Глава 9

Метод на извадките

9.1 Увод

Терминът **„статистика"** произхожда от латинската дума „статус" (status) – състояние. Първоначално, в 17–18 в., когато статистиката започва да се оформя като научна дисциплина, терминът статистика се е свързвал със система за събиране и описване на данни и факти, характеризиращи състоянието на държавата. Събиране на статистически данни, касаещи основно населението на страната, са правени доста отдавна: има сведения, че в 2238 г. до н.е. в Китай при императора Яо е било проведено преброяване на населението; такива преброявания са правени и в древния Египет, в древна Персия, в Римската империя.

Постепенно статистическите методи на изследване се усъвършенстват и навлизат във все повече области на науката и практиката и понастоящем те намират приложение практически навсякъде. Широкото разпространение на статистическите методи се обяснява с това, че статистическите проучвания в различни области на живота имат сходни предмет, цел и задачи на изследване. Предмет на статистическо проучване са големи съвкупности от обекти (предмети, явления, индивиди), които се характеризират с нееднородност по някакъв признак. Целта на статистическото проучване е да се открият общи характеристики, тенденции и закономерности в изследваните съвкупности и да се направят практически изводи. Основна задача на статистическото проучване е да се намери разпределението на елементите на изследваната съвкупност по избрани признаци. Например търси се разпределението на населението по възраст или разпределението на семействата по доходи.

Получените статистически данни придобиват значение, ако бъдат подложени на анализ, в резултат на който се направят практически изводи. Така например демографски статистически данни могат да бъдат използвани както в застрахователното дело, така и за съставяне на прогнози за демографски промени в населението.

Статистиката се състои от три раздела:

1. Събиране на статистически данни, т.е., данни, характеризиращи отделни единици на някакви масови съвкупности;

2. Статистическо изследване на получените данни, заключаващо се в изясняване на тези закономерности, които могат да бъдат установени въз основа на данните от масовото наблюдение;

3. Разработване на методи за статистическо наблюдение и анализ на статистическите данни. Последният раздел съставя съдържанието на математическата статистика.

Статистическите данни биват **числени** и **нечислени**. Числените статистически данни могат да бъдат числа или набор от числа (вектори), или функции, които характеризират признака, по който се различават обектите, подложени на статистическо проучване. Такива данни имат широко разпространение. Методите за събиране, обработка и анализ на тези данни са добре известни и утвърдени в статистическата практика. В настоящия курс ще боравим основно с числени статистически данни.

Статистическите методи са методи за анализ на статистически данни. Теорията на статистическите методи е насочена за решаване на конкретни задачи. Понеже постоянно възникват нови постановки на математически задачи за анализ на статистически данни, то постоянно се развиват и обосновават нови методи за анализ. Тяхното математическа обосновка се извършва основно със средствата на теорията на вероятностите.

Статистическите методи биват:

- методи с общо предназначение (приложната статистика), които могат да се прилагат във всички области на научните изследвания ;

- специфични методи, прилагането на които е ограничено в една или друга сфера на човешката дейност.

Приложната статистика намира място в почти всички области на живота:

• икономика (икономически прогнози);

• социология (проучване на общественото мнение и прогнози);

• банково и застрахователно дело (проучване на риска, прогнози);

• маркетинг (проучване на пазара и потреблението);

• демография (раждаемост, смъртност, демографски прогнози);

• медицина (прогнози за заболеваемост, ефективност на ново лекарство, откриване на фактори, които способстват за развитието на дадена болест);

• природни науки като физика, химия, биология (планиране на експерименти, откриване на зависимости и закономерности);

• техника и промишленост (контрол на качеството и на производството);

• селско стопанство (сравняване на качествата на различни сортове и породи, прогнози за добивите);

• метеорология и екология (климатични и екологични прогнози).

При използването на статистическите методи в конкретни области на науката и практиката са се обособили научно-приложни дисциплини от типа „статистически методи в икономиката", „статистически методи в медицината", „статистически методи в социологията". От математическа гледна точка тези дисциплини прилагат едни и същи статистически методи, като различията са само в тълкуването на вероятностно-статистическите модели, построени в съответствие с особеностите на конкретната област на приложение.

9.1.1 Генерална съвкупност и генерален признак

Статистическите изследвания се извършват, когато трябва да се намери разпределението на голяма съвкупност от сходни обекти (предмети, явления, индивиди) по някакъв признак като например:

• Разпределение на населението по възраст;

• Разпределение на семействата по доходи;

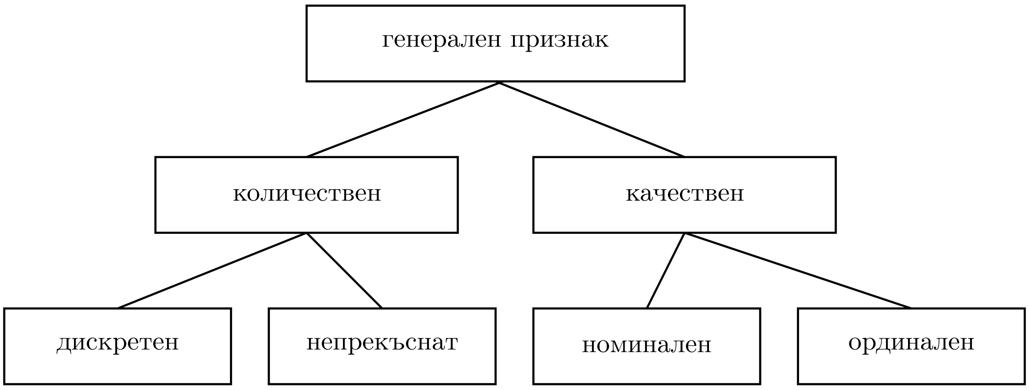
• Разпределение на мъжете (жените, децата) в дадена страна по ръст, размери на ходилото, гръдна обиколка и др.;

• Разпределение на болни от дадена болест според реакцията им на дадено лекарство;

• Разпределение на продукцията по качество;

• Разпределение на телевизионните зрители според оценките им за дадено предаване.

Множеството Ω от изследвани обекти се нарича **генерална съвкупност**, а броят N на тези обекти – **обем на генералната съвкупност**. Самите признаци, с които се характеризират обектите на една генерална съвкупност, се наричат **генерални признаци**. Те могат да бъдат **количествени (числени)** и **качествени (категориални, нечислени)**. Всеки количествен признак се характеризира с някаква числена стойност. Количествените признаци биват **дискретни** и **непрекъснати (интервални)**. Дискретни са онези от числените признаци, които могат да приемат краен брой или изброимо множество от възможни числени стойности. Например признаците „възраст в години", „брой на отказите на прибор", „брой на пасажерите" са дискретни. Непрекъснати са онези от числените признаци, които могат да приемат за стойност всички числа от даден интервал – например признака „височина", „тегло" и т.н. Ще отбележим, че числената стойност, която приема даден количествен признак има съдържателна стойност, ако е уточнено в каква мерна система и избрана мерна единица става нейното определяне. Например признакът „ръст" ще приема различни стойности, ако измерването е в „сантиметри", „метри" или „инчове". От своя страна, качествените признаци не могат да се характеризират числено. Те биват **номинални** и **ординални**. Номинални са онези от качествени признаци, които не подлежат на количествено сравнение. Например признакът „цвят на косите" приема следните, лишени от количествен смисъл, стойности: „кафяви", „руси", „рижи", „черни", „сиви". Ординални са онези качествени признаци, които подлежат на количествено сравнение. Например оценката за дадено телевизионно предаване е субективна и може да се характеризира със следните ординални стойности: „лошо", „средно", „добро", „много добро", „отлично", „възхитително" и др. Такива оценки може да се степенуват (ранжират), като на всяка от тях се съпостави число по някаква скала (от 1 до 6, от 1 до 10, от 1 до 100). Всяко едно такова съпоставяне, обаче, е условно. Цялостна класификация на различните генерални признаци е дадена на следната фигура.

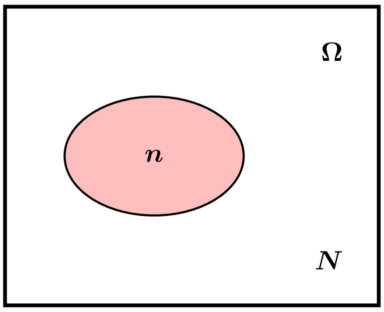


9.2 Извадка

Изчерпателно описание на генерална съвкупност по даден признак е възможно, ако се изяснят стойностите на признака за всички представители на генералната съвкупност без изключение. Например това е възможно при пълно преброяване на населението. Такова подробно изучаване обаче не винаги е възможно. Някои пречки за това са следните:

- Обемът на генералната съвкупност е голям, което води до оскъпяване на изследването;
- Генералната съвкупност е въображаема;
- Пълното изучаване води до унищожаване на генералната съвкупност и това обезсмисля изследването. Така например до унищожаване на генералната съвкупност би се стигнало, ако с пълно изучаване искаме да определим:
  – процента на годност на взриватели от дадена партида;
  – средното време на безотказна работа на прибори;
  – средния срок на годност на лекарства, хранителни продукти и др.

Затова от генералната съвкупност се избират краен брой представители (подбира се **извадка**), прави се изследване на разпределението на признака в извадката и по получените резултати се прави заключение за разпределението на признака в генералната съвкупност. Броят n на елементите в извадката се нарича **обем на извадката**.



Фиг. 9.1: Извадка с обем n

Удобно е данните от направените наблюдения за един или група признаци, с които се характеризират елементите на дадена генерална съвкупност, да се представят с помощта на таблица със следната структура: в първите два стълба на тази таблица се отбелязват номера и обекта на направеното наблюдение, а в следващите – стойностите на изследваните признаци (променливи) за съответния наблюдаван обект.
Както се вижда от дадената таблица, статистическото изследване може да обхваща както един, така и няколко от признаците на елементите на генералната съвкупност.

Наблюдения		Променливи				
№	Служител	Пол	Длъжност	Възраст	Стаж	Заплата
1	А.Б.В.	мъж	Директор	37	15	1860
2	Б.В.Г.	жена	Зам. директор	48	25	960
3	В.Г.Д.	мъж	Гл. инженер	45	20	675
4	Г.Д.Е.	жена	Счетоводител	42	19	720
5	Д.Е.Ж.	жена	Касиер	53	32	460
⋮	⋮	⋮	⋮	⋮	⋮	⋮
16	П.Р.С.	мъж	Бригадир	47	27	580
⋮	⋮	⋮	⋮	⋮	⋮	⋮
		номин.	ординални	дискр.	дискр.	непрек.

9.3 Способи за подбор на извадката

Способите за подбор на извадката се подразделят на два класа:

1. Подбори без разбиване на генералната съвкупност на части.
Това са случаите на:
– прост случаен подбор без връщане,
– прост случаен подбор с връщане.

2. Подбори с разбиване на генералната съвкупност на части.
Това са случаите на :
– типичен подбор

– механичен подбор,

– сериен подбор.

При подбор без разбиване на генералната съвкупност обектите на генералната съвкупност с обем N първо трябва да се номерират от 1 до N. След това от численото множество $\{1, 2, \ldots, N\}$ се прави случаен избор на n числа. Тези числа определят номерата на обектите от генералната съвкупност, които образуват нужната извадка с обем n.

От подборите с разбиване на генералната съвкупност на части заслужава да обърнем внимание на типичния подбор, което ще направим малко по-долу.

Изисквания към извадката. За да получим достоверни изводи от изследването, извадката трябва да отразява основните черти на генералната съвкупност. Затова при формирането на извадката трябва да се спазят следните изисквания:

Извадката не трябва да е пристрастна (преднамерена). Така например няма да получим реална представа за ръста и теглото на учениците в страната, ако проведем проучване за тези показатели в едно спортно училище. Отстраняването на пристрастността на избора, доколкото това е възможно, се извършва със случайна извадка.

Извадката трябва да е случайна, т.е. елементите на извадката трябва да се избират случайно и по такъв начин, че всички елементи на генералната съвкупност да имат равни шансове да попаднат в нея. Получаването на случайна извадка се извършва със статистическата **техника на случайния избор (процедура на рандомизация)**, като в много от случаите е полезно да се използват **таблици на случайните числа** [19], [18]. Ще отбележим, че напоследък, при получаване на случайна извадка, вместо таблици на случайни числа се използват т.нар. **генератори на случайни числа**, които са включени в много пакети от приложни програми, прилагани в съвременните персонални компютри.

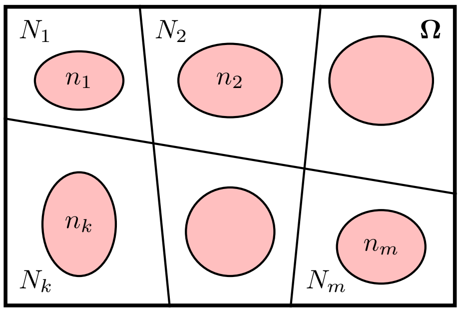
В случая, когато елементите на изследваната генерална съвкупност са номерирани, освен чисто случайна извадка, от въпросната генерална съвкупност може да се реализира т.нар. **„систематична извадка”**. За целта се избира случайно един произволен елемент на генералната съвкупност и дадено цяло число K, наречено „стъпка” на систематичната извадка. След това, започвайки от този случайно избран елемент се избират всички елементи на генералната съвкупност, които отстоят на цяло число стъпки от него. Лесно се съобразява, че когато от генерална съвкупност с N елемента трябва да се избере систематична извадка с обем n, стъпката K е равна на $[N/n]$−цялата част на частното N/n.

Пример 9.1. *Да се реализира систематична извадка с обем* 5 *от генерална съвкупност с обем* 16. *В този случай стъпката е цялата част на частното* $16/5 = 3,33\ldots$, *т.е.* $K = 3$. *Нека случайно избраният елемент е* 8. *Тогава елементите на систематичната извадка с начален елемент* 8 *и стъпка* 3 *ще са с номера:* $2, 5, 8, 11, 14$.

Обемът на извадката трябва да е достатъчно голям. При малък обем на извадката не може да се получи реална представа за свойствата на генералната съвкупност. Затова е желателно този обем да бъде по възможност по-голям. От друга страна, прекомерното увеличаване на обема на извадката не винаги е възможно или пък е нежелателно (поради оскъпяване на изследването или унищожаване на извадката при изследването). Изборът на обема n на извадката е разумен компромис между горните противоречиви изисквания и зависи от естеството на конкретното статистическо изследване.

Извадката трябва да е представителна (репрезентативна), т.е. изборът на представителите да се извършва така, че да отразяват структурата на генералната съвкупност. Съгласно закона за големите числа може да се твърди, че извадката ще бъде представителна, ако е случайна и нейният обем е достатъчно голям. Както видяхме по-горе, това не винаги е възможно при подбор без разбиване на генералната съвкупност на части. Затова в някои случаи е целесъобразно извадката да се извлече с **типичен подбор**, при който обектите се извличат не от цялата генерална съвкупност, а от всяка нейна „типична” част. Например ако лекарство се произвежда от две поточни линии, то подборът се прави не от цялата съвкупност от произведени лекарства, а от продукциите на всяка поточна линия поотделно. Типичният подбор се прилага, когато изследваният признак се колебае забележимо в различните „типични” части на генералната съвкупност. За да получим типичен подбор на извадката, първо разбиваме генералната съвкупност на краен брой подмножества, които са относително еднородни по отношение на изследвания признак. Нека обемите на тези подмножества са съответно $N_1, N_2, \ldots, N_m$, където $N_1 + N_2 + \cdots + N_m = N$. Нека от всяко такова подмножество по случаен начин са подбрани извадки с обеми съответно $n_1, n_2, \ldots, n_m$ и $n_1 + n_2 + \cdots + n_m = n$ (фиг. 9.2). Тогава подборът на обединената извадка с обем n ще бъде типичен, ако са изпълнени условията:

$$\frac{n_1}{N_1} \approx \frac{n_2}{N_2} \approx \cdots \approx \frac{n_m}{N_m}\,.$$



Фиг. 9.2: Типичен подбор на извадката

Като класически пример за ползата от използването на представителна извадка в социологично проучване може да се посочат прогнозите за спечелване на президентските избори в САЩ през 1936 г. с кандидати Ландън (Alf Landon) и Рузвелт (Franklin Roosevelt). Тогава списанието „Литературен преглед” (Literary Digest) провежда анкетно проучване с 4 милиона избиратели. Въз основа на върнатите с отговори над 2 милиона анкетни карти списанието предвижда, че убедителен победител в изборите ще бъде Ландън. В същото време, Галъп (George Gallup) предсказва, че победител в изборите ще бъде Рузвелт, като се базира на анкета, проведена само с 5000 избиратели. Изборите са спечелени убедително от Рузвелт. Причината за провала на прогнозата на „Литературен преглед” е, че анкетираните са били подбрани от телефонни указатели, а телефонът по онова време не е бил разпространен равномерно сред населението. Поради това от проучването са били изключени значими по обем социални групи на по-бедни потенциалните избиратели и проучването е дало изкривена представа за предизборните нагласи (въпреки големия обем на извадката!). Галъп отчита факта, че различните социални групи имат различни предизборни нагласи и като използват статистически данни за числеността на различните социални групи, организира своето анкетно проучване така, че относително малката извадка ($n = 5000$) да бъде представителна. Това обяснява успеха на неговата прогноза.

9.3.1 Първична обработка на данните

Когато обработваме числени статистически данни, всеки признак допуска количествена оценка. Затова вместо да говорим за разпределение на обекти по даден признак, можем да говорим за разпределение на някаква случайна величина X. Опитът, с който е свързана с.в. X, се състои в случаен избор на представител на генералната съвкупност, а стойността на X се определя от стойността на признака за този представител.

Предполагаме, че се изучава разпределението на с.в. X. Правят се n независими опита (наблюдения, измервания), при които с.в. X приема стойности

$$x_1, x_2, \ldots, x_n, \tag{9.1}$$

които образуват извадка с обем n (**статистически ред**).

Когато статистическите пресмятания с данните от реда (9.1) няма да се извършват с компютър, е полезно тези данни да бъдат подложени на предварителна обработка.

Първият етап от обработката на реда (9.1) е получаването на т.нар. **вариационен ред**, което се състои в подреждането на числата от (9.1) в нарастващ ред:

$$x^{(1)}, x^{(2)}, \ldots, x^{(n)} \quad (x^{(1)} \le x^{(2)} \le \cdots \le x^{(n)}), \tag{9.2}$$

където $x^{(k)}, \quad k = 1, 2, \ldots, n$ се нарича $k-$**та варианта** на вариационния ред (9.2). От реда (9.2) се определят минималната и максималната наблюдавани стойности

$$x_{\min} = \min\{x_1, x_2, \ldots, x_n\} = \min\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\} = x^{(1)},$$

$$x_{\max} = \max\{x_1, x_2, \ldots, x_n\} = \max\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\} = x^{(n)}$$

и размахът

$$R = x_{\max} - x_{\min}.$$

Определя се и следната **емпирична функция на разпределение**

$$F_n(x) = \frac{n_x}{n}, \tag{9.3}$$

където n_x е броят на вариантите, които са по-малки от x, а n е обемът на извадката. Емпиричната функция на разпределение $F_n(x)$ е аналог на функцията на разпределение $F(x)$ на разглежданата с.в. X. Разликата между тях е, че докато функцията на разпределение $F(x)$ определя вероятността на събитието $\{X < x\}$, емпиричната функция на разпределение $F_n(x)$ определя честотата на това събитие.

Вторият етап от обработката е групирането на данните.

9.4 Честотно групиране на данните

Честотното групиране на данните от една извадка може да се провежда в два различни варианта:

9.4.1 Честотно групиране на данните чрез указване на техните различните стойности и на честотите (дяловете) на тези стойности в реда от данни

Този метод на групиране е подходящ за дискретни числени променливи с малък брой различни стойности и за категориални променливи.

Групирането става, като се определят различните стойности, които приема с.в. X в реда (9.2)(или (9.1))

$$y_1, y_2, \dots, y_m \quad (y_1 < y_2 < \dots < y_m) \tag{9.4}$$

и също така се определят **честотите (абсолютните честоти)**

$$n_1, n_2, \dots, n_m,$$

където честотата n_k показва колко пъти y_k се среща в извадката и

$$n_1 + n_2 + \dots + n_m = n\,. \tag{9.5}$$

Тогава съставяме следната **таблица на честотите**

y_k	y_1	y_2	$\dots$	$\dots$	y_m
n_k	n_1	n_2	$\dots$	$\dots$	n_m
w_k	w_1	w_2	$\dots$	$\dots$	w_m

Тук $w_k = \dfrac{n_k}{n}$ е **относителната честота** на опитите, при които $X = y_k$.

Забележка 9.1. *От определението за статистическа вероятност следва, че* $w_k \approx P(X = y_k)$. *Освен това, от* (9.5) *следва, че*

$$w_1 + w_2 + \cdots + w_m = 1\,. \tag{9.6}$$

Пример 9.2. *От дневника са събрани данни (статистически ред) за броя на отсъствията на* 25 *ученици от един клас:*

$$2, 5, 0, 1, 6, 3, 0, 1, 5, 4, 0, 3, 3, 2, 1, 4, 0, 0, 2, 3, 6, 0, 3, 0, 1.$$

Вариационният ред се състои от числата

$$0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 5, 5, 6, 6.$$

Различните стойности са

$$0, 1, 2, 3, 4, 5, 6,$$

които се срещат съответно с честоти

$$7, 4, 3, 5, 2, 2, 2.$$

Тогава таблицата на честотите има вида

y_k	0	1	2	3	4	5	6
n_k	7	4	3	5	2	2	2
w_k	7/25	4/25	3/25	5/25	2/25	2/25	2/25

Понякога за нагледност се построява начупената линия, съединяваща точките (y_k, n_k), $\quad k = 1, 2, \ldots, m$, която се нарича **полигон на честотите** или начупената линия, съединяваща точките (y_k, w_k), $\quad k = 1, 2, \ldots, m$, която се нарича **полигон на относителните честоти**.

9.4.2 Честотно групиране на данните чрез разбиване на интервала между най-малката и най-голямата стойност в извадката на определен брой подинтервали

Този метод за интервално групиране е подходящ за непрекъснати числени променливи и за дискретни числени променливи с голям брой различни стойности.

Нека с.в. X е непрекъсната. Тогава в реда (9.1), като правило, няма повтарящи се стойности и във вариационния ред не се получава естественото групиране на данните, характерно за дискретните случайни величини. При голям обем на извадката пресмятанията с негрупирани данни е затруднително, ако не се използва компютър. За да групираме данните от реда (9.1), постъпваме по следния начин:

- Определяме минималната и максималната стойности $x_{\min}$ и $x_{\max}$ на числата от реда (9.1);

- Избираме броя m на интервалите, който препоръчително се определя по формулата

$$m = 1 + 3.32 \cdot \log_{10}(n)\,. \tag{9.7}$$

  или от следната таблица

n	15–24	25–44	45–89	90–179	180–359	360–719	720–1439
m	5	6	7	8	9	10	11

Таблица 9.1: Препоръчителен брой на интервалите

- Разбиваме интервала $[x_{\min}, x_{\max}]$ на m непресичащи се интервала $I_1 = [c_0, c_1), I_2 = [c_1, c_2), \ldots, I_m = [c_{m-1}, c_m]$ посредством избора на точките c_k: $c_0 \le x_{\min} < c_1 < c_2 < \cdots < c_{m-1} < x_{\max} \le c_m$.
  Ако искаме интервалите да са с равни дължини d, то избираме

$$d = \frac{c_m - c_0}{m} \approx \frac{x_{\max} - x_{\min}}{m} = \frac{R}{m}\,. \tag{9.8}$$

- Определяме честотите n_k, показващи какъв е броят на стойностите от реда (9.1), които попадат в интервала I_k.

- Определяме **относителните честоти** $w_k = n_k/n$, всяка от които приблизително е равна на вероятността с.в. X да попадне в интервала I_k: $w_k \approx P(X \in I_k)$.

Тогава съставяме следната **интервална таблица на честотите**

I_k	$[c_0, c_1)$	$[c_1, c_2)$	$\ldots$	$\ldots$	$[c_{m-1}, c_m]$
n_k	n_1	n_2	$\ldots$	$\ldots$	n_m
w_k	w_1	w_2	$\ldots$	$\ldots$	w_m

Забележки:

- Относителните честоти, изразени в проценти, се наричат **процентни честоти**.
- Може да възникне въпросът: към кой интервал да причислим стойността x_j от реда (9.1), ако тя съвпада с граничната точка на два съседни интервала, т.е. ако $x_j = c_k$ за някое k? За да няма объркване, предварително се приема да присъединяваме такива стойности или само към левия интервал, или само към десния интервал. Ние приемаме да присъединяваме такива точки към десния интервал.
- Препоръчва се интервалите да имат равни дължини, но не е задължително.
- Съседни интервали с малки честоти може да се обединяват.
- Препоръчва се броят на интервалите да е в границите от 6 до 20.
- Таблиците с честотите може да се разполагат вертикално.

Понякога за нагледност по интервалната таблица на честотите се построява хистограма на честотите или хистограма на относителните честоти.

Хистограмата на честотите представлява стъпаловидна линия; основата на $k-$то стъпало е интервалът I_k с дължина $d_k = c_k - c_{k-1}$, а височината му е равна на $h_k = n_k/d_k$ и се нарича **плътност на честотата**. Тогава лицето на $k-$то стъпало е равно на n_k, а сумата от лицата е равна на n.

Хистограмата на относителните честоти представлява стъпаловидна линия; основата на $k-$то стъпало е интервалът I_k с дължина $d_k = c_k - c_{k-1}$, а височината му е равна на $h_k = w_k/d_k$ и се нарича **плътност на относителната честота**. Тогава лицето на $k-$то стъпало е равно на w_k, а сумата от лицата е равна на единица.

Хистограмата на относителните честоти е аналог на плътността на разпределение на разглежданата (непрекъсната) с.в. X.

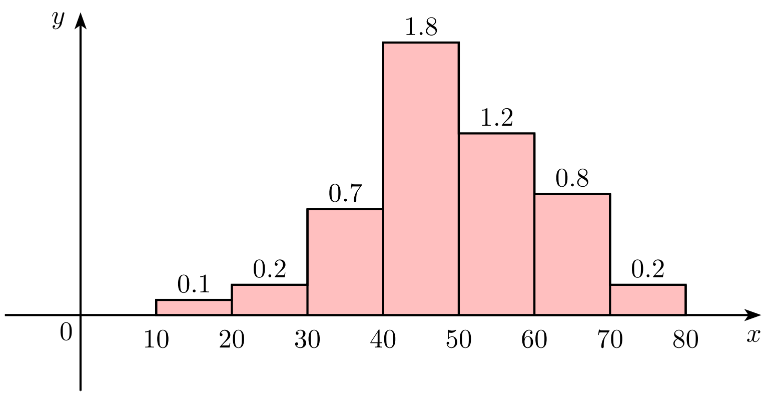
Пример 9.3. *Даден е вариационният ред с обем* $n = 50$*:*

$$\begin{gathered}12, 23, 26, 31, 33, 35, 37, 38, 38, 39,\\ 41, 41, 42, 42, 42, 43, 43, 44, 44, 44,\\ 44, 45, 45, 45, 46, 47, 47, 48, 50, 50,\\ 51, 51, 51, 53, 53, 54, 55, 56, 58, 59,\\ 61, 62, 64, 65, 65, 67, 68, 69, 73, 78.\end{gathered}$$

Понеже $x_{\min} = 12, x_{\max} = 78$*, то размахът на реда е* $R = 78 - 12 = 66$. *Според таблица* 9.1 *при* $n = 50$ *е препоръчително да разбием интервала* $[x_{\min}, x_{\max}]$ *на* $m = 7$ *подинтервала. От* (9.8) *имаме, че тяхната дължина приблизително е равна на* $d \approx 66/7 \approx 9.43$. *Ако изберем* $c_0 = 10$ *и* $c_7 = 80$*, то* $d = 10$ *и останалите точки на разделяне ще бъдат* $c_1 = 20$, $c_2 = 30$, $c_3 = 40$, $c_4 = 50$, $c_5 = 60$, $c_6 = 70$*. Тогава интервалната таблица на честотите ще има вида*

I_k	$10 - 20$	$20 - 30$	$30 - 40$	$40 - 50$	$50 - 60$	$60 - 70$	$70 - 80$
n_k	1	2	7	18	12	8	2
w_k	0.02	0.04	0.14	0.36	0.24	0.16	0.04

Хистограмата на честотите, построена по тази таблица, е дадена на фиг. 9.3.



Фиг. 9.3: Хистограма от пример 9.3

9.5 Кумулативни разпределения

Разгледаните по-горе честоти, относителни честоти и процентни честоти се наричат **некумулативни разпределения**. В статистиката се използват още следните **кумулативни разпределения**:

- **кумулативни честоти**, равни на
  $N_k = n_1 + \cdots + n_k, \quad k = 1, \ldots, m;$
- **кумулативни относителни честоти**, равни на
  $W_k = w_1 + \cdots + w_k, \quad k = 1, \ldots, m; \; W_0 = 0;$
- **кумулативни относителни процентни честоти**, равни на кумулативните относителни честоти, изразени в проценти.

Следващата таблица показва каква е връзката между тези честоти.

k	n_k	N_k	w_k	W_k	$w_k, \%$	$W_k, \%$
1	2	2	0.10	0.10	10	10
2	4	6	0.20	0.30	20	30
3	6	12	0.30	0.60	30	60
4	5	17	0.25	0.85	25	85
5	3	20	0.15	1.00	15	100

Емпиричната функция на разпределение на дискретна случайна величина се изразява с кумулативните относителни честоти по следния начин:

$$F_n(x) = \begin{cases} 0\,, & x \le y_1; \\ W_k\,, & y_k < x \le y_{k+1}, \quad k = 1, \ldots, m-1; \\ 1\,, & x > y_m. \end{cases}$$

Кумулативна крива (полигон на кумулативните относителни честоти) на непрекъсната случайна величина се нарича графиката на функцията

$$\overline{F}_n(x) = \begin{cases} 0\,, & x \le c_0; \\ W_{k-1} + \dfrac{w_k(x - c_{k-1})}{c_k - c_{k-1}}\,, & c_{k-1} < x \le c_k, \quad k = 1, \ldots, m; \\ 1\,, & x > c_m. \end{cases}$$

Тази функция е аналог на функцията на разпределение на разглежданата непрекъсната с.в. X. Нейната графика в интервала $[c_0, c_m]$ е начупена линия, която се получава, като последователно свържем с отсечки точките $(c_0, 0), (c_1, W_1), (c_2, W_2), \ldots, (c_{m-1}, W_{m-1}), (c_m, 1)$.

Таблицата на честотите от пример 9.2 заедно с кумулативните относителни честоти е дадена в таблица 9.2, а графиката на съответната емпиричната функция на разпределение $F_n(x)$ е показана на фиг. 9.4.

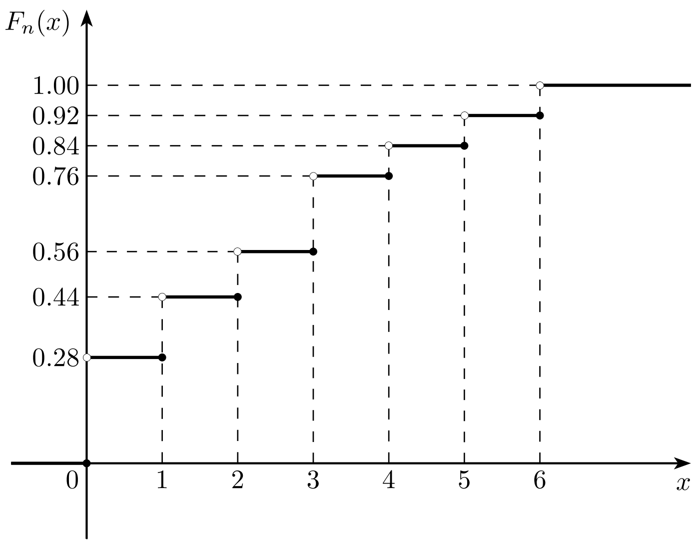
Интервалната таблица на честотите от пример 9.3 заедно с кумулативните относителни честоти е дадена в таблица 9.3, а графиката на съответната кумулативната крива $\overline{F}_n(x)$ е показана на фиг. 9.5.

k	y_k	n_k	w_k	W_k
1	0	7	0.28	0.28
2	1	4	0.16	0.44
3	2	3	0.12	0.56
4	3	5	0.20	0.76
5	4	2	0.08	0.84
6	5	2	0.08	0.92
7	6	2	0.08	1.00

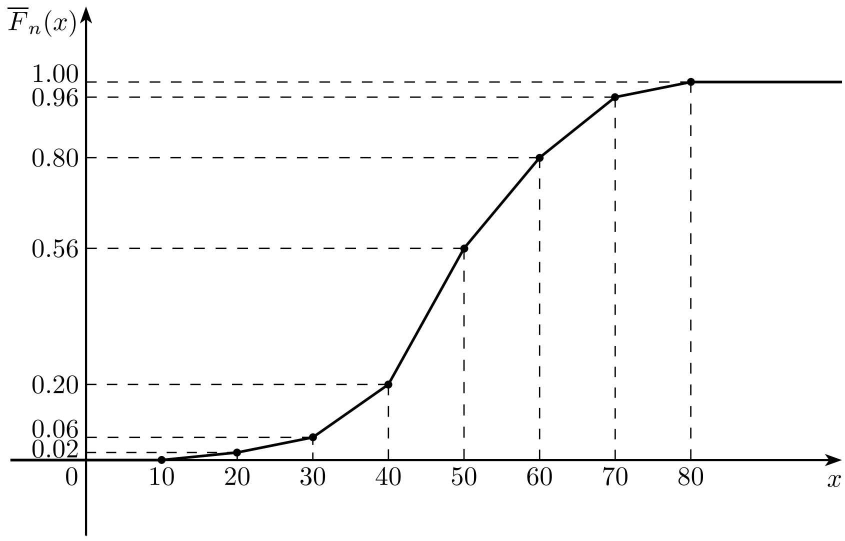
Таблица 9.2: Към пример 9.2

k	I_k	n_k	w_k	W_k
1	10 – 20	1	0.02	0.02
2	20 – 30	2	0.04	0.06
3	30 – 40	7	0.14	0.20
4	40 – 50	18	0.36	0.56
5	50 – 60	12	0.24	0.80
6	60 – 70	8	0.16	0.96
7	70 – 80	2	0.04	1.00

Таблица 9.3: Към пример 9.3



Фиг. 9.4: Графика на $F_n(x)$ от пример 9.2



Фиг. 9.5: Графика на $\overline{F}_n(x)$ от пример 9.3