Лекция 2. Разбиране на данни

Разбирането на данни е един от първите етапи във вече разгледаният модел на процеса за извличането на закономерности от данни. Целта е да бъдат научени определени характеристики на данни, позволяващи да бъдат формулирани първоначални хипотези за природата на скритите в данни зависимости, които от своята страна ще помогнат да бъдат избрани подходящи за тяхното извличане ИЗД алгоритми. В настоящата лекция ще се запознаем с някои статистически методи за разбиране на данни, както и с методи за оценяване на качеството на данни.

2.1. Категоризация на характеристиките на данни

Определено понятие за природата на данни може да бъде получено чрез *изчисляване* на някои техни *сумарни (агрегатни) характеристики*, като например, средните и/или най-често срещаните стойности на описващи данни атрибути, диапазон на техните стойности, степен на корелация между различни атрибути и т.н. Обаче, традиционните начини за изчисляване на подобни характеристики, произхождащи от математическата статистика и интензивно използвани в системи за управление на бази данни, се базират на неявно предположение, че всички данни са достъпни в момента на изчисляване, което в общия случай не е приложимо за случая на ИЗД (където са анализират бази от милиони и дори милиарди записи). Наличието на такива огромни масиви от данни налага ограничение, че данните могат да бъдат достъпни за обработка само *последователно* - на определени *порции*, когато те са, например, натоварени от някое външно запаметяващо устройство в оперативната памет (буфер). При изчисляване на някоя характеристика след натоварване в буфера на всяка следваща порция от данни, всички предишни данни се изтриват, освен определен (много малък) брой стойности, позволяващи да бъдат изчислени желаната сумарна характеристика на всички данни, минали през буфера.

И така, базирайки се на начина на своето изчисляване, всички характеристики на данни могат да бъдат класифицирани в следните три категории:

Дистрибутивна Една сумарна (агрегатна) функция е *дистрибутивна*, ако тя може да бъде изчислена по дистрибутивен (разпределен) начин. Да предположим, че всички наличните данни са разбити на k непресичащи се подмножества. Изчисленията на функцията върху всяко едно подмножество извеждат само по една сумарна нейна стойност. Ако резултатът, изведен от тези k сумарни стойности, съвпада с резултата, който би се получил при изчисляване на същата функция върху *всички* данни (без тяхното "порциране"), ще говорим, че такава функция може да бъде изчислена по разпределен начин. Например, дистрибутивни са такива функции като count() –

брой на записи в база данни, sum() – сума на всички стойности на един атрибут в база данни, min() и max() – минимална и максимална стойност на указания атрибут от база данни.
Една характеристика на данни е *дистрибутивна*, ако тя се получава чрез прилагане на някоя дистрибутивна агрегатна функция.

Алгебраична Една агрегатна функция е *алгебраична*, ако тя може да бъде изчислена чрез прилагане на някоя алгебраична функция с M аргумента (където M е някое крайно положително число), като всеки от аргументите е получен чрез прилагане на някоя дистрибутивна агрегатна функция. Например, алгебраична е функцията avr() – средната стойност, която се изчислява като sum()/count(), тъй като sum() и count() са дистрибутивни функции.
Една характеристика е *алгебраична*, ако тя се получава чрез прилагане на някоя алгебраична агрегатна функция.

Холистична Една агрегатна функция е *холистична*[1], ако не съществува фиксираната граница на размер на паметта, необходима за описание на един неин елемент, изчислен върху някое от подмножество на данни. С други думи, ако не съществува никоя алгебраична функция с M аргумента (където M е константа), характеризиращата необходимите изчисления. Примерите на холистичните функции са median() и mode() – най-често срещани стойности на атрибута.
Една характеристика е *холистична*, ако тя се получава чрез прилагане на някоя холистична агрегатна функция.

За изчисляване на алгебраични характеристики съществуват много различни ефективи похвати, което не е така за холистичните характеристики. Обаче, за някои от тях съществуват ефективни техники за *апроксимацията,* позволяващи да бъдат ефективно изчислени приближени стойности на такива характеристики с достатъчна точност.

2.2. Изследване на основни статистически характеристики на данни

Едни от най-често използваните характеристики на данни са *централната тенденция и разсейването*. Централната тенденция на данни се описва с такива характеристики като средна (аритметична) величина, медиана, мода и среден диапазон, докато характеристиките на разсейването на данни са квартили, екстремни стойности (крайности), вариация и стандартното отклонение. Всички тези характеристики са обект на интензивното изследване в математическата

[1] От *холизъм* – творчески процес на създаване, при който цялото е по-голямо от сумата на части, от което то е било съставено.

статистика, обаче от гледната точка на ИЗД, ние трябва да видим, как те могат да бъдат изчислени в случая на големи бази от данни.

2.2.1. Определяне на централната тенденция

Централната тенденция на данни се описва със средните величини, които са най-често използваните характеристики във всички области на практиката. Всяка фирма съхранява информацията за размера на доставките, продажбите, оборотния си капитал и т.н. Всички тези данни могат да се обобщят и да се представят в удобно за вземане на решение вид с помощта на средни величини. Това важи и за информацията на банките, отнасяща се до размера на представените кредити, лихвени проценти, размер на инфлацията и т.н.

Най-често използвана числова характеристика на "центъра" на едно множество от данни е *средната аритметична величина* (или просто аритметичното средно). Средната стойност на едно множество от *n* стойности $x_1,...,x_n$ се определя като:

$$\widetilde{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (2.1)$$

В релационните бази данни средната стойност се изчислява чрез вградената функция (напр. *avr()* в SQL).

Понякога, всяка стойност x_i от множеството може да има асоцииранo тегло w_i, отразяващо важността, значение или честотата на срещане на съответната стойност. В този случай се изчислява претеглена средна стойност:

$$\widetilde{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \quad (2.2)$$

Тъй като средната стойност се изчислява като алгебраичната функция sum()/count() от дистрибутивни характеристики, то тя сама е алгебраична, т.е. позволява ефективно дистрибутивно изчисление. Лесно е да се покаже, че претегленото средно е също алгебраичната характеристика.

За групирани данни (т.е. когато всички данни са разбити по групи), средната стойност на всички данни може да се изчисли като претеглената средна аритметична от средните стойности на групи:

$$\widetilde{x} = \frac{\sum_{i=1}^{k} n_i \widetilde{x}_i}{\sum_{i=1}^{k} n_i} \quad (2.3),$$

където :

k – е броят на групите,
n_i – е броят на стойности (случаи) в група i, а
$\widetilde{x}_i$ - е средната стойност на група i.

Пример 2.1. На база на проведената анкета, представена в табл. 2-1, да бъде намерена средна аритметична на продължителността на периода, през който безработните нямат работа.

Да предположим, че след подреждането в нарастващ ред, всички стойности на периода без работата са били разбити на 7 групи (подмножества), всяка от която има ширина 2 месеца:

№ на група	**Продължителност на Периода в месеци x_i**	**Брой безработни (честоти n_k)**	**Средни стойности $\widetilde{x}_i$**
1	(0, 2]	8	1
2	(2, 4]	72	3
3	(4, 6]	15	5
4	(6, 8]	14	7
5	(8, 10]	23	9
6	(10, 12]	21	11
7	(12, 14]	2	13
	Общо	**155**	

Табл. 2.1. Разпределение на 155 безработни по продължителност на периода, през който те са без работата

Чрез прилагане на ф-ла (2.3) получаваме:

$$\widetilde{x} = \frac{8 \cdot 1 + 72 \cdot 3 + 15 \cdot 5 + 14 \cdot 7 + 23 \cdot 9 + 21 \cdot 11 + 2 \cdot 13}{155} = \frac{816}{155} = 5.55$$

Средната стойност не е единствената и не е най-добрата характеристика на централната тенденция в данни, тъй като силно зависи от крайните (екстремалните) стойности. За подобни данни по-добра характеристика е *медиана – стойността на атрибута, притежавана от единицата, намираща се в средата на реда, когато единиците са подредени във възходящ или низходящ ред по стойността на изследвания атрибут.* Медианата се определя като стойността на единица с номер *(n + 1)/2,* когато брой на стойностите е нечетен, или като средно аритметично от две средини стойности – когато *n* е четно.

За преодоляване на проблеми, свързани с използване на традиционно дефинирана средната стойност, често се използва понятието *подрязаното (treamed) средно*. Се избира един процент *p* между 0 и 100, определяш *(p/2)%* от наредени в нарастващ ред данни отдолу и отгоре, които трябва да бъдат „изхвърлени" при изчисляване на средната стойност. От тази гледна точка медианата е подрязаното средно с *p = 100%*, докато стандартното средно отговаря на *p = 0%.*

Медианата е холистичната характеристика, т.е. тя не може да бъде изчислена чрез разбиването на едно множество на подмножества, изчисляването на техните медиани по-отделно и изчисляването на общата медиана на базата на получените стойности. Обаче, *приблизителната* стойност на медианата за големите бази данни може да бъде изчислена ефективно. Така, за групираните данни (т.е. данните, представени като съвкупността от стойности и асоциираните с тях честоти на срещането им в базата данни) медианата *Me* може да се интерполира чрез формулата:

$$Me = L_{Me} + \left(N_{Me} - C_{Me-1}\right)\frac{h}{n_{Me}} \quad (2.4)$$

където:

$N_{Me} = \frac{\sum_{i=1}^{k} n_k + 1}{2}$ - номер на медианната единица

L_{Me} – долната граница на медианната група (т.е. подмножество от данни, съдържащо медианата)

C_{Me-1} – кумулативен сбор на честотите, предшестващи медианната група

h – ширината на интервала на медианната група (разликата между максималната и минималната стойности на елементи, намиращи се в подмножеството от данни, съдържащо медианата)

n_{Me} – брой на случаите в медианната група

n_k - брой на случаите в група k

Пример 2.2. На база на проведената анкета, представена в табл. 2-2, да бъде намерена чрез медиана средна продължителността на периода, през който безработните нямат работа.

Да предположим, че след подреждането в нарастващ ред, всички стойности на периода без работата са били разбити на 7 групи (подмножества), всяка от която има ширина 2 месеца:

№ на група	**Продължителност на Периода в месеци x_i**	**Брой безработни (честоти n_k)**	**Кумулативни честоти C_i**
1	(0, 2]	8	8
2	(2, 4]	72	80
3	(4, 6]	15	95
4	(6, 8]	14	109
5	(8, 10]	23	132
6	(10, 12]	21	153
7	(12, 14]	2	155
	Общо	**155**	

Табл. 2.2. Разпределение на 155 безработни по продължителност на периода, през който те са без работата

За да приложим ф-ла (2.4), трябва да намерим в коя група се съдържа медианата. За тази цел, първо, се намира номерът на медианната единица: (155 + 1)/2 = 78, след това се

построява прогресивно кумулативно честотно разпределение (C_i) и се определя, в коя група се намира елемент с номер 78. Вижда се, че това е група 2 – с долната граница на съдържащи се в нея стойности – 2 и горната – 4. Подставяйки всички необходими данни в ф-ла (2.4) получаваме:

$$Me = 2 + \left(\frac{155+1}{2} - 8 \right) \frac{2}{72} = 3.96 \text{ месеца}$$

Очевидно е, че изчисляваната по тази формула характеристика е алгебраична, тъй като се базира на използване на дистрибутивната функция *sum()* – за създавана на кумулативно честотно разпределение.

Другата характеристика на централната тенденция е *мода – най-често срещана стойност на атрибута в изследваното множество от данни*. Възможно е няколко стойности да имат една и съща най-голяма честота на срещане. Данните с една единствена мода се наричат *унимодални*, а с повече от една – *мултимодални*. Ако всички стойности се срещат само по веднъж, то в такава база данни модата не съществува.

Модата е също холистичната характеристика, която обаче може да бъде ефективно апроксимирана по следната формула:

$$Mo = L_{Mo} + \frac{(n_{Mo} - n_{Mo-1})h}{(n_{Mo} - n_{Mo-1}) + (n_{Mo} - n_{Mo+1})} \quad (2.5)$$

където:

L_{Mo} – долната граница на модалната група (т.е. подмножество от данни, съдържащо модата)

h – ширината на интервала на модалната група (разликата между максималната и минималната стойности на елементи, намиращи се в подмножеството от данни, съдържащо модата)

n_{Mo} – брой на случаите (честотите) в модалната група

n_{Mo-1} - брой на случаите (честотите) в предмодалната група

n_{Mo+1} - брой на случаите (честотите) в следмодалната група

Пример 2.3. На база на проведената анкета, представена в табл. 2-1, да бъде намерена средна продължителността на периода, през който безработните нямат работа, с помощта на модата.

От таблицата се вижда, че модалната група е с номер 2. Следователно, получаваме:

$$Mo = 2 + \frac{(72-8)\cdot 2}{(72-8)+(72-15)} = 3.06 \text{ месеца}$$

Ако разпределение на данни е унимодално и средно асиметрично, то за такива данни е приложима следната емпирична формула, позволяваща лесно да се изчислява модата, по известни средната стойност и медианата:

$$\tilde{x} - Mo = 3 \times (\tilde{x} - Me) \quad (2.6)$$

*Средният диа*пазон е още една алгебраична характеристика на данни, представляваща средно между най-голямата и най-малката стойности на атрибута в изследваното множество от данни. Изчислението й е тривиално и се базира на вградените в повечето системи за управление на бази данни функции min() и max().

2.2.2. Измерване на разсейването на данни

Степента на различието между стойностите на един и същ наблюдаван признак (атрибут) се нарича *статистическо разсейване*. Най-често използвани характеристиките на разсейването са *пет-числено обобщение* (базирано на квартили), *межквартилното разстояние* и *стандартното отклонение*.

k-ти квантил (или *процентил*) на множество от числови данни, подредени във възходящ ред, се нарича стойността *x*, която има свойство, че *k* процента от всички данни имат стойности по-малки или равни на нея. Така разгледаната по-рано медиана представлява 50-ти процентил. Освен медианата, най-често използвани процентили са *квартили*. *Първият квартил* Q_1 е 25-ти процентил, а *третият квартил* Q_3 е 75-ти процентил. Квартилите, включително и медианата, дават представа за центъра, разсейването и форма на разпределението на данни. Разстоянието между първия и третия квартил е една проста характеристика на разсейването, която показва диапазон, покриван от средната половина от данни. Характеристиката се нарича *межквартилното разстояние (interquartile range – IQR)* и се дефинира като:

$$IQR = Q_3 - Q_1 \quad (2.7)$$

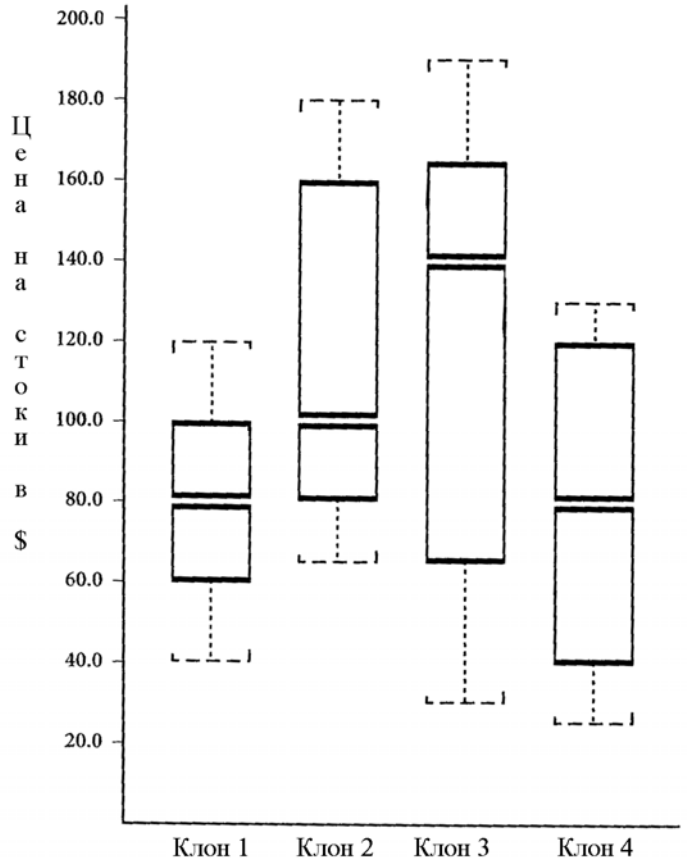
Тази характеристика се използва в емпиричното правило за определяне на "подозрителните" екстремни стойности – крайностите (outliers): за такива могат да се приемат стойностите, които са разположени най-малко на разстояние 1.5 x IQR зад третия квартил или под първия квартил.

Трябва да се има предвид, че една единствена характеристика от типа на IQR не е достатъчна за описание на несиметричното разпределение, тъй като в този случай разсейването от двете страни на разпределението не е еднакво. По тази причина една по-пълна картина за характера и форма на разпределението се дава от така наречено *пет-числено обобщение*, което включва данни за медиана *M*, квартилите Q_1 и Q_3 и най-малката и най-голямата стойности на наблюдавания атрибут – *Min* и *Max*. Това обобщение е удобно да се представи в графична форма с помощта на правоъгълниците (boxplot), а именно:

- Краища на правоъгълника са квартили, така че дължината на правоъгълниците съответства на межквартилното разстояние.
- Медианата се маркира като линия вътре в правоъгълника
- Двете пунктирни линии извън правоъгълника се простират до най-малката *Min* и най-голямата *Max* стойности на наблюдавания признак.

Подобни графики са удобни за сравнение на няколко различни множества от сравними (еднотипни) данни. Например, на фиг. 2.1 са показани графики на данните за продажба на един същ набор от стоки за определен период в 4 клона на една и съща фирма (AllElectronics).

И квартилите, и межквартилното разстояние са холистични характеристики. Ефективното изчисляване на пет-численото обобщение е предмет на текуща изследователска работа.



Фиг. 2.1. График на данни за продажби в четири клона на фирма AllElectronics

Дисперсия (вариация) и стандартно отклонение са характеристики, които измерват степента на отклонение от средната аритметична величина. *Дисперсията* е средна аритметична от квадрати на разликите между стойностите на атрибута и тяхната

средна аритметична стойност. Дисперсията *Var(X)* на *n* наблюдения $X = \{x_1,...,x_n\}$ се изчислява като:

$$Var(X) = \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \tilde{x})^2 = \frac{1}{n-1}\left[\sum x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2\right] \quad (2.8)$$

Стандартното отклонение σ е корен квадратен от дисперсията *Var(X)*.

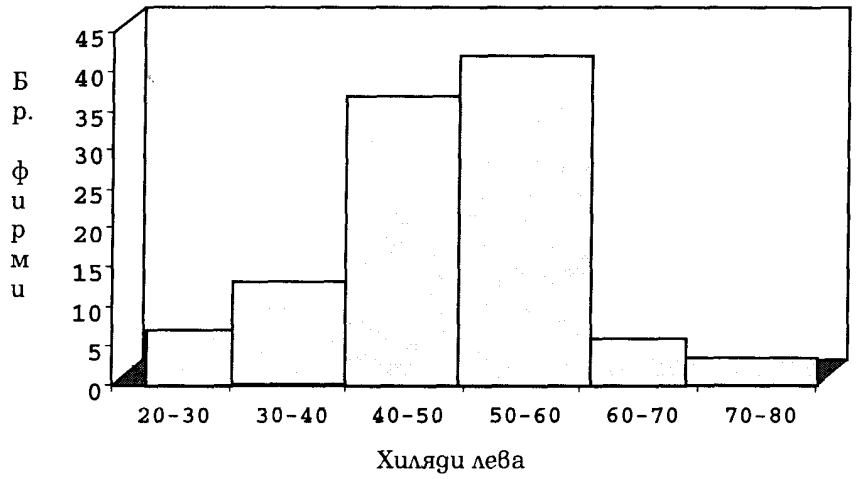
Основните свойства на стандартното отклонение като характеристиката на разсейването са:

- σ измерва разсейването относително *средната стойност* и трябва да се ползва само в случаи, когато тя е избрана като характеристика на центъра на разпределението.
- $\sigma = 0$ само тогава, когато няма никакво разсейване, т.е. когато всички наблюдения имат една и съща стойност; в противен случай $\sigma > 0$.

Обърнете внимание, че и стандартното отклонение, и дисперсията са алгебраичните характеристики, тъй като и *n*, и Σx_i, и Σx_i^2 могат да бъдат изчислени за всяко подмножество по-отделно и след това обединени чрез алгебраични изчисления съгласно ф-ла (2.8).

2.2.3. Графично представяне на базови статистически характеристики

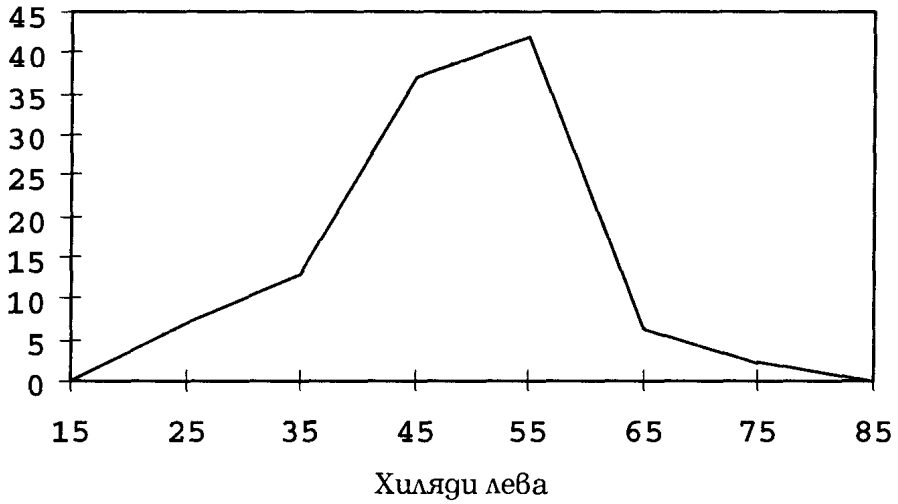
Честотни хистограми Един удобен и нагледен начин за представяне на разпределението на стойностите на един атрибут е *честотни хистограми* (виж пример по-долу).



Една хистограма се състои от множеството на правоъгълници, отразяващи броя или честотата на срещане на атрибутните "класове". За номиналните атрибути класове са самите номинални стойностите. За непрекъснатите атрибути – класове са

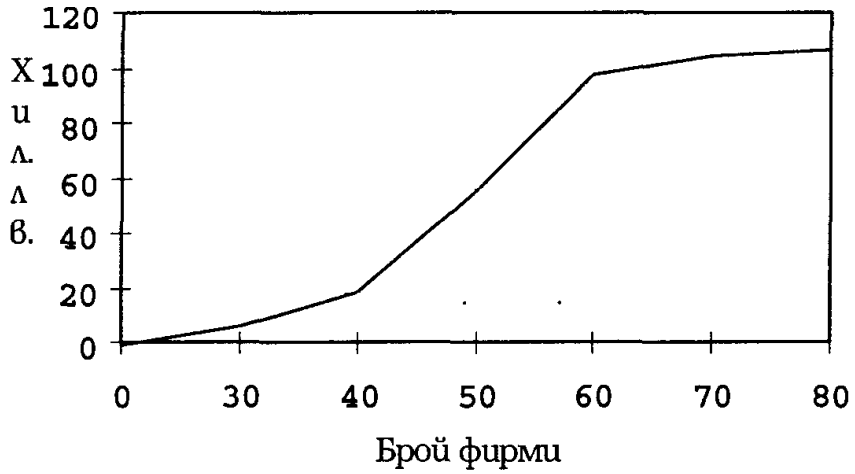
дискретизационни интервали, на които са разбити всички стойности на атрибута, срещани в база от данни. Основата на всеки от правоъгълниците е разположена върху хоризонталната ос и центрирана върху "етикета" на класа, а дължината на основата равна на дължината на класа. Обикновено за номинални атрибути ширината на класа е една и съща (тя не носи никаква смислена натовареност), а "етикетът" е самото име на стойността. Ширината на класове на непрекъснати атрибути се определя от използвания метод за дискретизация (виж следващата лекция), като най-често се използва метод за разбиване на дискретизационни интервали с еднаква ширина. "Етикетът" на класа в този случай е числовата стойност на средата на интервала. Когато височина на правоъгълника е равна на броя или относителната честота на срещане на съответния клас, такива хистограми се наричат *плоскостни диаграми от правоъгълници (bar chart)*. За непрекъснати атрибути класовете могат да имат и различна ширина. В този случай тя отразява ширината на дискретизационния интервал, а височината на правоъгълника – плътността на класа (т.е. броя или честота на срещане на класа, делена на ширината на класа). Хистограмите се използват вече над 100 години и са един много популярен метод за графичното представяне на едномерни данни.

Честотният полигон (полигон на разпределението) е друг обичаен метод за графично представяне на разпределението на едномерни данни. Построяването на честотния полигон става чрез нанасяне на честотите на вертикалната ос, а значенията на признака - на хоризонталната ос, по начина подобен на хистограмата. След това, от средата на всеки интервал се издига перпендикуляр с височина, съответстваща на честотата. Върховете на перпендикулярите се свързват с линия, която оформя честотния полигон. В двата края се добавят интервали с честоти нула за да може той да осигури началото и край върху хоризонталната ос.



Честотният полигон представя разпределението много ясно. Освен това, той е много добро средство за сравнение на няколко разпределения.

Диаграмата на кумулативното честотно разпределение се нарича *огива*. Огивата се използва, когато е нужно да се определи колко наблюдения има над или под дадена стойност в съответното разпределение. Огивата „по-малко от" показва колко единици от разпределението имат стойност по-ниска от горната граница на всеки интервал. Конструирането на огивата изисква, първо, да се построи кумулативно честотно разпределение. Второ, кумулативните честоти се отбелязват с точка в горната граница на всеки интервал, и накрая тези точки се свързват с права линия. Тази линия и представлява огивата.



2.3. Изследване на зависимости между атрибути

2.3.1. Използване на графично и таблично представяне

Първоначалната представа за наличие на зависимостите между избраните двойки от непрекъснати атрибути може да бъде получено с помощта на *графика на*

разсейване (scatter plot). За построяване на този график всяка двойка стойности на атрибути се интерпретира като двойка координати, съответстващи на определена точка от равнина. Този метод е удобен за визуализация на двумерни данни, например, за да се видят възможни клъстери от точки, екстремните точки и т.н. (виж, например, графикът на разсейване, показващ наличието на проста нелинейна зависимости между два атрибута (променливи) – цена на стока и брой на нейните продажби:

За изследване на съществуващите зависимости между два номинални атрибута може да се използват т.нар. *съвместни таблици (crosstab)*. При тях се разглежда честотното разпределение на стойностите едновременно на два атрибута. Броят на различните стойности, които двата атрибута могат да приемат, определя размера на таблицата (броя на колоните и редовете в нея). Във всяка клетка се разполага броят и/или процентът на примерите, имащи съответните стойности за двата атрибута. На фиг. 2.2 е даден пример за съвместна таблица за номиналните атрибути ,,клиентът е доволен от използваните услуги" и ,,клиентът ще използва нови услуги". На нея може да се открие зависимостта, която вторият атрибут има от първия. Освен за изследване на съществуващите зависимости, съвместните таблици могат да се използват и за откриване на грешки в използваните данни, тъй като се базират на честотното разпределение на атрибутните стойности.

	Клиентът е доволен от използваните услуги	**ДА**	**ДОНЯКЪДЕ**	**НЕ**	**ОБЩО**
Клиентът ще използва нови услуги					
ДА	Брой примери Процент	152 69.4%	43 19.6%	24 11%	219 100%
НЕ	Брой примери Процент	8 33.3%	5 20.8%	11 45.8%	24 100%
ОБЩО	Брой примери Процент	160 65.8%	48 19.7%	35 14.4%	243 100%

Фиг.2.2. Образец на съвместна таблица

2.3.2. Корелационен анализ

Корелационният анализ позволява да бъде измерена силата на връзка (зависимост) между избрана двойка *непрекъснати* атрибути *X* и *Y*. В основата на анализа лежи характеристика, наречена *ковариация* (съвместна вариация – изменение), показваща как *X* и *Y* се променят заедно:

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \tilde{x})(y_i - \tilde{y}) \quad (2.9)$$

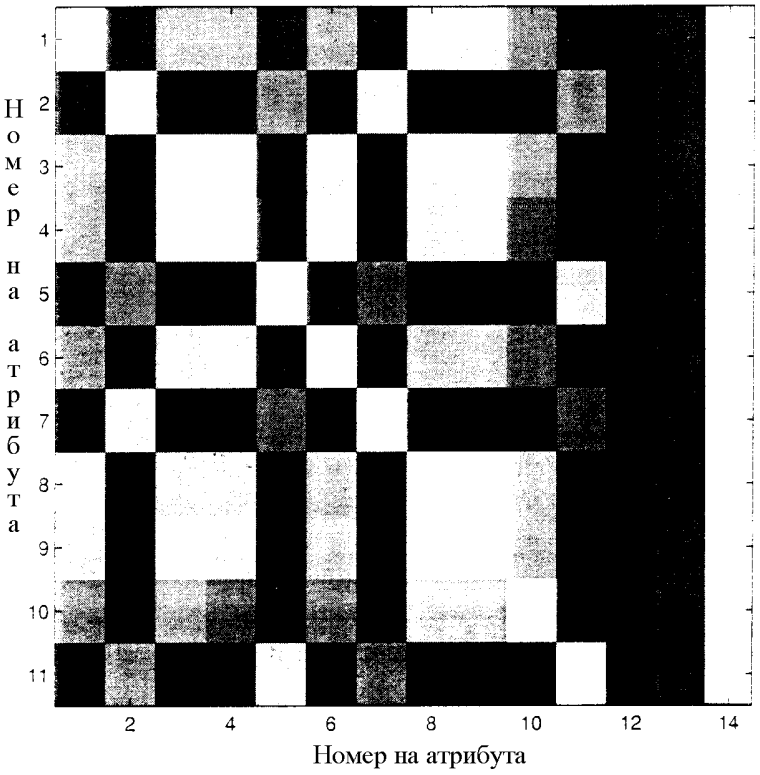
където $\tilde{x}$ и $\tilde{y}$ са, съответно, средни аритметични на множествата от стойности на съответните атрибути в изследваната извадка от данни. Ковариацията ще има голяма положителна величина, ако големите стойности на X имат тенденция да бъдат асоциирани с големите стойности на Y, а малките стойности на X – с малките стойности на Y. Когато големите стойности на X се асоциират с малките стойности на Y, то ковариацията приема отрицателната стойност.

Зависимостта на стойността на ковариацията от диапазоните на X и Y може да бъде премахната чрез стандартизация на атрибутните стойности, т.е. чрез тяхното разделяне на съответните стандартни отклонения. В резултата се получава характеристика, наречена *коефициент на корелация*:

$$r(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \tilde{x})(y_i - \tilde{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \tilde{x})^2 \sum_{i=1}^{n}(y_i - \tilde{y})^2}} \qquad (2.10)$$

Стойността на коефициента разна на +1 означава пълна положителна зависимост, -1 – пълна отрицателна (като тези стойности са почти невъзможни в реални ситуации), а 0 – липсва на линейна зависимост.

В общия случай, когато базата данни се описва с p атрибута, може да бъде съставена p x p корелационна матрица. На фиг.2.4. е показан образец на такава матрица за 11-мерна база от данни, визуализирана като пикселното изображение, позволяващо да се види, как корелират различни атрибути Така, например, всеки от атрибути 3 и 4 силно отрицателно корелира с атрибут 2 и положително корелира с атрибут 11. Атрибутите 8 и 9 са също така силно корелират.



Фиг. 2.4. Образец на корелационна матрица, представена във вид на изображение. Бяло съответства на +1, а черно – на –1. Три най-десни колонки съдържат стойностите –1, 0 и 1 съответно, за да служат за сравнение с интензивността на останалите пиксели. Останалите 11 x 11 пиксела представляват 11 x 11 корелационна матрица, описваща различни атрибути на предградията на г. Бостон.

Необходимо да се помни, че корелацията хваща *само линейната* зависимост между атрибутите. Така че независимост на атрибутите влече липсата на корелация, докато обратното в общия случай не е вярно.

2.3.3. Проверка на хипотези при корелационния анализ

В общия случай ние не работим с цялата генерална съвкупност от данни, а само с някоя нейна извадка. В този случай, както винаги, е актуален въпрос: доколко са достоверни са заключения, получени на базата на извадката? При корелационния анализ въпросът е дали може или не на основата на данни от извадка (размер на извадка n и изчислен коефициент на корелация $r(X, Y) \neq 0$) да се направи извод, че между два атрибута X и Y от генералната съвкупност (база данни с практически неограничен размер) съществува линейна зависимост (т.е., че $\rho(X, Y) \neq 0$).

Нулевата хипотеза H_0 е че в генералната съвкупност няма корелация (т.е. $\rho = 0$), а алтернативната $H_1 : \rho \neq 0$.

След като случайно се избира извадка от X-Y двойки с помощта на ф-ла (2.10). се изчислява корелационният коефициент r. Получената стойност се подлага на проверка за статистическата значимост (виж лекции 14-15 от курса по Машинно самообучение). Доказано е, че ако нулевата хипотеза е вярна, то стойностите на r се подчиняват на t-разпределението на Стюдент с $n-2$ степени на свобода. (Напомням, че t-разпределение има същата камбановидна форма както и нормалното, но е по-широко и по-късо. $t_{\alpha,\nu}$ има два индекса – първият определя желаното равнище на значимост α, а вторият параметър, който се нарича *степени на свобода* ν, е свързан с броя на независими случайни събития, които участват в определяне на стойността на случайната величина). Двете степени на свобода се губят, защото двата параметъра на генералната съвкупност (μ_1 и μ_2) са оценени на основа на извадката (X и Y). Стойността на оценената стандартна грешка се изчислява с помощта на уравнение:

$$s_r = \frac{1 - r^2}{\sqrt{n-2}} \quad (2.11)$$

За оценяване на хипотезите се използва статистиката:

$$t = \frac{r - \rho}{s_r} \quad (2.12)$$

Ако $t > t_{\alpha,n-2}$, то нулевата хипотеза се отхвърля, иначе – тя се приема.

Пример 2.3. Направена е случайна извадка от 30 студенти, за които са изчислени продължителност на посещаване на лекции по ИЗД и продължителност на подготовката за изпит по ИЗД. Въз основата на получените данни е изчислен коефициентът на корелация за двете променливи, чиято стойност е –0.23. Дали съществува линейна зависимост между тези две променливи в генералната съвкупност (т.е. дали това е вярно за всички студенти, посещавали курса по ИЗД)?

1. Проверява се нулевата хипотеза:
   $H_0 : \rho = 0$
   $H_1 : \rho \neq 0$
2. Равнище на значимост $\alpha = 0.05$ (т.е. достоверност 95%).
3. Осигуряване на необходимата информация за изчисления:

   $r = -0.23$

   $$s_r = \frac{1 - r^2}{\sqrt{n-2}} = 0.179$$
4. Изчисляване на емпиричната характеристика (статистика):
   $$t = \frac{r - \rho}{s_r} = \frac{-0.23 - 0}{0.179} = -1.285$$
5. По таблицата на t-разпределение се определя теоретичната характеристика

   $t_{\alpha,n-2} = t_{0.05,28} = 2.048$

6. Сравняват се емпиричната и теоретичната характеристика. Тъй като $t < t_{\alpha,n-2}$, то *нулевата хипотеза се приема.* Това означава, че според данни от извадката *липсва* връзка между продължителността на посещаваните лекции по ИЗД и продължителността на подготовката за изпит по ИЗД.