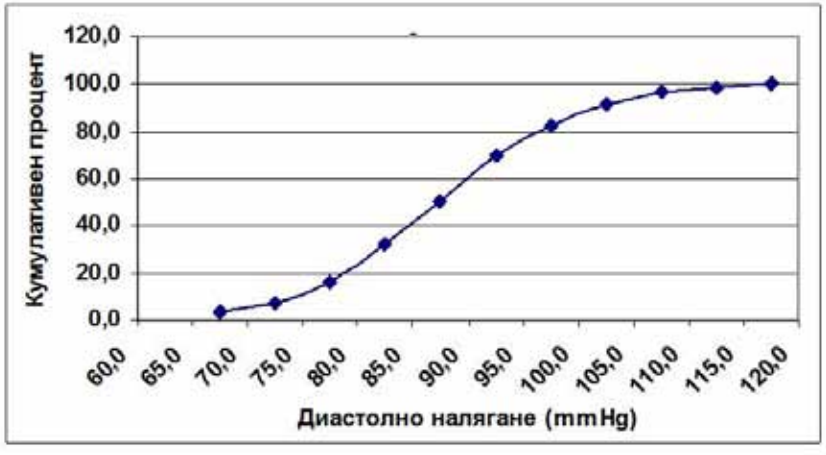
Несиметричните разпределения са по-рядко срещани. Те се характеризират с това, че имат голям брой стойности, разположени в една от страните и изтеглена в противоположна страна „опашка“ с малък брой случаи. Посоката на „опашката“ дава името на съответното разпределение.

При групиране на болшинството стойности към по-високите и изтегляне на „опашката“ към по-ниски стойности разпределението се нарича *отрицателно изтеглено или лява асиметрия*. Такъв вид има разпределението на резултатите от теста при 50 студента ***(табл. 5.3).***

Обратно, когато повечето са ниски и неголям брой стойности са високи, опашката е изтеглена към високите стойности и такова разпределение се нарича ***положително изтеглено или дясна асиметрия.***

Кумулативният полигон представлява линия, отразяваща кумулативно честотно разпределение ***(фиг. 5.6).*** Може да се използва за оценка на честотата на поява на стойност, която е по-малка или

равна на конкретна стойност на променливата (например, процент на мъжете с диастолно налягане по-малко или равно на 90 mmHg).



Фиг. 5.6. Кумулативен полигон на непрекъсната променлива

Заключение

Табличното представяне очертава много по-ясно характеристиките на изучаваните променливи, отколкото необработените данни. Графичното изобразяване, от своя страна, предоставя по-добро визуално възприемане на характеристиките на променливите величини, отколкото табличното представяне. Основният недостатък е в това, че в таблиците и графиките се представят обобщени данни и се губи индивидуалността на информацията.

При ***подбора на подходящи методи*** за представяне на данните трябва да се отчита ***конкретната ситуация и характера на данните***.