1. СТАТИСТИКАТА КАТО НАУКА

> "Развитието на съвременната наука върви под знака на интереса към масовите явления и скоро няма да има такъв клон от знанието, където с по-голям или по-малък успех да не разпростират своето влияние статистическите форми на познанието."
>
> А. А. Чупров

Както за всяка наука, така и за статистиката, фундаментален въпрос е въпросът за нейния предмет и метод. В тази глава читателят ще узнае какъв е характерът на обективните явления и на проявяващите се в тях закономерности, породили и пораждащи необходимостта от статистически подход при тяхното изучаване. Ще разбере какви са познавателните възможности на този подход, наречен статистически метод. Ще добие представа каква е същността на статистическата теория и какъв е нейният понятиен апарат. В практико-приложен аспект ще разбере кога и защо в своята практическа дейност трябва и може да си служи със статистически знания и със съответна статистическа информация.

Много науки разкриват повече или по-малко пълно своя предмет с наименованието си. Това не важи за статистиката като наука. Нейният предмет не може да бъде изяснен чрез разкриване на етимологията на думата "статистика". Предполага се, че тя произлиза от италианските думи *stato* – държава и *statista* – държавник. Според други предположения се свързва с латинската дума status – състояние, положение на нещата, с немската *der Staat* - държава и др. По повод на тези различия в тълкуването на произхода и смисловото съдържание на "статистика", известният немски статистик от втората половина на 18 в. ***А. Шльоцер (1735 - 1809)*** отбелязва, че тя не е нито латинска, нито италианска или немска, а е "хибридна" и е получавала различно съдържание в хода на развитието на статистическата наука и практика.

И днес думата статистика предизвиква различни представи: за дейността по събирането и обработването на статистически сведения; за

статистически данни, подредени в специални сборници; за определени методи за анализ и др. Следователно не може да се намери задоволителен отговор на въпроса относно предмета на статистиката в произхода на нейното наименование. Малко може да помогне в това отношение и проследяването на историческото й развитие. От нейното възникване като наука през втората половина на 17 в. до днес предметът й е дефиниран различно и дори противоречиво. Тя е определяна в различни варианти като методологична наука, като универсална наука, като обществена наука и пр.

Не е възможно и не е необходимо в тази книга да се изложат различните определения за предмета на статистиката и мотивите към тях, както и позициите, от които се подлагат на критика в хода на дискусиите едни или други схващания. Ще бъде изложен само един възможен подход за изясняване спецификата на статистическото изучаване, а въз основа на това и характера на теорията на статистиката като наука.

1.1. Масови явления и статистически закономерности

Заобикалящата ни действителност е многообразие от различни по характер явления, подчинени на определени вътрешни закономерности. Независимо от многообразието в съдържанието им, по форма на проявление биха могли да се разделят на единични и масови явления.

Единичните явления са отделно взети предмети, събития, процеси, форми на организация и др. Те, разбира се, не съществуват изолирано и независимо от други явления, намират се в сложни взаимодействия и причинно-следствени връзки. Когато обаче се изучават като единични явления, се установяват закономерности, които са валидни за всеки отделен случай, събитие, процес и др. Като се опознаят тези закономерности чрез наблюдения, лабораторни изследвания, експерименти и други, може с увереност да се твърди, че закономерността ще се прояви или явлението непременно ще настъпи при определени, известни вече условия. Иначе казано, тези явления са детерминирани. Тяхното поведение се описва точно от изведения един път закон, от установените вече закономерности. Такива закономерности се наричат ***динамични***.[1]

[1] Това наименование те са получили първо в класическата механика във връзка с изучаването на законите на механичното движение.

Известно е с каква точност астрономите предвиждат положението на небесните тела и свързаното с това точно определено време на изгрева и залеза на слънцето, появата на слънчеви и лунни затъмнения и т.н. Възхищава ни точността, с която днес се извеждат в желана орбита космическите кораби. Никой не се съмнява, че законът на Архимед е валиден за всяко тяло, потопено във вода. Известно е предварително при какви условия кипи и замръзва водата, какво съединение ще се получи, ако се смесят солна киселина и цинк и т.н. Биха могли да се посочат още много примери на динамични закономерности в различни области. Тяхното разкриване и използуване има огромно значение за науката и за практическата човешка дейност. Но те са само една от формите на проявление на вътрешната същност на явленията.

Масовите явления се състоят от множество единици (случаи) или имат многократно повторение във времето и притежават свойства, качества и "поведение", каквито нямат и не могат да имат отделните единици (случаи). Иначе казано, в тях се проявяват закономерности, които са им присъщи именно като масови явления. Тези закономерности не са строго детерминирани, не се проявяват по еднакъв начин при дадени условия, а като обща тенденция, с по-големи или по-малки колебания, т.е. имат вероятностен характер. Те се наричат ***статистически закономерности***.

Когато се разглеждат например родените деца в отделни семейства, изглежда че съществува пълен безпорядък в съотношението между момчета и момичета. Когато обаче се обхванат ражданията като масово явление, установява се поразителна закономерност - раждат се повече момчета при едно относително трайно съотношение. У нас например, през последните 50-60 години съотношението между родените момичета и момчета е 100:106.

Логично е да се предполага, че има зависимост между доходите и потреблението на определени продукти в домакинствата, но нейната форма и сила се проявява в достатъчно голяма съвкупност от домакинства като статистическа закономерност. Същото може да се каже за зависимостта между производителността на труда и трудовия стаж на работниците, между замърсяването на околната среда и някои видове заболявания, между наторяването и добивите при селскостопанските култури и т.н.

Интересът към статистическите закономерности се обуславя не само от вечния стремеж на човека към познание, но и от практическата

необходимост да се управляват управляемите масови явления или да се регулира поведението на хората и обществото в съответствие със закономерностите на неуправляемите, но обективно съществуващи явления. Този интерес е предизвиквал и продължава да предизвиква търсенето на такъв подход и изследователски апарат в познавателния процес, който да е адекватен на характера и особеностите в проявлението на статистическите закономерности.

Статистическите закономерности не могат да се разкриват като се изучават отделни единици, колкото и подробно да е такова изучаване. Щом са закономерности на масовите явления, те могат да се изучават само ако явленията се разглеждат като единство на множество еднотипни единици (случаи).

Много явления в ***икономиката*** и в ***социалната област*** са масови в посочения смисъл, в тях се проявяват статистически по форма закономерности и за изучаването им е необходим подходящ подход. Това учените са установили още през 17-18 в. От средата на 19 в. напълно се утвърждава убеждението, че науката не може да проникне в тези закономерности като изучава само отделни случаи.

Статистическите закономерности се разкриват и в ***естествените науки***. В областта на биологията от времето на ***Чарлз Дарвин (1800-1882)*** редица явления в органичния свят се изучават като масови явления със закономерности, които не могат да се проявят в отделните индивиди.

Чрез статистически анализ ***Грегор Мендел (1822-1884)*** разкри законите на наследствеността.

Във ***физиката***, където за пръв път се дефинирани динамичните закономерности, също се установява, че много явления, изучавани от нея, имат масов характер и са подчинени на статистически закономерности. От началото на 20 в. се утвърждава нов подход към тези явления. Пример в това отношение е квантовата физика, в която убедително се доказва, че в света на микрочастиците се проявяват статистически закономерности. Бащата на квантовата физика ***Макс Планк (1858-1947)*** откри, че лъчевата енергия се състои от елементарни частици – кванти, чието поведение не може да се изследва с методите на класическата физика, а изискват статистически подход. ***Вернер Хайзенберг (1901-1976)***, друг първостроител на квантовата физика, формулира "принципа на несигурността" и също стигна до извода, че при изследване на микросистемите може само с помощта на статистическия метод да се установят закономерностите и да се правят предвиждания.

Невъзможността да се изследват закономерностите в движението на микрочастиците с методите на класическата механика и необходимостта от прилагането на статистическия подход е илюстрирана образно от френския физик ***Паские*** със следния пример: "За да се изучи напълно движението на молекулите в един кубически сантиметър газ при 0° по Целзий и обикновено налягане, би следвало съгласно със законите на класическата механика да се състави система от диференциални уравнения, всяко от които съдържа милиарди членове, отразяващи взаимодействието на всички молекули... Ако бихме пожелали да изследваме движението на всяка от тези молекули в течение само на една секунда, би трябвало да изгубим за това 10 милирда века, т.е. около 20 милиарда човешки поколения"[1] .

Примери за много явления, които имат масов характер и в които се проявяват статистически закономерности, могат да се посочат и в областта на ***медицината***, ***метеорологията***, ***лингвистиката***, ***технологичните процеси*** и др.

Разбира се, динамичните и статистическите закономерности не си противостоят, тъй както не си противостоят единичните и масовите явления. Статистическите и динамичните закономерности са ***различни форми*** на проявление на всеобщата закономерна връзка между явленията. Природата на явленията и процесите в обективната действителност намират израз в различни ***по същество*** закономерности - социални, биологични, физични и други, които се изучават в съответните науки. Те обаче се проявяват ***във формата*** на динамични или статистически закономерности. Според тази форма се прилага различен подход при изучаването им.

1.2. Познавателна същност и характерни особености на статистическия метод

Закономерностите изобщо не се проявяват непосредствено, открито, видимо. Те се разкриват чрез научно изследване с помощта на подходящ изследователски апарат. Затова и статистическите закономерности не могат да се разкрият и изучат без научно изследване, без съответен научен подход. Този подход е ***статистически***.

[1] Цитирано по: ***Пасхавер, И. С.*** Закон больших чисел и закономерности массового процесса. М., 1966, с.20.

Изучаването на статистическите закономерности е ***статистическо изучаване***.

Терминът статистически метод често се употребява с две значения.

Първо, като специфичен подход към разкриване и опознаване на конкретното проявление на статистическите закономерности в масовите явления.

Второ, като определен конкретен начин за получаване на статистически характеристики в процеса на статистическото изучаване. В този смисъл се говори например за индексен метод, репрезентативен метод и др. В същия смисъл, употребен в множествено число, терминът статистически методи се използува за означаване на целия апарат от понятия, правила, формули, процедури и други, прилагани при статистическото изучаване.

Тук ще разгледаме статистическия метод в смисъла на първото определение. Различните конкретни статистически методи, прилагани при статистическото изучаване, се разглеждат в следващите глави на книгата.

Статистическото изучаване трябва да се разбира като ***форма на познавателен процес***, необходима във всички области, в които явленията са масови в посочения смисъл и закономерностите по форма на проявление са статистически. Според характера на изучаваните обективни явления и същностната страна на резултатите, изучаването може да бъде икономическо, социологическо, демографско, биологично и т.н., но според метода и конкретния изследователски апарат е статистическо.

За да се разберат познавателната същност на статистическия метод и специфичния характер на статистическото изучаване изобщо, необходимо е да се имат предвид някои основни положения относно начина на проявление на статистическите закономерности.

Всяко масово явление е под действието както на основни, трайни и систематични причини, които имат определяща роля, така и на причини със случаен характер. Основните причини са свързани с вътрешната същност на явленията и определят основно техните закономерности. Случайните, краткотрайните причини влияят в една или друга степен и в една или друга посока върху отделните единици (случаи) на масовите явления. Те предизвикват ***случайни отклонения*** от общата закономерност.

Известно е от философията, че необходимостта и случайността са неотделими, намират се в единство. Във философски смисъл случайността е форма на проявление на необходимостта, а необходимостта се проявява чрез случайността. Тази връзка между случайността и необходимостта е обективна и предопределя необходимия подход при изучаването на закономерностите. Тя следователно е обективната основа, върху която се опира статистическият метод като своеобразен подход за опознаване на статистическите закономерности.

"Мостът", свързващ абстрактно-философския аспект на връзката между необходимото и случайното със статистическия метод, е ***законът за големите числа***. В логическата си същност той е израз на посочената обективна връзка и единство на необходимостта и случайността, но я конкретизира като отношение между единичните случаи, формиращи масовото явление, и общата закономерност, която ги обединява. Законът за големите числа е доведен до формата на общ принцип на познавателния процес при изучаване на масовите явления. Той е намерил израз и в математически теореми с практическо приложение при статистическото изучаване.[1] Логическият смисъл на закона за големите числа и неговото приложно-практическо значение се състои най-общо в следното.

Отделният единичен случай на проявление на масовото явление не разкрива общата статистическа закономерност, но той съдържа частица от нея в случайна и завоалирана форма. Когато се обхванат ***достатъчно голям брой случаи*** (единици), отделните случайности се неутрализират и се проявява закономерността. Не би могло например да се установи, че съществува закономерност да се раждат повече момчета, отколкото момичета при относително трайно съотношение, ако не се наблюдават достатъчно голям брой случаи. Ако се наблюдават само отделни домакинства, не би могло да се установи зависимост между доходи и потребление, но ако се наблюдават достатъчно голям брой домакинства, такава зависимост може да се разкрие.

Законът за големите числа, като израз на единството и връзката между случайното и необходимото, между единичното и общото, определя ***първата характерна особеност*** на статистическия метод - ***масовостта на статистическото изучаване***. Това означава, че опознаването на закономерностите на масовите явления по статистически

[1] Някои основни положения, свързани със закона на големите числа, се разглеждат в глава 4.

път изисква да се обхващат достатъчно голям брой случаи, чрез които се проявява масовото явление, за да се прояви действието на закона за големите числа.

Втората характерна особеност на статистическия метод се състои в това, че чрез него се изучава ***количествената страна*** на масовите явления. Трябва да се има обаче предвид, че качествената и количествената страна на явленията се намират в единство. За статистическото изучаване това има значение в две отношения.

Първо, статистическото изучаване трябва да се опира върху знанията за качествената природа на явленията. Статистическите понятия, които се формират при статистическото изучаване, отразяват действителността не непосредствено, а опосредствано от съответните теоретични понятия, те са техни статистически аналози.

Второ, изучаването на количествената страна на явленията води до установяване на статистически закономерности и до изводи, чрез които могат да се изяснят важни страни на качествената същност на явленията, може да се проникне по-дълбоко в тяхната природа или да се открият дори неизвестни преди това техни качества и свойства.

Третата характерна особеност на статистическия метод се състои в това, че чрез него се установява ***конкретно проявление*** на закономерностите в дадени времеви и пространствени граници, изразено в числа и мерки. Знанията, които той осигурява, са конкретни, получени въз основа на осигурен по съответен начин емпиричен материал. Трябва обаче да се има предвид, че чрез обобщаване на тези конкретни знания може да се стигне в някои случаи до изводи относно общовалидни закономерности, независими от времеви и пространствени граници, т.е. да се стигне до общотеоретични обобщения и изводи.

1.3. Теория на статистиката

Статистическото изучаване, в каквато и да е област, не може да бъде успешно, ако не се основава върху научни знания относно: познавателната същност и особеностите на статистическия метод; условията за коректното му използване; принципите на планиране и организация на събирането, систематизирането и обобщаването на статистическите сведения; същността, изчисляването, анализирането и интерпретирането на обобщаващите статистически характеристики и т.н. Изобщо статистическото изучаване трябва да се основава на научна система, на ***статистическата теория***.

Специфичните особености на отделните области на действителността налагат особености и на статистическото й изучаване. Макар че изхожда от някои общовалидни принципи, то в една или друга степен се модифицира. Различните понятия на науките, изучаващи по същество съответните явления, трябва да се специфицират в ***статистически понятия***. Съществуват особености и по отношение на приложимите средства за анализ. Всичко това е наложило възникването и развитието на цяло семейство статистически науки: обща теория на статистиката, икономическа статистика, социална статистика, демографска статистика и др. Затова понятието статистическа наука или статистическа теория трябва да се разглежда като събирателно, обхващащо всички отрасли на тази наука. Обединяващо звено е общата теория на статистиката.

Общата теория на статистиката е наука за познавателната същност и общите методологични и организационни основи на статистическото изучаване.

От това определение следва по-конкретно, че общата теория на статистиката:

1. Изяснява същността, възможностите и условията за приложение на статистическия метод като специфичен подход за изследване на конкретното проявление на закономерностите на масовите явления.
2. Предлага разработени и непрекъснато усъвършенствани общи принципи и правила за планиране, организиране и осъществяване на статистическите изучавания.
3. Съдържа общи статистически понятия като специфицирани статистически отражения, като статистически по форма познавателни образи на действителността.
4. Предлага конкретни методи за получаване на обобщаващи числови статистически характеристики и за техния анализ.

Системата от знания, които формира общата теория на статистиката, е теоретична основа на статистическите изучавания във всички области на действителността. Тези знания се трансформират в конкретни ръководни положения при емпиричните статистически изучавания чрез теорията на съответните отраслови статистики.

Това е в общи линии съвременната теория на статистиката. Но, както всяка друга наука, тя е изминала сравнително дълъг път на развитие. Началото й е поставено във втората половина на 17-ти век. За

първи път през 1660 г. в Германия ***Херман Конринг (1606-1681)*** въвежда преподаването на нова университетска дисциплина, наречена ***държавознание***. Тя е съдържала описание на всичко за държавата, което според Конринг, трябва да знаят държавните дейци – територия, население, армия, финанси, държавно устройство и др. Един от най-активните последователи на Конринг и първостроители на новата наука ***Готфрид Ахенвал (1719-1772)*** я нарича статистика през 18-ти век, а ***Август Шльоцер*** произнася популярната фраза "статистиката е застинала история, а историята е текуща статистика". Все още тогава е запазен обаче нейният описателен характер. Затова в историята на статистиката нейното съдържание от онова време се нарича описателно направление или ***описателна статистика***.

По същото това време, когато в континентална Европа се заражда и развива описателното направление, в Англия възниква друго, наречено ***политическа аритметика*** по заглавието на една книга на ***Уилям Пети (1623-1687)***. Именно в изследванията на Пети в областта на икономиката и на неговия съвременник ***Джон Граунт (1620-1674)*** в областта на демографските процеси се прилага един нов подход, чрез който се разкриват непознати дотогава закономерности в масовите явления. Този подход, доразвит и в много отношения усъвършенстван понататък от науката, наричаме днес ***статистически метод***. Наред с това се изгражда цялостната статистическа методология и стройната система на теорията на статистиката.

1.4. Основни статистически понятия

Както всяка наука, така и статистиката има свой понятиен апарат, свои специфични понятия и съответстващата им терминология.

1.4.1. Статистическа съвкупност

Основно понятие в статистиката е понятието ***статистическа съвкупност***. Беше подчертано, че масовите явления са определени общности от множество единици (случаи), в които се проявяват статистически закономерности.

Статистическата съвкупност е общността от единици (случаи), чрез които се проявява дадено масово явление, изучавано по

статистически път, дефинирана винаги по същество и в определени пространствени и времеви граници.

Масовите явления съществуват обективно. Когато се изучават статистически, се обединяват отделните елементарни форми на проявление (единици, случаи) по нещо общо за всички тях, което ги прави еднородни, еднотипни. Така се формират статистически съвкупности от еднородни в някакво отношение единици, чрез конкретните характеристики на които може да се стигне до обобщаващи характеристики на целите съвкупности и по този начин се стига до разкриване на конкретното проявление на интересуващите ни закономерности в масовите явления. Трябва при това да се има предвид, че еднородността на съвкупността се разглежда винаги в точно определено отношение, от някаква гледна точка.

Всяка съвкупност се състои от определен брой единици, които образуват ***обема на съвкупността***. Съвкупността има своя вътрешна ***структура***. Нейните единици имат някакво ***разпределение***. Вътре в съвкупността се проявяват и определени ***връзки и зависимости***. Съществуват връзки и зависимости и между различни съвкупности. Изобщо в статистическите съвкупности се проявяват онези общи свойства и закономерности, които могат и трябва да бъдат обхванати с оглед целите на статистическото изучаване.

Статистическите съвкупности се класифицират по видове от различни гледни точки.

1. Съвкупностите могат да бъдат формирани от единици (случаи), съществуващи в точно определен момент, или от единици (случаи), възникващи в рамките на определен период. От тази гледна точка те биват ***моментни и периодни***. Моментна съвкупност е например населението на Република България на определена дата към определен критичен момент при преброяване на населението. Сключените бракове в София през 2007 г. образуват периодна съвкупност.

При дадено статистическо изучаване могат да се дефинират и да се наблюдават и моментни, и периодни съвкупности. Често те са свързани в балансови равенства и това позволява да се изучи възпроизводството на съвкупностите и интензивността на определени потоци. В такива случаи се използва понятието ***динамични съвкупности***.[1] Например, заетите лица

[1] Вж. **Сугарев, З.** Развитие във времето на статистически съвкупности и структури, сп. Статистика, бр. 5, С., 1977.

лица във фирма "А" към 1 януари 2005 г. и към 31 декември 2005 г. са моментни съвкупности. В периода между двата момента се формират периодни съвкупности като потоци: приети (входящ поток) и напуснали (изходящ поток) през годината. Тези моментни и периодни съвкупности са свързани в балансово равенство: като се прибави към броя на заетите в началото на годината броят на назначените и се извади броят на напусналите, се получава броят на заетите в края на годината.

2. Съществено значение има делението на съвкупностите на генерални и представителни. ***Генералната съвкупност*** обхваща всички единици (случаи) на даденото явление. Представителната съвкупност, наричана обикновено ***извадка***, обхваща част от единиците на генералната съвкупност, която при известни условия представя генералната съвкупност, т.е. характеристиките, получени от извадката, могат да се приемат като приближения (оценки) на параметрите на генералната съвкупност.

3. В някои случаи има смисъл деленето на съвкупностите на реални и хипотетични. ***Реалната съвкупност*** е действително съществуваща, състояща се от краен брой единици, които могат реално да бъдат обхванати и наблюдавани. ***Хипотетичната съвкупност*** е въображаема безкрайна съвкупност. Ние мислено си я представяме като такава, за да поставим реалната съвкупност в отношение спрямо нея и да изведем положения, отнасящи се до "поведението" на реалната съвкупност. Обикновено представата за хипотетична съвкупност е свързана с формирането на извадка при условията на безкрайно голям брой възможности единиците на генералната съвкупност да попаднат в извадката.

4. Когато явленията се разглеждат в развитие и се сравняват съвкупности за едни и същи по характера си явления, но обхванати чрез статистически наблюдения през различните периоди или в различни моменти, съвкупностите се делят на диференциални и интегрални. ***Диференциалната съвкупност*** се състои от единици (случаи), които при следващото изучаване не могат да влязат в аналогична съвкупност. Например съвкупността на родените и съвкупността на умрелите през 2007 г. са диференциални съвкупности. През следващата година се формират други съвкупности на родени и умрели. Нито една единица от родените и умрелите през 2007 г. не влиза в съвкупностите през 2008 г. При ***интегралните съвкупности*** обикновено всички или част от единиците, обхванати при едно наблюдение, могат да се съдържат и в съвкупностите при наблюдение на същото явление в някой следващ

момент или период. Например населението, домакинствата и др. към определен момент образуват интегрални съвкупности в посочения смисъл. Очевидно е, че това разграничение на съвкупностите е тясно свързано с разграничението им на моментни и периодни и с понятието динамични съвкупности, но то се прави от друг аспект и има значение при изучаване измененията във вътрешната структура на съвкупностите. Тези изменения могат да стават поради ***вътрешни потоци***, когато едни и същи единици променят положението си в съвкупността и поради ***външни потоци***, когато има зараждане (вливане отвън) на нови единици и изчезване (излизане навън) на съществуващи преди това единици.

5. Съществува условно разграничение на съвкупностите на общи и частни. Една съвкупност е ***обща***, ако се разглежда като такава спрямо друга, която е ***частна***. Трябва да се има предвид, че дадена съвкупност ще бъде частна в посочения смисъл, ако е подложена на допълнително самостоятелно статистическо изучаване наред или след изучаването на общата съвкупност.

1.4.2. Статистическа единица

Единиците (случаите), чиято общност образува дадена статистическа съвкупност, се наричат статистически единици.

Статистическата единица (наричана още единица на статистическата съвкупност) се разглежда при всяко статистическо изучаване като конкретна, елементарна и неделима по-нататък форма на проявление на даденото масово явление. Тази неделимост не трябва да се схваща в абсолютен физически смисъл. Единицата може да е отделен човек, домакинство, фирма, трудова злополука и т.н.

Съвкупността не може да се дефинира добре без точно определяне на статистическата единица. В редица случаи е необходимо да се фиксират критерии, по които се преценява дали дадена единица има качества да бъде третирана като единица на дефинираната статистическа съвкупност. Ако например статистическата единица е домакинството, то трябва да се определят критерии, по които се дефинира тази единица.

1.4.3. Статистически признаци

Статистическите единици ни интересуват при статистическите изучавания дотолкова, доколкото имат определени белези, свойства,

качества, прояви и пр., проучването на които е необходимо, за да се изведат някакви обобщаващи характеристики за съвкупността. Тези особености, свойства и пр., които ни интересуват при статистическото изучаване, се наричат ***статистически признаци***.

Всеки признак на статистическата единица получава при статистическото изучаване съответна ***характеристика (значение, определение).*** Под характеристика, значение или определение на признака трябва да се разбира отговорът, който се дава на въпроса относно дадения признак. Ако на въпроса относно възрастта на едно лице отговорът е "25 години", то 25 години е характеристика (значение, определение) на признака възраст.

Статистическите признаци са различни по вид, което налага различен подход при статистическото им изучаване.

1. Основното им разграничение е на вариационни и категорийни. ***Вариационните признаци***, наричани още количествени или метрирани, са тези, чиито характеристики се изразяват с числа. Те могат да бъдат прекъснати (дискретни, дисконтинюитетни) и непрекъснати (индискретни, континюитетни). ***Прекъснати*** са вариационните признаци, които могат да приемат отделни, изолирани една от друга характеристики. Такива са броят на работниците в отделно предприятие, броят на членовете на отделно домакинство и др. ***Непрекъснати*** са признаците, които могат да получават всякакви характеристики в даден числов интервал. Такива са: себестойността на единица изделие, месечната заплата на отделните работници и др.

Категорийните признаци, наричани още атрибутивни, качествени или неметрирани, са признаци, чиито характеристики нямат числов израз, а се дават словесно, описателно. Такива са например пол, семейно положение, отраслова принадлежност на фирмата и др.

Особен вид категорийни признаци са тези, които имат две възможни характеристики, представляващи проста алтернатива. Такива признаци се наричат ***дихотомни, алтернативни*** или ***бинарни***. Такъв е например признакът пол.

Има категорийни признаци, които могат да получават при отделни статистически единици две или повече характеристики едновременно. Например признакът, зададен с въпрос: "Какъв чужд език ползвате?" може при отделна единица (например студент) да има два или повече отговора. Такива категорийни признаци се наричат ***кумулативни***.

2. В зависимост от това, дали характеристиките на признаците при отделни единици остават неизменни, или се изменят с течение на времето, признаците се делят на ***постоянни*** и ***непостоянни***.

3. Когато признаците се намират в определена връзка помежду си, при която едни от тях влияят като фактори върху други, които в някаква степен са резултат от първите, се делят на ***факторни*** и ***резултативни***. Това разграничение има значение при статистическия анализ.

4. Статистическите признаци могат да се разграничават и според това, дали ги притежават всички единици на съвкупността или само част от тях. От тази гледна точка те се делят на ***всеобщи*** и ***невсеобщи***.

5. Когато статистическите признаци се отнасят за хора, понякога се делят на ***естествени*** и ***социални*** според това, дали са естествено обусловени (пол, възраст), или социално (професия, образование).

1.4.4. Обобщаващи числови статистически характеристики

Многократно беше отбелязано, че чрез статистическото изучаване се стремим да разкрием конкретното проявление на закономерностите в масовите явления, че логическият път на статистическото познание е от единичното към общото, от случайното към необходимото.

Статистически израз на общото, типичното, закономерното за съвкупностите са ***обобщаващите числови статистически характеристики***. Те са обобщаващи, защото се получават чрез обобщаване на множество характеристики на отделните единица и дават концентриран числов израз на една или друга важна страна на конкретното проявление на изучаваните явления. Те могат по същество и вид да бъдат различни и се получават посредством подходящи процедури. Продукт са обикновено на статистическия анализ.

1.5. Практикум

1.5.1. Въпроси за самопроверка

1. Кои закономерности се наричат статистически?
5. Кои са характерните черти на статистическия метод?
6. Каква е логическата същност на закона за големите числа?
7. Какъв е предметът на теорията на статистиката?

8. Каква статистическа съвкупност са фирмите в България към 1 януари 2008 г.?
9. Коя е статистическата единица при статистическо изучаване на безработицата?
10. Кои признаци се наричат вариационни?
11. Кои признаци са бинарни (дихотомни)?
12. Какъв признак е признакът "семейно положение"?
13. Какво се разбира под обобщаващи статистически характеристики?

1.5.2. Работата не е в числата

"Вие не сте ми разказали – забеляза лейди Нател – с какво се занимава Вашият жених.

- Той е статистик – отговори Леймия ...

Лейди Нател очевидно беше зашеметена ...

- Но лельо Сара, това е много интересна професия - каза с жар Леймия.

- Аз не се съмнявам в това – отговори нейната леля, която очевидно се съмняваше. – Да се изрази нещо значително само в едни числа ... Но не мислите ли Вие, че животът със статистик би бил в известна степен скучен?

Леймия мълчеше. На нея не й се искаше да говори за поразителната дълбочина на емоционалните възможности, които тя беше открила под числовата външност на занятието на Едуард.

- Работата не е в числата - каза тя най-после – а в това, какво правите с тях."[1]

[1] Цит. по: **Kendall, M., A. Stuart**, The Advanced Theory of Statistics, London, 1961, p. 9.

2. СТАТИСТИЧЕСКО ИЗУЧАВАНЕ

> "Ще дойде време, когато статистическото мислене ще е толкова необходимо на цивилизования човек, колкото и умението му да чете и пише."
>
> Х. Уелс

В тази глава се изясняват логиката и основните процедури на реалния процес на статистическото изучаване. Тяхното познаване е необходимо за всеки специалист, който в своята професионална област си служи със средствата на статистическата наука или с информация от статистически изследвания. Читателят ще узнае по-конкретно през какви фази преминава емпиричното статистическо изучаване, какви са неговите видове и форми. Ще се научи как да обобщава и представя изходните сведения от своите наблюдения, ще знае кога и как да си служи с различните статистически скали, какви рискове за грешки го съпътстват, какъв е характерът на различните статистически величини, които се получават от статистическите изследвания и каква е тяхната интерпретация.

2.1. Същност и основни фази на статистическото изучаване

В гл. 1 беше дадена характеристика на статистическото изучаване като познавателен процес. Бяха описани гносеологичните основи на този процес и неговите особености. Но за да се обхване и изучи по статистически път конкретното проявление на закономерностите в масовите явления, необходимо е да се планират и реализират по определени правила и в определена форма набирането, обработването, обобщаването и анализирането на конкретната статистическа информация за интересуващите ни явления, т.е. да се извърши ***емпирично статистическо изучаване***. То обхваща серия от процедури и етапи на работа в цялостния процес - от получаването на първичните сведения за отделните единици

(случаи) до изчисляването на обобщаващите числови статистически характеристики и тяхното логическо интерпретиране по същество.

Всяко емпирично статистическо изучаване има за обект (в най-широк смисъл) някакво масово явление, проявяващо се в дадени пространствени и времеви граници. Затова то винаги има определен териториален обхват и продължителност по време и се осъществява по цялостна ***програма*** при съответна организация.

Обикновено когато се говори за емпирично статистическо изучаване, има се предвид общността от всички операции, които трябва да се изпълнят, за да се постигне неговата крайна цел. След като тази цел е изяснена и е извършена цялата предварителна проучвателна и подготвителна работа, целият следващ процес на статистическото изучаване се разделя на три условно отграничени фази: 1) статистическо наблюдение; 2) статистическа групировка; 3) статистически анализ.

Статистическото наблюдение е началната фаза, в която се получават по съответно организиран начин първичните сведения за отделните единици.

Статистическата групировка като втора фаза обхваща онези процедури, посредством които единичните сведения се обобщават и се получават статистически данни за обособени на някакво основание групи.

Статистическият анализ обхваща действията, посредством които се получават обобщаващи числови характеристики за изучаваните явления, разкриващи тяхната същност, изменения, вътрешни връзки и зависимости и др., съобразно крайните цели на статистическото изучаване.

Такива са логическият път и основните фази на едно емпирично статистическо изучаване, когато то трябва да осигури изходните сведения, последващата им обработка до "крайния продукт". Разбира се, не всяко изучаване, което по форма е статистическо, трябва да започва с регистриране на първичните сведения. Неговата начална точка могат да бъдат обобщените вече данни от извършени преди това наблюдения и групировки. Съвременните автоматизирани регистри и банки от данни са широка информационна основа за по-нататъшни статистически изследвания, удовлетворяващи едни или други познавателни цели. Но дори и тогава се използува натрупана информация в друго време и по друга програма, която е продукт на извършено статистическо наблюдение и съответна групировка, подчинени на определени принципи и правила.

Всяко статистическо изучаване изисква предварителна подготовка и изпълнение в строго съответствие с предварително съставени програма и организационен план.

Програмата трябва да съдържа ясно формулирана цел, точно и изчерпателно определен обект, конкретно и еднозначно дефинирани статистически съвкупности, статистически единици, статистически признаци и др. Програмата трябва да съдържа също макетите на предстоящите групировки, основните насоки на статистическия анализ и формите (сборници, доклади, технически носители и др), в които ще се предоставят за ползване резултатите от изучаването.

Планът на статистическото изучаване съдържа редица положения относно организацията на изучаването през отделните му фази.

2.2. Обект на статистическото изучаване

Обект на всяко емпирично статистическо изучаване по принцип е някакво масово явление. Това общо положение, макар и вярно не означава, че когато се проектира едно статистическо изучаване и се назове явлението, обектът е фиксиран и не се нуждае от изясняване. Има много случаи, при които е трудно интуитивно да се определи обхватът на явлението, обект на статистическото изучаване, и да се разграничи то от останалите, близки до него по характер явления. Необходими са по-строги дефиниции и критерии, за да се определи обектът недвусмислено и точно и съответно да се дефинират статистическите съвкупности и статистическите единици. Ако престои например да се изучи по статистически път интелигенцията в страната, за да се установят нейният брой, състав, разпределение по териториални единици, в селата и градовете и т.н., едва ли веднага, без колебание може да се определи обектът на статистическото изучаване, и то не въобще, а с такава конкретност, че да може да се формира статистическата съвкупност. Необходимо е точно понятие за интелигенция, а следователно и критерий, по който всяко лице ще се отнася или не към обекта на изследването. Същият въпрос възниква и ако обект на изучаване е средната класа в обществото. Тя едва ли може да се дефинира еднозначно и безспорно. Очевидно е, че за правилно определяне на обекта на статистическото изучаване в най-конкретен смисъл са необходими знания, които по характера си се формират главно от науката, изследваща по същество дадената област.

2.3. Видове статистически изучавания

Характерът на изучаваните явления, целите на изучаването, разполагаемото време, финансовите средства и други обуславят различни по вид статистически изучавания. Класифицирането им може да се извърши от различни гледни точки.

1. Според обхвата статистическите изучавания биват изчерпателни и частични.

Изчерпателните статистически изучавания обхващат всички единици (случаи), чрез които се проявяват дадените явления, т.е. обхващат целите генерални съвкупности.

Частичните статистически изучавания обхващат само част от единиците на съвкупностите. Това са обикновено ***извадкови изучавания***. На понятието за пълноценно статистическо изучаване в строгия смисъл на думата отговаря само такова извадково изучаване, което осигурява възможност въз основа на характеристиките, получени от наблюдаваните единици в извадката, да се правят достатъчно надеждни оценки за параметрите на целите (генералните) съвкупности. Това е репрезентативното (представителното) статистическо изучаване. То има строги научни основи и намира широко приложение в различни области. (вж. гл. 5)

Друг вид частично статистическо изучаване е изучаването на ***основния масив***. Чрез него се обхваща преобладаващата част от единиците на съвкупността и се прилага в случаите, когато останалата част не оказва влияние върху резултатите и освен това наблюдението й е свързано с трудности или изобщо е невъзможно.

Към частичните статистически изучавания често се причисляват статистическата оценка, статистическата монография и статистическата анкета. Строго погледнато, те не отговарят точно на понятието за статистическо изучаване и затова обикновено се наричат сурогати (заместители) на статистическите изучавания. В наше време те почти са изгубили практическото си значение.

2. Статистическите изучавания биват еднократни и текущи.

Еднократните статистически изучавания се организират и провеждат като правило по твърде широка програма, за да се получи подробна информация за явленията. Това са обикновено явления, които не се изменят бързо и текущото им изучаване не е целесъобразно или ако се изучават и текущо, това става по ограничен брой признаци. Еднократно не значи едно единствено изучаване на даденото явление. Такива изучавания могат да се извършват и през някакви периоди. Особеното

Особеното обаче е това, че за даденото изучаване се изготвя програма, извършват се наблюдението и следващите фази на изучаването, с което то приключва. Когато отново възникне необходимост да се изучи даденото явление, отново изучаването се планира, организира и провежда по нова програма. Преброяванията например на населението, на търговската мрежа и други са еднократни, независимо, че не се извършват само един единствен път.

Текущите статистически изучавания се извършват непрекъснато (текущо) по установена програма и организация. Според характера на явленията регистрирането на сведенията за статистическите единици се извършва при тяхното възникване или през определени интервали (периоди). Тези интервали за различните явления могат да бъдат с различна продължителност, но са предварително определени и наблюдението се извършва в тази периодичност и по съставената програма, докато възникне нужда да се направят изменения. Така се наблюдават например потребителските цени на стоките и услугите, производството на определени изделия, вносът и износът на стоки, раждания и умирания и др.

2.4. Статистическо наблюдение

И най-съвършената програма, и най-добрите намерения за всестранно анализиране на статистическите данни ще останат нереализирани, ако наблюдението не гарантира качествени изходни сведения. Ето защо на статистическото наблюдение трябва да се гледа като на изключително важен етап (фаза) от целия процес на статистическо изучаване. Необходимо е не само да се намери най-добро решение на всички организационни и други въпроси, но да се приложи и най-подходящата за случая форма на статистическо наблюдение.

Възможни са ***две основни форми на статистическо наблюдение***: статистическа отчетност и специално организирано статистическо наблюдение.

Статистическата отчетност се състои в предоставяне на държавните статистически органи на сведения от страна на предприятия, учреждения, институти и други в определени срокове по утвърдени статистически отчетни формуляри. Тя е правно регламентирана и е съобразена с необходимостта от ритмичен поток от информация за определени икономически и социални процеси.

Специално организираните статистически наблюдения имат за цел да обхванат явления, които не се наблюдават чрез отчетността, или да осигурят по-подробна информация, каквато не би могла да се осигури от отчетната форма на наблюдение. Преброяването на населението и жилищния фонд, на търговската мрежа, на селскостопанските животни и други са типични примери на такива специално организирани статистически наблюдения. Такива са наблюденията и при социологическите анкети, различните извадкови изследвания в областта на икономиката, в социалната сфера и др.

При планиране и организация на статистическото наблюдение се решават множество въпроси относно единицата на наблюдението, времето и мястото на наблюдението, източниците на сведенията, метода на наблюдението, органите, формулярите и др.

Единицата на наблюдението е онази клетка, от която се получават сведения за статистическата единица (единицата на съвкупността). При отчетната форма единица на наблюдението е стопанската или друга единица (фирма, организация), която е задължена да представя отчетни статистически сведения в съответни статистически формуляри. В този случай тя обикновено се нарича ***отчетна единица***. При специално организираните статистически наблюдения единицата на наблюдението трябва да бъде определена предварително, като се има предвид коя единица от всички възможни ще осигури в организационно и друго отношение най-добри възможности за получаване на необходимите сведения за статистическите единици.

Въпросът за ***времето на наблюдението*** има различни страни. При текущото изучаване е необходимо да се определи какви периоди ще обхваща всяко наблюдение или през какви интервали от време ще се фиксира състоянието на моментните съвкупности.

При еднократните изучавания особено значение има определянето на най-подходящото време за наблюдение, след като е възникнала необходимост от изучаване на явлението. Основното съображение е да се определи времето, когато явлението се намира в своето нормално състояние и условията за наблюдение са най-благоприятни.

Времето на наблюдението трябва да се конкретизира и в точно определен календарен период за регистриране на сведенията за отделните единици.

Много голямо значение има ***критичният момент*** на наблюдението. Той има смисъл при периодни и при моментни съвкупности.

При моментни съвкупности това е онзи конкретен момент, към който се фиксира състоянието на съвкупността, т.е. значенията на признаците на статистическите единици се определят към този момент и се наблюдават само тези единици, които са съществували в този момент. Без определянето на критичния момент не е възможно изобщо обхващането на моментните съвкупности. Сведенията за всички единици се отнасят към този момент, независимо кога те практически се регистрират в определения период за наблюдение.

Във връзка с ***мястото на наблюдението*** трябва да се разглеждат: 1) териториалният обхват на наблюдението; 2) мястото на регистрацията. В първия случай въпросът се свързва преди всичко с териториалното разположение на единиците на изучаваните съвкупности и трябва да се реши още когато се изяснява обектът на изучаването и се определят статистическите съвкупности при съставянето на програмата на статистическо изучаване. При специално организираните статистически изучавания, имащи характер на преброявания, територията, върху която се извършва наблюдението, се разделя на участъци с точно определени граници, за да се организира правилно наблюдението и точно да се определи териториалният обхват на дейността на лицата, ангажирани като органи на наблюдението.

Въпросът за мястото на регистрацията се решава също конкретно при всяко наблюдение, но по принцип се предпочита мястото, където се намират единиците на наблюдението.

Успехът на всяко наблюдение, а следователно и на цялостното статистическо изучаване зависи в много голяма степен от ***органите на наблюдението***. Имат се предвид органите, които изпълняват определени функции при практическото осъществяване на наблюдението (това не са само постоянните държавни статистически органи на страната).

При отчетната форма органи на статистическото наблюдение са съответните служебни лица във фирмите, организациите и други, които имат задължението да съставят статистическите отчетни формуляри и отговарят за достоверността на даваните сведения.

При редица специално организирани еднократни наблюдения се формират органи за даденото наблюдение и те престават да бъдат такива след приключването му. Например при преброяването на населението се създават централна комисия, областни и общински комисии, които ръководят преброяването на дадената територия. Подбират се и лица, които непосредствено извършват наблюдението и регистрацията. Те се

наричат преброители, а при анкетните изследвания - анкетьори и интервюери. Освен това се назначават и контрольори, които подпомагат преброителите (респ. анкетьорите) и извършват контролните наблюдения в определените им райони (участъци).

Държавните статистически органи или други институти, които организират и ръководят изучаването, осъществяват подготовката на органите на наблюдението.

Статистическите органи, както и лицата, участващи само при наблюдението, са длъжни да спазват един особено важен принцип - за тайната на индивидуалните сведения, които физическите и юридическите лица дават при наблюдението. Този принцип е израз на защита правата и свободата на личността и интересите на фирмите, и е правно регламентиран със Закона за статистиката.

Статистическото наблюдение в точния смисъл на думата започва с ***регистрация на сведенията за единиците***.

Възможни са общо три основни начина за регистриране: самонаблюдение, кореспондентски начин и експедиционен начин.

Самонаблюдението е начин, при който отделните лица, които са единици на наблюдението, сами дават сведения и обикновено сами ги записват в статистическите формуляри. Практически с оглед на удобство този начин се прилага във варианти. Възможно е формулярите да се доставят на лицата от преброителите (анкетьорите) и да им се предостави сами да направят записванията в тях. В друг случай преброителите (анкетьорите) водят разговор с лицата, които дават сведенията, и правят записванията. В трети случай формулярите се изпращат по пощенски път до лицата, даващи сведения, те ги попълват и ги връщат обратно. Кой вариант конкретно ще се приеме зависи от характера на изучаваните явления и от увереността, че именно по този начин могат да се получат в предвидения срок обективни и точни сведения.

Кореспондентският начин е този, при който определени лица или органи - специалисти в съответната област, изпълняват функциите на кореспонденти на статистическите или други органи, организиращи изучаването. Те дават сведения за статистическите единици. Този начин се прилага, когато е необходима компетентна преценка от страна на специалисти поради особения характер на изучаваното явление. В някои страни например подбрани вещи лица дават мнение относно промените в цените във връзка с измененията в качеството на стоките. По този начин в

някои страни се събират сведения и относно добивите и реколтите в селското стопанство.

Експедиционният начин се състои в изпращането на специално подготвени за даденото наблюдение лица (експедиции), които извършват на място наблюдението и записват сведенията. Този начин се прилага рядко, когато наблюдението има по-специален характер, изисква специална подготовка и особен подход и няма увереност, че по друг начин ще се получат достоверни сведения.

Единичните сведения се регистрират при всяко наблюдение по определен начин в ***статистически формуляри*** (наричани още въпросници, отчетни формуляри, преброителни карти, анкетни карти и др.). Съставянето на тези формуляри не е само технически въпрос. Необходимо е много внимателно и компетентно да се преценяват неговата форма и размер, начинът на подреждане на въпросите (признаците), начинът на редактиране (по същество и технически) и т.н. Трябва да се изхожда от характера на явлението, от възприетия начин на регистрация, от предвидения начин за обработка на сведенията, от установените държавни стандарти и др.[1]

В практиката все по-широко приложение получава непосредственото въвеждане на получаваните при наблюдението сведения в портативни персонални компютри.

Източниците на сведенията за единиците са различни при отделните статистически изучавания. Общо те могат да бъдат: 1) непосредствено наблюдение; 2) документи; 3) интервю (диалог, беседа).

При ***непосредственото наблюдение*** се извършва броене или измерване, за да се получат необходимите сведения. Така се постъпва в експерименталното дело, при определяне на жилищната площ на жилищата, при статистическия контрол на качеството и др.

Документални източници се използват обикновено при отчетната форма на статистически наблюдения.

Интервюто (беседата) е основен източник на сведения при статистически наблюдения, когато единици на наблюдението са лица, като преброяване на населението, социологически анкети и др.

[1] Относно функциите на въпросниците и основните изисквания и подходи при съставянето им вж. **Станев, С.,** Въпросникът – ключов елемент на статистическото изучаване, сп. *Статистика*, 1993, кн. 1.

2.5. Статистическа групировка

2.5.1. Същност на статистическата групировка

Като продукт на статистическото наблюдение се получава голямо количество единични сведения. В такъв вид те не могат да характеризират изучаваните масови явления. Необходимо е по-нататък тези сведения да бъдат обобщени по съответен начин, за да се приведат във вид и форма, подходящи за анализ. При това обобщаване трябва да се премине от сведения за отделните единици към ***статистически данни***, отнасящи се за качествено еднородни в някакво отношение групи. Това се постига посредством статистическата групировка.

Статистическата групировка не е само техническа процедура за обобщаване на единичните сведения. Тя има дълбоко съдържателна същност и голямо значение в процеса на статистическото изучаване. Чрез нея се прави една крачка от единичното към общото, от отделните характеристики на признаците към обобщаващи статистически характеристики. Групировката създава възможност да се проникне по-нататък във вътрешната структура на явленията, да се разкрият закономерностите в разпределението на единиците на съвкупностите по характеристиките (определенията) на интересуващите ни признаци, да се изследват зависимости между отделните признаци и т.н. В този смисъл тя може да се разглежда и като средство за анализ.

Статистическата групировка обхваща практически ***три основни процедури***:

Първо, образуване на групите въз основа на определени признаци (групировъчни признаци), т.е. съставянето на ***скали***, по които ще се разпределят статистическите единици. Това е същественият момент при всяка групировка.

Второ, отнасяне (разпределяне) на отделните единици (случаи) по скалата според конкретните значения на признаците, преброяване и записване на общия брой на единиците за всяка група.

Трето, определяне чрез сумиране или други аритметични действия на съответните статистически величини, които се отнасят за отделните групи, например сумарните значения на признаците.

2.5.2. Скалиране

Беше посочено, че част от статистическите признаци се характеризират с числа. Наличието на число и мярка е важно условие за понататъшна обработка на сведенията за отделните единици и за получаването на обобщаващи числови статистически характеристики за съвкупностите. За колкото повече признаци теорията е намерила начин за измерване, толкова повече се разширяват възможностите за проникване във вътрешната същност на явленията, в сложните връзки и зависимости между тях. Във връзка с това може да се припомни един девиз на ***Галилей***: **"Измервай измеримото и прави неизмеримото измеримо!"**. Придаването на числови значения на признаците създава възможност за извършване на математически операции с числата и за извеждане по този начин на обобщаващи числови статистически характеристики.

Използват се прости и сложни (комбинирани) единици за измерване. Простите са брой, килограм, тон, литър, час, ден, киловат и т.н. Сложните (комбинираните) съдържат две прости единици. Такива са тонкилометър, човекоден, човекочас и др.

Но въпреки стремежа към числови измерители, такива не са възможни за всички признаци. Каквито и да са признаците обаче, необходимо е да се съставят подходящи скали, за да се направят разпределения (групировки). Скалирането е необходимо и за прилагането на необходимите и възможни методи за анализ.

В статистическите изследвания най-често се прилагат следните скали:

1. ***Номинална скала***. Съдържа наименования на възможните и необходими определения на съответните категорийни признаци, наименования на териториални единици и др. Например за признака семейно положение номиналната скала съдържа 4 възможни поделения, съответстващи на четирите определения на признака: неженен (неомъжена), женен (омъжена), разведен (разведена), вдовец (вдовица).

При много на брой характеристики (определения) на признаците обикновено се използва предварително съставено ***класификации*** и ***номенклатури*** и според целите на групировката скалите се отнасят за по-общи или по-детайлни "етажи" на класификациите. Такива са например класификацията на професиите, класификацията на икономическите дейности, класификацията на продуктите по икономически дейности, номенклатурата на промишлената продукция и др.

Разновидност на номиналната скала е ***дихотомната (бинарната)***, прилагана за дихотомни (бинарни) признаци. По тази скала за двете алтернативни определения на признаците се дават условни числови значения 0 и 1 и по тях съвкупността се разделя на две групи. Например, при статистическия контрол на качеството на продукцията може стандартните изделия да се означат с 1, а нестандартните с 0.

2. ***Рангова скала***. По нея на единиците (обектите) се определят рангове (поредни места) от 1 до n. Ранжирането може да се основава на обективни критерии, свързани непосредствено с дадени признаци, или на субективни експертни оценки. Трябва да се има предвид, че числата (ранговете) по тази скала изразяват само поредното място на единиците.

3. ***Ординална скала***. Съдържа такива поделения, които не са числови, но изразяват различие, степен в дадено качество или свойство на единиците. Такава би била скалата при разделяне на анкетирани лица по степен на владеене на чужд език, ако тази степен се определя примерно с оценките много добре, добре, средно и слабо. Ординалната скала често се разглежда като разновидност на ранговата, но очевидно е, че съществува значителна разлика между двете скали, която се проявява и в подхода при анализа.

4. ***Интервална скала***. Това е числова скала, отнасяща се за вариационни признаци. Характеристиките на признаците не само имат числов израз, но разликата между числата тук измерва различията между единиците по дадения признак. Ако работната заплата на един работник е 700 лв., а на друг е 750 лв., разликата от 50 лв. измерва напълно определено различието между двамата работници по работна заплата.

За съставянето на интервални скали, когато признаците имат много на брой характеристики се прилагат три принципа – аритметичен, геометричен и целеви.

Аритметичният принцип се изразява в съставянето на скали с еднаква ширина на интервалите. Например скала за разпределение на населението по възрастови групи с ширина на интервала 5 навършени години. За съставяне на такива скали се прилагат две формули.

Първо, когато специалистът, съставящ скалата, предварително е преценил и е решил колко групови интервали да формира, тогава ширината на интервалите (h) се определя, като разликата между максималното (x_{max}) и минималното (x_{min}) значение на признака се раздели на броя на групите (k), т.е. по формулата:

$$(2.1) \quad h = \frac{x_{\max} - x_{\min}}{k}.$$

Второ, по формулата на ***Х. Стърджес***, която е изведена чрез формализиране на връзката между ширината на интервалите, максималното и минималното определение на признака и броя на единиците в съвкупността (N), които ще се разпределят по скалата:

$$(2.2) \quad h = \frac{x_{max} - x_{min}}{1 + 3{,}322 \log N}.$$

По ***геометричния принцип*** се образуват интервали с ширина, изменяща се в геометрична прогресия, т.е. всеки интервал има ширина определено число пъти по-голяма или по-малка от предходната.

По ***целевия принцип***[1] се формират скали с различни интервали според конкретните цели, за които ще служат скалите. Например, при групировка на населението по възраст е необходимо по скалата да се обособят 3 подсъвкупности – население под трудоспособна, в трудоспособна и в над трудоспособна възраст.

Приложението на една или друга скала се определя както от характера на съответните статистически признаци, така и от конкретните цели, които се поставят при анализа. Най-добри възможности за приложение на по-съвършени методи предлага интервалната скала и затова тя се нарича силна скала.

2.5.3. Видове статистически групировки

Тъй като групировъчните признаци са различни по вид, и се прилагат различни скали, могат да се получат и различни по вид групировки: вариационна, категорийна (атрибутивна), териториална (пространствена), темпорална (по време).

Вариационната групировка има за резултат разпределение на единиците на съвкупността по характеристиките (значенията) на вариационен признак. Следователно скалата, по която се прави разпределението, е интервална. Например разпределение на фирмите по брой на заетите в тях лица, разпределение на домакинствата по размер на доходите и др.

[1] В областта на икономическата статистика се нарича икономически принцип.

Категорийната (атрибутивната) групировка представя разпределение на единиците на съвкупността по категорийни (атрибутивни) признаци. Скалата, по която се прави тази групировка, е номинална или ординална. Ако признаците са дихотомни (бинарни, алтернативни), прилага се бинарна (дихотомна) скала.

Териториалната (пространствената) групировка задоволява интереса към териториалното разположение на единиците на изучаваните съвкупности. Скалата при териториалната групировка е номинална, съдържаща обособени териториални поделения, които могат да се формират по различни обективни или други критерии в зависимост от предназначението на групировката. Това могат да бъдат икономически райони, области, общини, географски райони и др.

Темпоралната групировка (по време) осигурява разпределение на единиците на съвкупността по време (периоди) на тяхното възникване. Периодите могат да бъдат различни и се определят от конкретни съображения при всяка групировка. Типичен пример е групировката на жилищата по периоди на тяхното построяване.

Групировките могат да се правят както само по един групировъчен признак (по една скала), така и по два или повече признака (скали) едновременно (в комбинация). В първия случай тя се нарича ***единична или едномерна (проста)***, а във втория - ***множествена (сложна, комбинирана)***, или ***двумерна, многомерна***. При множествената групировка обособените по един признак групи се подразделят на подгрупи и по други признаци. По такъв начин всъщност единиците се разпределят едновременно според комбинациите в характеристиките на групировъчните признаци. Множествената групировка разкрива по-големи възможности за анализ и затова понякога се нарича аналитична.

Съществената част от процедурите при групировката е съставянето на скалите. Останалите операции се отнасят до техниката на групировката, която в наше време се изпълнява автоматизирано от компютърната техника, снабдена със съответни софтуерни приложения.

2.6. Статистически анализ

2.6.1. Същност и главни насоки на статистическия анализ

Колкото и голямо да е значението на групировката като необходим етап на статистическото изучаване, това изучаване не може да спре дотук. Данните, получени непосредствено от групировката, могат да дадат само ограничена характеристика на явленията. Те не разкриват непосредствено дълбокото вътрешно съдържание, структурата, тенденциите и зависимостите в явленията. Тези данни са в голяма степен само "суров материал", от който по-нататък чрез анализа трябва да се "изгради" цялостното познание на изучаваните съвкупности.

Статистическият анализ обхваща получаването посредством подходящи методи на обобщаващи числови характеристики и смисловото им интерпретиране. Той обаче не трябва да се схваща ограничено само като заключителна фаза на статистическото изучаване, а да се разбира и по-широко - като целенасочена обработка на фактически числов материал с приложение на статистически методи за целите на научни изследвания, за обосновка на управленски решения и др.

Конкретните цели и задачи на анализа се определят от характера, целите и задачите на даденото изследване, както и от спецификата на явленията. Изследванията в различните области имат различни аспекти и познавателни задачи. В един случай те изискват ***диагностичен анализ***, който да разкрие статуса на явлението в даден момент или период, неговата структура, взаимозависимости и др. В други случаи анализът е ***ретроспективен*** и удовлетворява интереса към поведението на явленията в миналото, условията на тяхното зараждане, етапите на развитие и др. В трети случай анализът е ***прогностичен***, когато изследването е насочено към развитието на явленията в бъдеще, измененията, които ще настъпят в неговите мащаби, състав, скорост на изменение, въздействие върху други явления и т.н. Винаги обаче (когато се изследват съвкупности) освен съдържателна страна, която се обхваща и изучава от съответната наука, анализът има и статистическа форма и статистическо съдържание. Статистическият анализ следователно е еднакво необходим при изследвания, свързани с миналото, с настоящето и бъдещето. Той има своя "територия" и при чисто описателни изследвания и при изследвания, чиято основна цел е разкриването на причинно-следствени връзки, на законите на строежа и развитието на явленията.

Специфичните функции на статистическия анализ обаче се състоят в това ***да даде конкретни числови характеристики*** на явленията въз основа на обработени по съответен начин статистически данни.

Многообразието на целите и задачите, които задоволяват различните изследвания по същество, определят многообразието в насоките, аспектите и средствата на статистическия анализ. Независимо от това многообразие, могат да се очертаят схематично три главни направления на статистическия анализ.

Първо, ***анализ на разпределението*** на единиците на съвкупностите по значенията на интересуващите ни признаци (по съответните скали). В тази извънредно важна област на статистическия анализ се установяват обобщаващи числови характеристики за цели съвкупности относно центъра на разпределението, степента на разсейването (вариацията) на единиците по значенията на признаците, формата (закона) на разпределението и др.

Важно познавателно значение в тази област имат обобщаващите характеристики на ***структурата на явленията***, на степента на концентрация, на териториална локализация, на различията (или сходствата) между отделните структури и др.

Второ, ***анализ на динамиката*** на явленията (динамичен статистически анализ). Това е твърде широка област, в която статистическият анализ разкрива и измерва скоростта на развитието, основните тенденции, цикличните и сезонните вариации и др.

Трето, ***анализ на зависимостите*** между явленията. Това е широко по обхват и богато по съдържание направление на анализа. Взаимната връзка и причинната обусловеност на явленията приема при статистическия анализ конкретна числова форма.

Посочените основни направления на статистическия анализ обхващат множество конкретни задачи. Те се осъществяват с помощта на разнообразни методи, които предлага общата теория на статистиката.

2.6.2. Предпоставки на статистическия анализ

Научната обоснованост и познавателната стойност на резултатите от статистическия анализ зависят в голяма степен от някои важни ***предпоставки***. Тяхното пренебрегване при конкретните изследвания може да стане причина за неверни изводи, колкото и прецизно да е в техническо отношение приложението на съответните методи.

1. На първо място, много голямо значение има ***приоритетът на качествения анализ***. Това означава, че статистическият анализ трябва да се основава преди всичко на достатъчно пълно познаване на качествената страна на явленията, които се анализират - характер, форми на развитие, закони, на които са подчинени, качествено своеобразие, обусловено от специфични условия и т.н. Статистическият анализ трябва да има за основа онези знания за явленията, които дава науката, изучаваща ги по същество

Разбира се, резултатите от всестранния и правилно насочен статистически анализ често разкриват нови страни, нови закономерности и зависимости и служат като основа за формулиране на нови теоретични положения.

2. Важни предпоставки за получаването на верни и обосновани резултати от статистическия анализ са пълнотата, сравнимостта, съпоставимостта и достоверността на статистическите данни.

Пълнотата на статистическите данни се преценява съобразно задачите на анализа и характера на явленията. Необходимо е по-конкретно да се установи дали данните се отнасят за цялата съвкупност, или само за част от нея. Когато се отнасят само за част от съвкупността, трябва да се установи дали могат да се използват за характеризиране на цялата съвкупност, т.е. дали те са представителни (репрезентативни). Трябва да се знае и начинът, по който тези данни са били получени, а следователно и възможността да бъдат използвани именно за определената цел на анализа.

Сравнимостта на статистическите данни е качество, което се определя от възможността вярно и правилно да се установява колко са големи различията в обема на съвкупностите и на техните характеристики. От само себе си се разбира, че такива сравнения са възможни само за едноименни величини, изразяващи еднородни явления. В строгия смисъл на думата сравнимостта изисква най-пълно единство в дефинирането на сравняваните величини съобразно вътрешното съдържание на явленията. Може например да се сравнява броят на населението на София от две преброявания, за да се установи неговото увеличение. На пръв поглед изглежда, че при такава елементарна задача няма опасност да се сравняват несравними величини. Може обаче да се окаже, че едната величина изразява броя на наличното население, а другата - на постоянното и те следователно са несравними.

Съпоставимостта на статистическите данни се определя от възможността да се отнасят (съпоставят) едни към други и да се получават смислени производни величини. Съпоставимостта трябва да се преценява не от формална гледна точка, а от гледна точка на вътрешното съдържание на данните и познавателния смисъл на резултатите от съпоставянето.

Несравнимостта и ***несъпоставимостта*** на статистическите данни могат да възникнат по различни причини: различие в определението на статистическата единица и на съвкупността при различни наблюдения, различно съдържание, влагано в едни и същи термини в отделни страни или през различни периоди от време, специфични за дадена територия или за даден период обстоятелства, повлияли върху състоянието на изучаваните явления и др. За да се знае дали данните са сравними и съпоставими, необходимо е да се знае и по какъв начин са получени, коя е статистическата единица, какво е териториалното и повременното отграничаване на съответните статистически съвкупности.

Безусловно изискване (предпоставка) на статистическия анализ е ***достоверността*** на анализираните данни. Тя може да се разглежда в широк и тесен смисъл. В широк смисъл се свързва със степента, в която статистическите понятия съответстват на теоретичните понятия. Това ще рече преди всичко дали статистическите понятия, използвани при статистическото изучаване, са адекватни операционни аналози на теоретичните понятия на науката, изучаваща по същество интересуващите ни явления, за които се отнасят анализираните данни. В този широк смисъл на достоверността се включва и точността, т.е. достоверността в тесен смисъл, свързана със ***статистическите грешки***.

Колкото и прецизно да е подготвено и проведено едно статистическо изучаване, рискът от грешки не е изключен (особено когато изучаването е голямо и в него вземат участие много хора). Тук се имат предвид такива грешки, които се наричат ***грешки на регистрацията.*** От тях трябва да се разграничат ***грешките на репрезентативността (***стохастичните грешки), които са от съвсем друго естество и се допускат при извадковите статистически изучавания. Те се разглеждат в гл. 5.

Грешките на регистрацията може да се дължат на различни причини. Те могат да произтичат от недостатъци на програмата на изучаването - лошо изяснена и неточно дефинирана статистическа

съвкупност и статистическа единица, неясно формулирани статистически признаци и др. Могат да бъдат допуснати и от органите на наблюдението - поради некомпетентност, небрежност или съзнателно регистриране на неверни сведения. Неверни сведения могат да се дадат и от лицата (съзнателно или несъзнателно), които са единици на наблюдението.

Грешките могат да се изразяват в неточен обхват (ненаблюдавани или повторно наблюдавани единици) или в неверни сведения за единиците. Освен това могат да бъдат логически и аритметически. Те са ***логически***, когато дадените сведения са логически неприемливи, не са верни, не отговарят на истинските характеристики на единиците. ***Аритметически*** са, когато погрешно са извършени някакви аритметични действия - събиране, изваждане, умножение и др.

Грешките могат да бъдат ***съзнателни*** (преднамерени) и ***несъзнателни*** (непреднамерени). Освен това, според отражението им върху обобщените резултати от изучаването, те биват случайни и систематични.

Случайни грешки са тези, които при една част от единиците увеличават характеристиките на признаците, а при друга част ги намаляват. По такъв начин за съвкупността като цяло те в голяма степен се компенсират.

Систематични са грешките, които изменят истинските характеристики в една посока. Те се натрупват и могат значително да повлияят (изопачат) в една посока върху обобщаващите статистически характеристики на съвкупностите. Поради това тези грешки са особено опасни за крайните резултати от статистическото изучаване.

И все пак, колкото и да са съвършени програмата и цялата организация на изучаването, могат да възникнат грешки. Затова много важно значение има контролът върху достоверността на сведенията, проверката на тези сведения и отстраняването на евентуално допуснатите грешки, преди да се пристъпи към анализ.

След като е завършена регистрацията на сведенията при наблюдението и са събрани попълнените статистически формуляри необходимо е да се проверят съдържащите се в тях сведения. Проверката бива аритметическа и логическа. ***Аритметическата проверка*** има за цел да открие и отстрани аритметичните грешки. ***Логическата проверка*** открива логическите грешки. Съвременната електронно-изчислителна техника има възможности (ако се състави съответна програма) да извърши както аритметична, така и логическа проверка и да открива

грешките. Когато при установяване на грешка верният отговор е очевиден, грешката се коригира. Ако обаче това не е възможно, трябва да се потърси единицата на наблюдението и да се установят истинските сведения.

Добре разработените софтуерни програми предотвратяват грешките при статистическата групировка и при изчисляването на съответните обобщаващи статистически характеристики.

2.7. Статистическите редове

При изследване на дадено явление почти никога не се анализират едновременно всички негови страни и следователно не винаги се използва едновременно цялото количество разнообразни статистически данни, получени от статистическите изучавания. Обикновено анализът протича последователно, като се разглеждат поотделно, макар и по обща програма, различните страни на явленията. От друга страна, явленията се разглеждат в тяхното развитие и изменение и във връзка с други явления. Това налага при анализа често да се обхващат данни не само от едно, а от различни статистически изучавания, извършвани по различно време, както и за различни териториални единици. Поради това е необходимо преди статистическия анализ да се подберат необходимите статистически данни и да се подредят по съответен начин съобразно целите на анализа и изискванията на методите, които ще се приложат. Това означава да се съставят статистически редове. Разбира се, много статистически редове се получават като непосредствен продукт на статистическа групировка.

Статистическите редове следователно са подбрани и целесъобразно подредени статистически данни за целите на анализа.

Според съдържанието им статистическите редове могат да изразяват или някакво разпределение на единиците на съвкупностите по избрани скали, или изменението на явленията във времето. Според това те биват ***редове за разпределение*** (статични редове) и ***динамични (хронологични, повременни) редове***.

1. ***Редовете за разпределение*** имат за основания териториални поделения, определения на категорийни (атрибутивни) признаци, характеристики на вариационни признаци или периоди на възникване на единиците на моментни съвкупности. В зависимост от това те биват териториални (пространствени), атрибутивни (категорийни), вариационни и темпорални (редове за разпределение по време).

Териториалните (пространствени) редове изразяват териториалното разположение на единиците на съвкупностите или различията в някакви характеристики по териториалните поделения. Като правило те представят разпределения по номинални скали. Например разпределение на специалистите с висше икономическо образование по области на страната.

Атрибутивните (категорийните) статистически редове съдържат разпределения на единиците на съвкупностите по определенията на атрибутивни (категорийни) признаци. Такъв е например редът, съдържащ разпределението на заетите лица по професионални групи. Разпределенията по атрибутивни признаци са всъщност разпределения по номинална или ординална скала.

Вариационните статистически редове изразяват разпределения на единиците на съвкупностите по вариационни признаци (разпределения по интервална скала). Такова е разпределението на населението по възрастови групи, разпределението на предприятията по брой на заетите лица, разпределението на заетите по размер на работната заплата и др.

Когато вариационният ред изразява разпределението по дискретен признак, чийто значения са малко на брой и показват обикновено някакви степени, отграничаващи една група от друга, статистическият ред се нарича ***степенен (дискретен)*** вариационен ред. Например разпределението на студентите по изпитни оценки по шестобалната система. Когато значенията на признака са много и при групировката са обособени групови интервали с долна и горна граница, редът е ***интервален***, образуван по интервална скала.

Темпоралните редове (за разпределение по време) са резултат на темпоралната групировка и изразяват разпределение на единиците на моментни съвкупности по периоди на тяхното възникване.

2. ***Динамичните*** (хронологичните) статистически редове изразяват изменение, развитие на явлението във времето. Те биват моментни и периодни.

Моментните динамични редове съдържат данни за моментни съвкупности. Например дълготрайните активи на дадена фирма към първо число на всеки месец през 2007 г.

Периодните динамични редове се отнасят за периодни съвкупности. Например брутният вътрешен продукт на Р. България по години през периода 1990 – 2007 г.

Когато в динамичните редове се проявява определена тенденция на

на развитие, се наричат ***нестационарни*** (регуларни), а когато липсва такава тенденция - ***стационарни*** (ирегуларни). Разкриването на конкретните тенденции е предмет на анализа на динамичните редове.

2.8. Статистически таблици

Огромното количество статистически данни, които се получават при статистическите изучавания, е необозримо, ако тези данни не се представят в подходяща форма, улесняваща използването им, най-лесното им възприемане. Една рационална форма за такова представяне са ***статистическите таблици***.

2.8.1. Същност и елементи на статистическите таблици

Думата таблица в широко употребявания смисъл е общо наименование на разнообразни по съдържание построения, които по форма са мрежи от пресичащи се хоризонтални редове и вертикални колони, запълнени с числа и съответно озаглавени. Терминът ***статистически таблици*** обаче има по-специфично съдържание. То се отнася не само и не главно до формата, а преди всичко до същността на това, което се представя в табличен вид.

Статистическите таблици съдържат ***числови статистически характеристики***. В статистически таблици се "отливат" резултатите от статистическата групировка, представят се данните, които подлежат на анализиране, а често и резултатите от статистическия анализ.

Статистическите таблици не трябва да се разглеждат само като удобна форма за сбито и прегледно подреждане на статистически данни, а преди всичко като рационален начин за съчетание на статистически величини с оглед характеризирането на явленията. С основание често се говори не просто за статистически таблици, а за ***табличен метод*** и за таблична статистика.[1]

Всяка статистическа таблица има два вида елементи: формални и логически. ***Формалните елементи*** се отнасят до външната форма на таблицата, до нейния макет. Те са заглавие, заглавен ред, челна колона, редове, колони и клетки.

[1] В историята на статистиката се откроява като разновидност на нейното описателно направление табличната статистика, чиито представители са датският статистик ***Й. Анхерсен (1700-1765)*** и руският статистик ***Ив. Кирилов (1689-1737)***.

Логическите елементи на статистическата таблица се отнасят до нейното вътрешно съдържание. Всяка статистическа таблица съдържа някакви числови характеристики, изложени по такъв начин, че при четенето им могат да се правят съждения за характеризираните явления. Езиковата "обвивка" на тези съждения са логически построените изречения, които мислено или гласно се произнасят при четене на таблицата.

Простите и сложните съждения, които се съставят при четенето на дадена таблица, се отнасят за някакви съвкупности, или групи от единици, за които са дадени съответни характеристики. Същността, вътрешното съдържание на статистическата таблица са именно съжденията, представени чрез числа и "облечени" в определена външна форма, която образува макета. Поради това като логически елементи на статистическата таблица се определят ***статистическият субект*** (статистическият подлог) и ***статистическият предикат*** (статистическо сказуемо).

Статистическият субект са онези съвкупности, групи, териториални единици и др., които са представени в таблицата и за които са дадени съответни характеристики.

Статистическият предикат са онези признаци, обобщаващи характеристики и др., чрез които се описва, характеризира се статистическия субект.

2.8.2. Видове статистически таблици

Статистическите таблици са разнообразни по съдържание, форма и конкретно предназначение. Разграничаването им по вид може да се извърши от различни гледни точки.

В зависимост от това, дали съдържат първични статистически данни, получени непосредствено от наблюдението и групировката, или специално подбрани и целесъобразно подредени данни за решаване на определена изследователска или друга задача, статистическите таблици биват ***първични*** и ***производни***.

Според това, дали явленията се характеризират в тяхното статично състояние, или се представя развитието им във времето, таблиците биват ***статични*** и ***динамични***. Особена комбинация на статични и динамични таблици са ***балансовите таблици***, в които се представя състоянието на дадена съвкупност (понякога разчленена на части) към определен момент,

който е начален за разглеждания период, единиците, които излизат от съвкупността през периода (изходящ поток), единиците, които влизат в съвкупността през същия период (входящ поток), и състоянието на съвкупността в края на периода. По такъв начин балансовите таблици съдържат две моментни и две периодни съвкупности.

Важно значение има разграничението на таблиците според строежа на статистическия субект. От тази гледна точка те се делят на ***прости*** и ***сложни***.

Прости статистически таблици са тези, чийто статистически субект по своя строеж е прост - представлява само наименование на териториални поделения или други обекти, като например наименованието на области, общини, отрасли, ведомства, държави и др.

Сложни статистически таблици са тези, в които статистическият субект се състои от групи, подсъвкупности, обособени по определени признаци, по една или повече скали.

Статистическият предикат също може да има различен строеж и различно съдържание. В един случай предикатът съдържа признаци или обобщаващи характеристики, които дават само най-обща информация за състоянието или развитието на явленията. Съжденията, които могат да се направят в този случай са прости. С тях само се описва статистическият субект. В друг случай статистическият предикат може да се състои от свързани помежду си признаци и др., които дават възможност не само да се опише явлението, изразено в статистическия субект, но да се разкрият и вътрешните връзки, зависимости и др., т.е. да се анализира явлението, да се направят изводи, заключения. В първия случай таблиците са ***описателни***, а във втория - ***аналитични***. Аналитични са например корелационните таблици, в които се представя разпределението на единиците на съвкупностите по значенията на два признака едновременно, между които съществува зависимост (вж. гл. 8).

2.9. Статистически величини

Чрез статистическото изучаване се получават ***статистически величини***. Всяка статистическа величина получава при дадено изучаване конкретен израз в определено число.[1] Статистическите величини могат да

[1] Често се допуска смесване на понятията величина, число и цифра. Величините са количествени дадености от определен вид с определено съдържание. Те се изразяват конкретно с определени числа. Цифрите са само условни писмени знаци за означаване на числата.

бъдат различни по своето съдържание, по форма и по начина на получаване. Основно те биват: абсолютни, относителни и средни статистически величини.

В тази глава се разглеждат абсолютните и относителните статистически величини. Средните величини се разглеждат в гл.3.

2.9.1. Абсолютни статистически величини

Абсолютните статистически величини изразяват размера (обема, равнището и др.) на изучаваните явления в дадени времеви и пространствени граници, измерени чрез съответна мярка съобразно натурално-веществената им форма и физическите им качества. Според съдържанието им се делят на обеми, равнища и маси.

Обемите са количества единици на съвкупности - брой на заетите лица в промишлеността към определен момент, продадени количества от даден вид стока през 2005 г. и др.

Равнищата са значения на признаци на статистическите единици или средни значения на признаци за определени съвкупности, като например цени на продадените стоки, себестойност на единица изделие, производителност на труда средно на един зает в дадена фирма и др.

Масите са такива абсолютни величини, които се формират от обеми и равнища, например износът в млн. лв. през 2007 г. се е формирал от продадените (изнесените) количества стоки (обеми) и техните износни цени (равнища).

Абсолютните величини могат да бъдат единични (елементарни, индивидуални) и агрегатни (сумарни).

Единичните (елементарните, индивидуалните) абсолютни величини се отнасят за отделни единици.

Агрегатните (сумарните) абсолютни величини се отнасят за цели съвкупности или за техни части (подсъвкупности, групи) и се получават по пътя на агрегирането, обобщаването на единичните величини.

Особен вид абсолютни величини са ***условните абсолютни величини***. Те се получават чрез привеждането към избрана условна единица, посредством определени преводни коефициенти или по друг начин. Например броят на тракторите в селското стопанство се привежда в брой условни трактори с мощност 15 конски сили; товарните вагони се привеждат в условни двуосни вагони.

Абсолютните величини са винаги наименовани, изразени са в съответна мярка.

2.9.2. Относителни статистически величини

Относителните статистически величини се получават като отношения между абсолютни величини, а в някои случаи, и като отношение между други относителни величини.

При изчисляване на относителни величини аритметичното действие е деление. По принцип могат да се делят една на друга както едноименни (изразени в една мярка), така и разноименни (изразени в различни мерки) величини. Във всеки случай обаче, между статистическите величини трябва да има смислена връзка - частното, което ще се получи, да има определен познавателен смисъл. Освен това, трябва строго да се съблюдават изискванията за реална съпоставимост на величините от гледна точка на обхвата, методологията и други, които бяха разгледани в т. 2.6.2.

Според характера и познавателното им значение относителните величини са: планово-прогнозни; структурни; динамични (хронологични); териториални; интензивни; координационни.

1. Планово-прогнозните относителни величини характеризират планирани или прогнозирани изменения спрямо определен период, както и степента на постигнатото изпълнение на планираното (прогнозираното). Изразяват се обикновено в проценти.

2. Структурните относителни величини характеризират структурата (строежа) на съвкупностите. Те се прилагат широко в емпиричните изследвания. Това се обяснява преди всичко с интереса към структурата на изучаваните явления и нейното изменение. Делят се на разчленителни и съотносителни.

Разчленителните структурни относителни величини са отношения на частите към цялото и изразяват относителния дял на отделните части в тяхната обща сума, приета за 1 или 100. Разчленителни са например относителните величини, представляващи отношение на засетите площи с пшеница, ечемик, царевица и др. към общия размер на площите, засети със зърнени култури.

Съотносителните структурни относителни величини се получават като отношение на частите на цялото към тази част, която от някаква гледна точка се третира като основна или особено важна. Например

отношението на отделните групи помощен медицински персонал към броя на лекарите.

3. Динамичните (хронологичните) относителни величини характеризират относителни изменения във времето, т.е. темповете на развитие. Те се наричат още динамични (хронологични) индекси. Получават се като отношение на абсолютни величини (обем, равнище и др.) през даден период към абсолютни величини (обем, равнище и др.) през друг период (базов). Например относителната величина, получена като се отнесе цената на единица стока от даден вид през февруари 2008 г. към цената на същата стока през януари 2008 г.

Когато се измерват относителни изменения, отнасящи се за съвкупности, конструират се по съответен начин множествени (сложни, съвкупностни) индекси. Те се разглеждат в гл. 11.

4. Териториалните относителни величини характеризират относителни териториални различия. Абсолютната величина (на равнище, обем, маса) за даден район, се отнася към съответната абсолютна величина за района, с който се прави сравнението (базов район).

Тези относителни величини се наричат още ***териториални индекси*** (единични и множествени). Особеностите на множествените (сложните) териториални индекси се разглеждат в гл. 11.

5. Интензивните относителни величини представляват отношения на определени съвкупности, които произлизат от други съвкупности. Съпоставяните съвкупности се намират във връзка помежду си, като среда и резултат, произлязъл от тази среда. Тези относителни величини показват всъщност с каква интензивност възниква съвкупността-резултат от съвкупността-среда. Те имат широко приложение в статистиката и в много емпирични икономически, социални и други изследвания. По характер интензивни относителни величини са редица демографски коефициенти: на раждаемост, на смъртност, на брачност и др.

Интензивните относителни величини се подразделят на генерални (брутни) и специфични.

Генерални са интензивните относителни величини, когато съвкупността-резултат се отнася към такава съвкупност-среда, не всички единици на която участвуват във формирането на резултата. Ако например броят на родените деца през дадена година се раздели на средния брой на цялото население през годината, ще се получи генерална (брутна) интензивна относителна величина - общ или брутен коефициент на раждаемост. Ако същият брой родени деца се отнесе само към средния

средния брой на жените, образуващи родилен (фертилен) контингент (на възраст от 15 до 49 г.), ще се получи ***специфична относителна интензивна величина*** - специфичен коефициент на раждаемост.

Както генералните, така и специфичните интензивни относителни величини могат да бъдат общи и частни (групови).

Общите интензивни относителни величини се отнасят за цели съвкупности, а ***частните*** - за обособени по някакъв признак групи.

Изчисляването на генерални (брутни) и специфични, на общи и частни (групови) интензивни относителни величини разширява възможностите за анализ. Може например да се установи доколко върху общия генерален (брутен) коефициент на раждаемост влияят повъзрастовите (частните) коефициенти на раждаемост, възрастовата структура на родилния контингент и относителният дял на родилния контингент в цялото население.

6. Координационните относителни величини са отношения на различни по своя характер абсолютни величини, които са независими една от друга. Връзката между тях не е нито връзка между части и цяло, нито между среда и резултат. И все пак отношението между тях има определен познавателен смисъл. Всъщност тук се има предвид отношение между различни съвкупности, които в математическия смисъл не се намират във връзка, но връзката между тях е логическа. Разширяването или "свиването" на едната съвкупност изменя в някаква степен условията, при които съществува другата.

Ако например се раздели произведената електроенергия в страната през 2007 г. на средния брой на населението, ще се получи координационна относителна величина. Такъв вид относителни величини са още: брутният вътрешен продукт средно на лице от населението; леглата в болничните заведения на 10 000 души от населението и др.

2.10. Практикум

2.10.1. Въпроси за самопроверка

1. Какъв вид статистическо изучаване е преброяването на населението?
2. Какво значи критичен момент при статистическото наблюдение?
3. Кой начин на наблюдение се нарича експедиционен?

4. Какво означава аритметичен принцип при съставянето на интервална скала?
5. Коя скала се нарича рангова?
6. Кои динамични редове се наричат стационарни?
7. Коя групировка е вариационна?
8. Какво означава статистически предикат в една статистическа таблица?
9. Каква е разликата между разчленителните и съотносителните структурни относителни величини?
10. Каква относителна величина са студентите на 1000 души от населението?
11. Каква относителна величина е коефициентът на брачност, т.е. сключените бракове през даден период на 1000 души от населението?

2.10.2. Задачи за упражнение

Задача 1. Направено е наблюдение на 53 фирми. Между наблюдаваните признаци е признакът възраст на ръководителя на фирмата. Установено е, че най-младият ръководител е на 30 години, а най-възрастният – на 69 години. Необходимо е да се направи интервална скала по аритметичен принцип за разпределение на ръководителите по възраст, като се формират 8 възрастови групи.

Задача 2. При статистическия контрол на качеството на продукцията в едно предприятие е установено, че 5 % от проверените изделия са нестандартни (останалите 95 % са стандартни). Необходимо е да се състави дихотомна (бинарна) скала за това разпределение.

2.10.3. И статистиците се шегуват

Бележитият български учен проф. д-р ***Асен Златаров*** беше казал: "Смехът е кислород на живота. Той е културна придобивка; робът не се смее, дивакът – също. Смехът е велико творческо начало." Статистиците споделят тези мисли. Но те твърдят, че има „неоспорими" данни, които „доказват" вредата и дори фаталните последици от консумацията на краставици.

1. Почти 100 % от хората, които страдат от хронични заболявания, са консумирали краставици.

2. Около 99 % от хората, умрели от рак, през живота си са консумирали краставици.

3. 100 % от войниците, загинали в сражения, през живота си са консумирали краставици.

4. 99,7 % от всички лица, станали жертви на автомобилни или самолетни катастрофи, са употребявали краставици през последните шест месеца преди злополуката.

5. 93,1 % от всички непълнолетни престъпници произлизат от семейства, в които са консумирани краставици.

6. Сред хората, родени през 1875 г., които са яли краставици, смъртността е 100 %.

7. Почти всички живи лица, родени през 1910-1920 г., които са консумирали краставици, имат набръчкана кожа, отслабнало зрение и изпадали зъби.

8. Експериментално е установено, че морски свинчета, които принудително са хранени с краставици в продължение на един месец, са загубили апетита си.

Единственият начин за предпазване от вредното действие на краставиците е да се заменят с безвредни зеленчуци, каквито са например блатните орхидеи.[1]

[1] Вж. Физиците продължават да се шегуват. С., 1983, с. 227.

3. ЕМПИРИЧНИ СТАТИСТИЧЕСКИ РАЗПРЕДЕЛЕНИЯ

> "Статистическите числа не управляват света, но те показват как светът се управлява."
>
> Й. В. Гьоте

Читателят ще се запознае от тази глава със смисъла и съдържанието на едно основно статистическо понятие "разпределение" и в частност с разпределенията при емпиричните изследвания, както и с техните обобщаващи характеристики. Той ще разбере какъв е познавателният смисъл на различните видове средни величини, на измерителите на статистическата вариация и на коефициентите, характеризиращи формата на разпределенията. Тези знания дават възможност на специалиста, на изследователя, на бизнесмена и др. да прави верен избор на адекватния измерител, правилно да го интерпретира и да се предпазва от грешки.

3.1. Обща постановка

В предходните глави многократно беше използуван терминът ***разпределения***. От контекста читателят е получил представа за смисловото съдържание на този термин. Тук е необходимо да се направят допълнителни уточнения и да се разгледат някои видове и форми емпирични разпределения с оглед на изложението в следващите глави.

При разглеждането на статистическата групировка беше установено, че единиците (случаите) на статистическата съвкупност се разпределят на групи по значенията на групировъчните признаци по определени скали. Според вида на скалите се получават различни разпределения. В тази глава се разглеждат разпределения на статистическите единици по значенията на вариационни признаци, т.е. по интервални скали. Те следователно се представят във вариационни редове.

Тъй като това са разпределения, получени при конкретни, емпирични статистически изучавания, те се наричат ***емпирични разпределения***, за да се разграничат от теоретичните разпределения, които се разглеждат в гл. 4. Когато разпределението е направено по един признак (една скала), то се нарича ***едномерно***. Когато е направено по два признака е ***двумерно***, а при повече признаци (скали) - ***многомерно***.

В тази глава по-нататък се разглеждат едномерните разпределения. Пример на такова едномерно разпределение по интервална скала се съдържа в табл. 3.1.

Таблица 3.1

Разпределение на заетите лица във фирма "Н" по размер на месечната работна заплата през м. октомври 2008 г.

Групови интервали – лв.	Брой на заетите (абсолютни честоти)	Относителен дял на заетите в общия брой (относителни честоти)	Кумулативни честоти	Абсолютна плътност на разпределението
	f	$v = \frac{f}{\sum f} 100$	C	e
500 - 540	10	4,7	10	0,250
540 - 580	19	9,0	29	0,475
580 - 620	27	12,8	56	0,675
620 - 660	38	18,0	94	0,950
660 - 700	46	21,8	140	1,150
700 - 740	35	16,6	175	0,875
740 - 780	24	11,4	199	0,600
780 - 820	12	5,7	211	0,300
	211	100,0	x	x

Единиците, които се съдържат в отделните групи, т.е. попадат в отделните групови интервали, се наричат ***честоти***. Когато са в абсолютни величини, те се наричат ***абсолютни честоти***, а когато са в

относителни величини – ***относителни (релативни) честоти***. Затова емпиричните разпределения се наричат още ***честотни разпределения***, за разлика от теоретичните разпределения, които са разпределения на вероятности, т.е. те са ***вероятностни разпределения***.

В някои случаи при статистическия анализ е необходимо разпределението да се представи така, че всеки групов интервал да обхваща честотите на всички предходни или на всички следващи. В примера (табл. 3.1) това означава колко заети получават заплата до 540 лв., до 580 лв., до 620 лв. и т.н. или колко са получаващите заплати над 500 лв., над 540 лв., над 580 лв. и т.н. Така получените честоти се наричат ***кумулативни***. В първия случай те са прогресивно-кумулативни или само кумулативни, а във втория случай - регресивно-кумулативни.

Когато скалата не е с еднаква ширина на интервалите, непосредственото сравняване на абсолютните или на релативните честоти се затруднява и не може да се получи вярна представа за разпределението, тъй като честотите се отнасят за значения на признака, представени в различен диапазон, и следователно зависят от различията в ширините на интервалите. Ето защо абсолютните, респ. относителните честоти, се редуцират към единица интервал. По такъв начин се получава ***плътността на разпределението*** - абсолютна и относителна.

Абсолютната плътност на разпределението се получава, като се разделят абсолютните честоти на съответните ширини на интервалите (вж. табл. 3.1).

Относителната плътност на разпределението има същия смисъл, но по отношение на релативните (относителните) честоти. Тя се получава, като се разделят относителните честоти на ширините на интервалите.

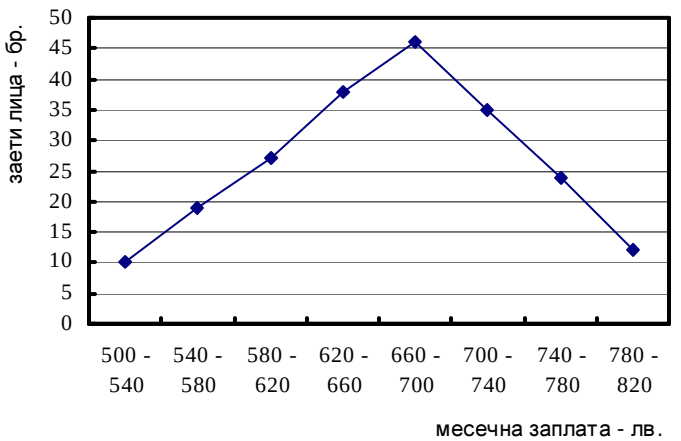
Емпиричните разпределения могат да се представят чрез техния графичен образ (графичен модел). За такъв при едномерните разпределения служат обикновено полигонът на разпределението и хистограмата. Те дават визуална представа за формата на разпределенията, която също е предмет на анализ.

Полигонът и ***хистограмата*** се строят в първия квадрант на ортогонална координатна система с помощта на скали по двете оси, образувани с предварително определен мащаб. На скалата по абсцисната

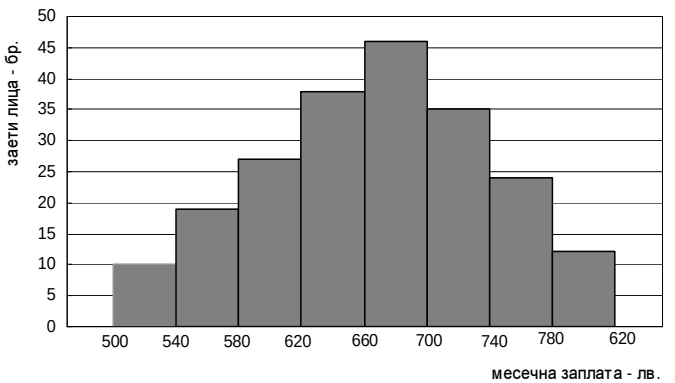
ос се нанасят ширините на груповите интервали, а по ординатната ос се отчитат честотите (абсолютни или относителни).

Полигонът описва формата на разпределението с линия, а хистограмата - с поставени един до друг правоъгълници, чиито основи съответстват на ширините на груповите интервали, а височините - на честотите.

На фиг. 3.1 и фиг. 3.2 е представено разпределението, съдържащо се в табл. 3.1, с полигон и хистограма.



Фиг. 3.1



Фиг. 3.2

Едно от преимуществата на полигона е възможността да се представят едновременно две и повече разпределения, за да бъдат

сравнени, както и да се представи емпирично разпределение и неговият теоретичен модел (съответното теоретично разпределение).

Хистограмата е подходяща при нееднакви по ширина групови интервали.

Възможно е графично да се представят и кумулативните честоти с линия, която се нарича ***кумулата***.

3.2. Типични форми на емпирични разпределения

Масовите явления имат свой вътрешен строеж. За различните конкретни съвкупности той е различен и винаги конкретен по време и място. Затова, ако се представят графично изучаваните емпирични разпределения, ще се получат разнообразни по своята конфигурация графични образи.

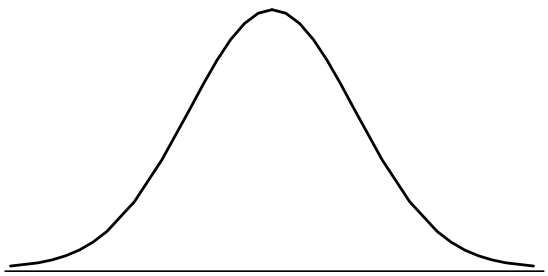
Ако си представим, че при даден интервален вариационен ред (интервална скала) се намалява неограничено ширината на интервалите и честотите са достатъчно големи, то линията, която описва формата на разпределението, вместо начупена (полигон) ще стане гладка, но ще има различна форма. Тя се нарича ***крива на разпределението***.

Всред голямото многообразие могат да се отделят определени ***типични форми на разпределения***. Някои от тях се срещат по-често при емпиричните изследвания, а други служат като теоретични модели на емпиричните разпределения.

1. Равномерно разпределение. Това е разпределение, при което всички значения на признака имат еднаква честота. Графично то може да се представи с линия, успоредна на абсцисната ос. Такова разпределение се среща сравнително рядко - обикновено в експерименталното дело, когато появата на всяко събитие (всяко значение на признака) е еднакво вероятна. Като реално емпирично разпределение, изискващо анализ, то почти няма значение, но като хипотетично разпределение, спрямо което може да се разглежда емпиричното, то е полезно при анализа.

2. Симетрично унимодално (едномодално) разпределение. Описва се от симетрична крива, която има формата на камбана (фиг.3.3). Характерно за него е, че в средата на разпределението честотата е най-голяма, а от двете й страни честотите намаляват постепенно и са най-

малко при минималното и максималното значение на признака. Кривата на това разпределение има един връх (една мода; вж. за модата в т. 3.4). В обективната действителност рядко се среща идеално симетрично разпределение. Някои емпирични разпределения обаче могат да бъдат твърде близки до разглежданото симетрично разпределение. Такова би било например разпределението на голям брой младежи от една възраст по ръст. Или да предположим, че автоматичен струг обработва детайл с определен размер 25 мм с допустими отклонения ±1 мм. Ако стругът е добре настроен и работи нормално, произведените от него детайли биха се разпределяли по размер симетрично с точка на най-голямо натрупване 25 мм. Ако се окаже, че разпределението приема друга форма, това е указание, че освен случайните допустими отклонения, действа и някаква друга причина, например в настройката на машината. Симетрични разпределения от типа на разглежданото се срещат още във физиката, биологията и експерименталното дело. На симетричното емпирично разпределение съответствува нормалното теоретично разпределение (разпределение на Гаус - Лаплас), което има строго математическо описание и важни свойства (вж. гл.4).

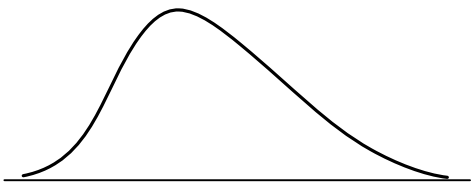


Фиг. 3.3

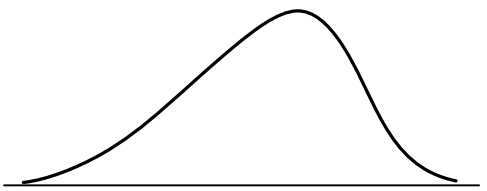
3. Умерено асиметрично разпределение. Описва се от крива, която не е симетрична, а е изтеглена вляво или вдясно от точката на най-голямото натрупване.

При дясно изтеглена крива (дълго дясно рамо) максимална е честотата на значение на признака, по-малко от средното значение (фиг. 3.4).

При ляво изтеглена крива (дълго ляво рамо) най-голямото натрупване (най-голяма честота) е при значение на признака, по-голямо от средното (фиг. 3.5).



Фиг. 3.4

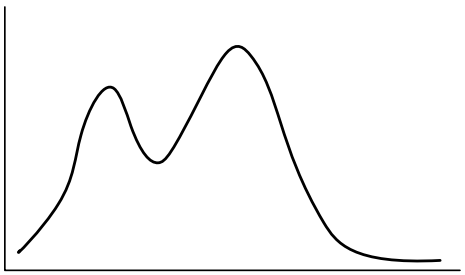


Фиг. 3.5

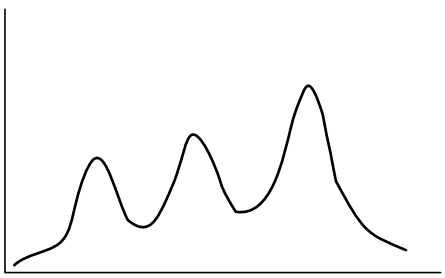
Умерено-асиметрични разпределения се срещат сравнително по-често при емпирични изследвания. Приблизително такава форма на разпределение имат за определен период домакинствата по размер на дохода средно на лице в домакинството.

4. Бимодално (двумодално) разпределение. Характеризирира се с това, че при две значения на признака (два групови интервала) има голямо натрупване на единици (честоти) и поради това кривата има два върха (две моди), между които се образува седловина (фиг. 3.6).

5. Мултимодално разпределение. Кривата му има три или повече върхове (фиг. 3.7).



Фиг. 3.6



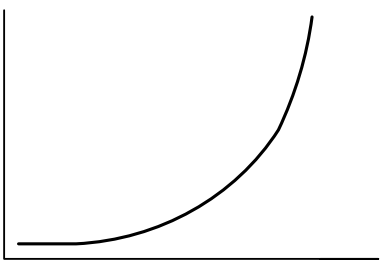
Фиг. 3.7

Когато се получи бимодална или мултимодална форма на разпределение, това е сигнал, че съвкупността не е еднородна по отношение на закономерността на разпределението, че тя вероятно обхваща две или повече съвкупности, всяка от които има своя форма на разпределение.

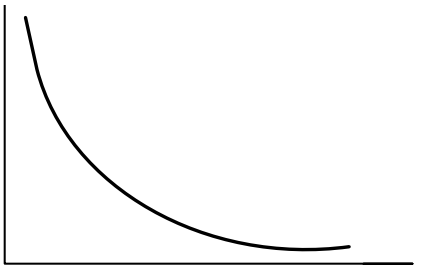
6. Крайно асиметрично разпределение. То има две типични форми:

а) ***J-разпределение (йот-разпределение***, наричано още ***обратно L-разпределение)***. При това разпределение броят на единиците (честотите) е най-малък при най-малките значения на признака, постепенно се увеличава и достига своя максимум при най-големите значения на признака (фиг. 3.8). Близко до J-разпределението е например разпределението на починалите от ракови новообразувания по възрастови групи от населението (на 1000 души от съответната възрастова група).

б) ***Обратно J-разпределение*** (наричано още ***L–разпределение***). При него броят на единиците е най-голям при най-малките значения на признака, постепенно намалява и достига своя минимум при най-големите значения на признака (фиг. 3.9).

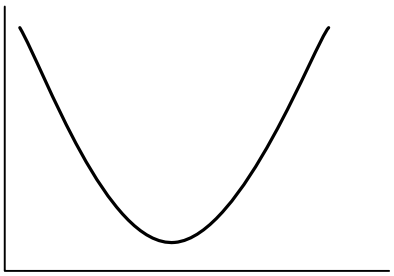


Фиг. 3.8



Фиг. 3.9

7. U-разпределение. Характерно за него е това, че при най-малките значения на признака има голям брой единици, след това единиците постепенно намаляват, достигат своя минимум около средното значение на признака и след това отново се увеличават до най-големия размер. Кривата на това разпределение наподобява латинската буква "U", откъдето е получило и наименованието си (фиг. 3.10).



Фиг. 3.10

Изследването на формите на емпиричните разпределения има голямо значение при статистическия анализ, особено когато се разглеждат не само като статично състояние, а се сравняват разпределения, отнасящи се за едни и същи явления през различни периоди или моменти.

3.3. Основни обобщаващи характеристики на емпиричните разпределения

Графичното представяне на разпределението дава полезна информация за неговата форма. То показва към коя от изложените типични форми или някоя друга може да се отнесе дадено емпирично разпределение. При анализа обаче това не е достатъчно. Необходимо е да се намерят ***обобщаващи числови характеристики на емпиричните разпределения***.

Първо, необходима е обобщаваща числова характеристика за ***центъра на разпределението (централната тенденция)***. Този център може да се характеризира със значението на признака, което се среща най-често в съвкупността (има най-голяма честота), със значението на признака на единицата, която заема централно положение, или с такова значение, което е средно на всички единични значения. Това по същество са различни средни величини.

Второ, много е необходимо при анализа да се измери степента на ***вариацията (разсейването)*** на единиците на съвкупността по значенията на дадения признак. Възможно е две или повече емпирични разпределения да имат еднаква форма и еднакъв център на разпределение, но значенията на признака да варират в различни граници.

Трето, необходими са числови характеристики относно формата на разпределението в две отношения:

а) по отношение на симетрията, т.е. в каква степен даденото емпирично разпределение се доближава или се отклонява по форма от симетричното теоретично разпределение, прието за еталон, или каква е степента на неговата ***асиметрия***;

б) по отношение на максималната ордината на кривата, т.е. дали върхът на кривата на емпиричното разпределение съвпада с върха на стандартното нормално разпределение, дали той е по-висок (кривата е по-изострена), или е по-нисък (по-притъпена), което значи измерване на ***ексцес***.

При конкретния анализ на двумерни и многомерни разпределения са необходими и други обобщаващи числови характеристики. Такива са например характеристиките за силата на зависимостта между два или повече признака, за зависимостта на разсейването по единия признак от разсейването по другия и т.н.

Посочените основни обобщаващи числови характеристики на емпиричните разпределения описват различни техни страни (свойства) и взети общо, във връзка една с друга, дават много информация за изследваните разпределения. Въз основа на тях могат да се правят съждения относно вътрешните закономерности на процеса на формирането на разпределенията. Те са много необходими и когато въз основа на извадки трябва да се направят изводи относно разпределенията в генералните съвкупности.

3.4. Средни величини

Когато се анализират емпиричните разпределения, средните величини са едни от обобщаващите характеристики на тези разпределения. Чрез тях се измерва ***центърът на разпределението*** (централната тенденция). Спрямо средна величина обикновено се характеризира варирането (разсейването) на единиците на съвкупността по значенията на изучавания признак. Има също средни, които изпълняват други функции. Когато например се изследва развитието на явленията, чрез средна величина се измерва средната скорост на това развитие. Във

всички случаи средната величина е ***обобщаваща характеристика*** на съвкупността от единици (случаи).

Подчертавайки, че средната величина е обобщаваща числова характеристика, трябва да се има предвид и разпространената погрешна представа, че тя е някаква представителна единица, притежаваща средни качества. Тази абсурдна представа няма нищо общо с истинската същност на средната. Уместно е да се подчертае и мисълта на бележития белгийски учен-статистик ***Адолф Кетле (1796-1874)***: "Понятието за средна величина съществува извън науката, която само му придава определеност и точност."

3.4.1. Видове средни величини

Средните величини могат да се класифицират от различни гледни точки.

1. Според характера на осредняваните величини те биват ***вариационни*** и ***хронологични***.

Вариационните средни изразяват средни значения на вариационни признаци.

Хронологичните средни изразяват средно значение на тези осреднявани величини, които характеризират състоянието на явленията през определени периоди или към определени моменти от времето. Те следователно се изчисляват от динамични (хронологични) редове.

2. В зависимост от това, дали се отнасят за цяла съвкупност или за обособени подсъвкупности (групи) на съвкупността, средните биват ***общи*** и ***групови***.

3. Според това, дали се изчисляват за генерални съвкупности, или за извадки, излъчени по принципите на репрезентативното изучаване, средните величини биват ***средни на генерални съвкупности*** и ***средни на извадки.***

4. В зависимост от това, дали се изчисляват от всички дадени значения на признака, респ. от всички членове на динамичния ред чрез определени алгебрични действия, или се определят от положението, което заемат в редовете, средните величини се делят на ***алгебрични средни*** и ***позиционни средни***.

В зависимост от формата на осредняването, свързана с определени алгебрични операции, ***алгебричните средни*** биват: средна ***аритметична***, средна ***хармонична***, средна ***геометрична***, средна ***квадратична***, средна ***кубична*** и др. При статистическия анализ се използуват обикновено първите четири средни.

Позиционните средни се определят според положението, което заемат в статистическия ред. Това са, от една страна, ***медианата*** и свързаните с нея ***квартили***, ***квинтили***, ***децили*** и ***центили***, обединени в общото понятие квантили или градиенти, а от друга страна - ***модата***.

Всяка средна има свое познавателно значение, свои особености и области на приложение. Във връзка с емпиричните разпределения е необходимо да се подчертае, че средната аритметична, медианата и модата, изчислявани от вариационни редове, са измерители на ***центъра на разпределенията***, наричан често в литературата ***централна тенденция***.

3.4.2. Средна аритметична величина

Средната аритметична е най-често употребяваната средна величина в статистиката. Тя е в такава степен популярна, че когато се говори за средна величина, най-често се има предвид тази средна. Нейната популярност и широко приложение се обясняват преди всичко с това, че по своето съдържание и смисъл отговаря на познавателните задачи, които най-често се поставят при осредняването.

Средната аритметична се изчислява както за вариационни признаци (вариационна аритметична средна) и представлява център на разпределението, така и за динамични редове (хронологична аритметична средна). Тук ще бъдат разгледани общите въпроси относно същността, формата, свойствата и др. на средната аритметична, имайки предвид главно нейното приложение при осредняване значенията на вариационни признаци, т.е. като център на разпределението. Специфичните особености и изчисляването на хронологичната аритметична средна се разглеждат в гл. 10.

Ако отделните значения на признака се означават с x_1, x_2, x_3, ..., x_N, а броят им - с N, изчисляването на средната аритметична ($\bar{x}$) може да се представи така::

$$(3.1) \qquad \bar{x} = \frac{x_1 + x_2 + x_3 + ... + x_N}{N}.$$

Ако за означаване на сума се приложи общоприетият знак - гръцката буква Σ (сигма), формулата може да се запише в следния по-кратък вид:[1]

$$(3.2) \qquad \bar{x} = \frac{\sum x}{N}.$$

Това е формулата на ***непретеглената средна аритметична.*** Нарича се непретеглена, защото при изчисляването ѝ се вземат отделните значения на признака без да се придава някакво тегло, т.е. без да се умножават по други величини. От формулата следва, че $N\bar{x} = \sum x$, което означава, че като се заместят конкретните осреднявани величини със средната аритметична, ще се получи същата обща сума Σx (определящо свойство).

Ако отделните значения на признака се срещат при различен брой единици, т.е. дадено е разпределение със съответни честоти, посочената формула е неприложима. Необходимо е всяко значение на признака да се претегли (умножи) с честотите, т.е. да се вземе толкова пъти, колкото пъти се среща в съвкупността.

Ако честотите (теглата) се означат f_1, f_2, f_3, ..., f_N, формулата на ***претеглената средна аритметична*** ще има вида:

$$(3.3) \qquad \bar{x} = \frac{\sum xf}{\sum f}.$$

Когато не са дадени отделните значения на признака, а са в групови интервали, за осреднявани величини (x) се приемат средите на интервалите.

[1] Тъй като се разбира, че се сумират всички значения на x, тук и в следващите формули не се означават границите на сумирането към Σ.

Да приемем, че е дадено разпределението на заетите лица в една фирма, съдържащо се в табл. 3.1. Необходимо е да се изчисли средната работна заплата като средна аритметична величина. Данните и изчисленията са дадени в табл. 3.2.

Таблица 3.2

Разпределение на заетите лица във фирма "Н" по размер на месечната работна заплата през м. октомври 2005 г.

Групови интервали – лв.	Среди на интервалите	Брой на заетите (абсолютни честоти)	Произведения на средите на интервалите и честотите
	x	f	xf
500 - 540	520	10	5200
540 - 580	560	19	10640
580 - 620	600	27	16200
620 - 660	640	38	24320
660 - 700	680	46	31280
700 - 740	720	35	25200
740 - 780	760	24	18240
780 - 820	800	12	9600
		211	140680

Претеглената средна аритметична, изчислена по формула 3.3 е:

$$\bar{x} = \frac{\sum xf}{\sum f} = \frac{140680}{211} = 666{,}73\,лв\,.$$

Това е център на даденото разпределение, около който се намират конкретните заплати. Казано по друг начин, ако общата изплатена сума от 140680 лв би била разпределена поравно, то заплатата на всеки един би била 666,73 лв, т.е.

$$\sum xf = \bar{x}\sum f \quad \text{или } 140680 = 666{,}73 \,.\, 211.$$

В примера за изчисляване на претеглена средна аритметична като тегла са дадени честотите, т.е. броят на единиците на съвкупността (заетите лица), разпределени по значенията на изучавания признак (заплата). Тук няма никакво съмнение, че именно с тези тегла трябва да се изчисли претеглената средна аритметична. Има обаче случаи, при които теглата трябва да се определят по пътя на разсъждения и преценки относно природата на изучаваните явления и познавателния смисъл на изчисляваната средна аритметична. Това важи особено при изчисляване на средна аритметична от средни и относителни величини. Погрешният избор на теглата може да измени смисъла на изчисляваната средна величина и да я лиши от истинското й съдържание. Този избор на теглата не може да бъде произволен. Той трябва да бъде логически обоснован. ***Като правило теглата при изчисляването на средната аритметична (адекватните тегла) са онези единици (честоти - абсолютни или относителни), които са носители на признака, чиито значения се осредняват*** и към съвкупността на които е адресирана изчисляваната средна аритметична. Ако трябва да се изчисли среден брой заети лица, падащи се на ***1 предприятие***, средната трябва да се изчисли с тегла броят на предприятията; ако се изчислява среден добив на пшеница от ***1 декар***, за тегла ще се вземат съответните реколтирани площи в декари и т.н.

В известни случаи (особено при изчисляване на средната аритметична от относителни величини, като например при изчисляването на средните аритметични индекси), е възможно от едни и същи осреднявани величини да се изчисляват претеглени средни аритметични с различни тегла. Но и тогава теглата трябва да отговарят на посоченото правило за определяне на адекватните тегла.

Между средните аритметични величини, изчислени от едни и същи осреднявани величини, но с различни тегла, обикновено съществува разлика независимо от това, че и едните, и другите тегла могат да имат напълно определен смисъл. И все пак, въпреки различията в теглата, при определени условия двете средни могат да съвпаднат по числова стойност.

При статистическия анализ е много важно да се знаят обстоятелствата, от които зависят различията между претеглените с различни тегла средни аритметични.[1]

Без да разглеждаме подробно този въпрос и без да привеждаме математически доказателства, ще формулираме някои ***основни положения и изводи***, които имат значение за правилно разбиране на важни методологични положения в други части на книгата.

Да приемем, че осредняваните величини са $x_1, x_2, x_3, ..., x_N$ и от тях са изчислени две претеглени средни аритметични, от които едната с тегла $f_1, f_2, f_3, ..., f_N$, а другата с тегла $m_1, m_2, m_3, ..., m_N$. Формулите на двете средни ще имат вида:

$$\bar{x} = \frac{\sum xf}{\sum f}; \qquad \bar{x} = \frac{\sum xm}{\sum m}.$$

За ***съотношението между две претеглени средни аритметични с различни тегла*** е доказано следното равенство:

$$\frac{\sum xf}{\sum f} : \frac{\sum xm}{\sum m} = 1 + V_x . V_{\frac{f}{m}} . r_{x\frac{f}{m}}, \qquad (3.4)$$

където:

V_x е коефициентът на вариацията на осреднените величини, който характеризира степента на различията между тях;

$V_{\frac{f}{m}}$ - коефициентът на вариацията на съотношението между различните тегла, който по същество характеризира степента на различията между структурата на теглата *f* и на теглата *m*;

$r_{x\frac{f}{m}}$ - коефициентът на линейната корелация между осредняваните величини и съотношенията между различните тегла, т.е. линейната зависимост между осредняваните величини $x_1, x_2, x_3, ..., x_N$ и съотношенията на теглата $\frac{f_1}{m_1}, \frac{f_2}{m_2}, \frac{f_3}{m_3}, ..., \frac{f_N}{m_N}$. (По-подробно относно коефициентите на

[1] Според изследванията на В. Цонев това съотношение е разгледано за първи път в чисто математически план в: ***Bowley, A.,*** Elements of Statistics, London, 1905 (Вж. ***Цонев, В.,*** По повод на статията "Структури и ефекти", *Икономическа мисъл*, 1985, кн.10.)

вариацията вж. т.3.5.4, а относно коефициента на линейната корелация - гл. 8. В дадения случай е важно не изчисляването на тези коефициенти, а техният логически смисъл).

От формулата следва, че разликата между различно претеглените средни аритметични се дължи на вариацията (различията) на осредняваните величини, на вариацията (различията) в структурата на теглата и на линейната зависимост между осредняваните величини и съотношението на теглата.

Ако поне един от трите множителя отдясно на равенството е равен на нула, отношението на двете средни ще бъде равно на единица, т.е. двете средни ще бъдат еднакви. Иначе казано, за да бъдат еднакви двете средни аритметични, изчислени с различни тегла, трябва да е налице поне едно от следните три условия:

1) ако осредняваните величини са еднакви, т.е. $x_1 = x_2 = x_3 = ... = x_N = const$, тогава средната аритметична е равна на постоянната величина x, а $V_x = 0$;

2) ако структурата (съотношението) на теглата f напълно съвпада със структурата на теглата m, т.е. $\frac{f_1}{m_1} = \frac{f_2}{m_2} = \frac{f_3}{m_3} = ... = \frac{f_N}{m_N} = const$, тогава $V_{\frac{f}{m}} = 0$;

3) ако липсва линейна зависимост между значенията на x и съотношенията на теглата $\frac{f}{m}$, тогава $r_{x\frac{f}{m}} = 0$.

Ако не е налице нито едно от трите условия, двете средни аритметични ще се различават. Абсолютният размер на разликата зависи от абсолютните стойности на трите множителя отдясно на равенството. Тъй като коефициентите на вариацията са винаги положителни, коя от двете средни ще бъде по-голяма зависи от алгебричния знак на коефициента на линейната корелация. Този коефициент може да приема различни стойности в границите от 0 до ± 1.

Когато $r_{x\frac{f}{m}} > 0, \frac{\Sigma xf}{\Sigma f} > \frac{\Sigma xm}{\Sigma m}$, и обратно, когато $r_{x\frac{f}{m}} < 0, \frac{\Sigma xf}{\Sigma f} < \frac{\Sigma xm}{\Sigma m}$.

Коефициентът на линейната корелация ще бъде положителен и следователно $\frac{\Sigma xf}{\Sigma f} > \frac{\Sigma xm}{\Sigma m}$, когато отношенията $\frac{f}{m}$ са по-големи, при по-големите значения на *x* и по-малки при по-малките значения на *x*. Обратно, коефициентът на линейната корелация $r_{x\frac{f}{m}}$ ще бъде отрицателен и следователно $\frac{\Sigma xf}{\Sigma f} < \frac{\Sigma xm}{\Sigma m}$, когато на по-големите значения на *x* отговарят по-малки $\frac{f}{m}$, а на по-малките значения на *x*, по-големи $\frac{f}{m}$.

Формула 3.4 може да се адаптира и за съотношението между претеглена и непретеглена средна аритметична.

$$\frac{\sum xf}{\sum f} : \frac{\sum x}{N} = 1 + V_x . V_f . r_{xf} , \qquad (3.5)$$

където:

V_x е коефициентът на вариацията на осредняваните величини, както във формула 3.4;

V_f - коефициентът на вариацията на теглата;

r_{xf} - коефициентът на линейната корелация между осредняваните величини *x* и теглата *f*.

Тези формули за съотношението между различно претеглени средни аритметични и между претеглена и непретеглена средна аритметична имат съществено значение за измерване и оценяване влиянието на определени структури и структурни изменения върху формирането и динамиката на обобщаващи икономически, социални и др. характеристики.

Средната аритметична (претеглена и непретеглена) притежава ***свойства***, които могат да се дефинират и като теореми. Ще посочим само някои от тях.

Първо. Сумата от разликите (отклоненията) между осредняваните величини (значенията на признаците) и тяхната средна аритметична е винаги равна на нула:

$$\sum (x - \bar{x}) = 0 ; \quad \sum (x - \bar{x}) f = 0 . \qquad (3.6)$$

Второ. Сумата от квадратите на разликите (отклоненията) между осредняваните величини (значенията на признаците) и тяхната средна аритметична е минимум, т.е. число, винаги по-малко, отколкото е сумата от квадратите на разликите между осредняваните величини и всяко друго постоянно, произволно избрано число (A), различно от средната аритметична.

$$(3.7) \qquad \sum (x-\bar{x})^2 < \sum (x-A)^2 \;\; ; \; \sum (x-\bar{x})^2 f < \sum (x-A)^2 f \; .$$

Тези две свойства откроявят най-ярко средната аритметична като център на разпределението.

Трето. Ако теглата при изчисляването на средната аритметична се разделят или се умножат на постоянно, произволно избрано число (A), средната аритметична не се изменя. От това следва, че средната аритметична не зависи от абсолютните стойности на теглата, а от съотношението между тях. Затова като тегла могат да се използват относителните (релативните) честоти, тъй като за постоянно число може да се вземе сумата на абсолютните честоти (Σ*f*).

$$(3.8) \qquad \bar{x} = \sum \left(x \frac{f}{\sum f} \right).$$

В случая сумата на теглата (относителните честоти) е единица. Ако $\frac{f}{\sum f}$ се означи с *v*, формулата на средната аритметична ще приеме вида:

$$(3.9) \qquad \bar{x} = \sum xv \; .$$

Ако относителните честоти са представени в %, ($v = \frac{f}{\sum f} 100$), формулата на средната аритметична ще бъде:

$$(3.10) \qquad \bar{x} = \frac{\sum xv}{100} \; .$$

Това свойство създава практически удобства при изчисляването на средната аритметична.

3.4.3. Средна хармонична величина

В редица случаи в зависимост от разполагаемите данни се налага използуването на друг вид алгебрична средна, наречена ***средна хармонична величина.***

Пример. За да се установи средният разход на работно време за изработването на ***единица продукция*** от даден вид през първия работен час на един работен ден, са наблюдавани трима работници. Първият е изработвал 1 изделие за 3 минути, вторият - за 4 , а третият - за 6 минути.

Ако се изчисли средната аритметична, ще се получи

$$\frac{3+4+6}{3}=\frac{13}{3}=4,33 \text{ мин.}$$

Това обаче не може да бъде средният разход на работно време за производството на единица продукция, тъй като сумата от разхода на работно време е разделена на броя на работниците. Логично е, ако разполагаме с данни за произведената продукция и общо употребеното работно време за нейното произвеждане, средният разход на работно време за единица продукция да се намери като отношение на употребеното работно време към произведената продукция. Тези данни липсват, но биха могли да се изчислят от данните в условието.

Знае се, че е наблюдавана работата на работниците в продължение на 1 работен час, т.е. в продължение на 60 минути. Всеки работник за 1 минута е произвеждал: първият - 1/3 бр., вторият - 1/4 бр., третият - 1/6 бр. За 60 минути съответно са получавани

1/3 . 60 = 20 бр., 1/4 . 60 = 15 бр. и 1/6 . 60 = 10 бр.

При тези данни средният разход на работно време за единица продукция може да се изчисли като средна аритметична:

$$\frac{3.20+4.15+6.10}{20+15+10}=\frac{60+60+60}{45}=4 \text{мин.}$$

Тъй като 20 = 1/3 . 60, 15 = 1/4 . 60 и 10 = 1/6 . 60, изчисленията могат да се запишат така:

$$\frac{60+60+60}{\frac{1}{3}.60+\frac{1}{4}.60+\frac{1}{6}.60}=\frac{60(1+1+1)}{60(\frac{1}{3}+\frac{1}{4}+\frac{1}{6})}=\frac{3}{\frac{1}{3}+\frac{1}{4}+\frac{1}{6}}=\frac{3}{0,75}=\frac{180}{45}=4 \text{мин.}$$

Вижда се, че същата средна може да се изчисли направо по данните от условието, без предварително отделно да се преизчислява количеството произведена продукция. При този случай обаче осредняването се извършва по друга форма на средната величина - ***средната хармонична***. Тя в примера е непретеглена. При N осреднявани величини $x_1, x_2, x_3, ..., x_N$, непретеглената средна хармонична ($\bar{x}_h$) може да се представи с формулата:

$$\bar{x}_h = \frac{N}{\sum \frac{1}{x}} \quad . \tag{3.11}$$

Нека се върнем към примера и да приемем, че две групи работници произвеждат еднакви изделия. В първата група при общо отработени 240 минути, за 1 изделие са изразходвани 8 минути. Във втората група при общо отработени 600 минути, за 1 изделие са изразходвани 12 минути. В такъв случай средният разход на работно време за единица продукция е:

$$\bar{x}_h = \frac{240+600}{\frac{1}{8}\cdot 240 + \frac{1}{12}\cdot 600} = \frac{840}{80} = 10{,}5 \text{ мин.}$$

Това е ***претеглената средна хармонична***.

Ако този среден разход на работно време се изчислява като средна аритметична, тя трябва да бъде претеглена с количеството произведена продукция:

$$\bar{x}_h = \frac{8\cdot 30 + 12\cdot 50}{30+50} = \frac{240+600}{30+50} = \frac{840}{80} = 10{,}5 \text{ мин.}$$

От този пример се вижда съществената разлика в характера на теглата на претеглената средна аритметична и претеглената средна хармонична. Вместо теглата на средната аритметична f (бройките изделия) за теглата при средната хармонична се използват xf (отработеното време). Тъкмо тази разлика в теглата определя и различната форма на осредняване.

Ако теглата, с които се изчислява средната хармонична, се означат с f^*, формулата на претеглената средна хармонична ще бъде:

(3.12) $$\bar{x}_h = \frac{\sum f^*}{\sum \frac{1}{x} f^*} .$$

От примера следва, че претеглената средна хармонична в статистиката се прилага тогава, когато са дадени като тегла такива величини, които по своето съдържание са произведения от осредняваните величини, и онези тегла, с които би трябвало да се изчисли претеглената средна аритметична ($f^* = xf$). При използване на правилно избрани, логически обосновани тегла, осредняването и по средната аритметична, и по средната хармонична води до еднакъв резултат, тъй като при $f^* = xf$

$$\frac{\sum f^*}{\sum \frac{1}{x} f^*} = \frac{\sum xf}{\sum \frac{1}{x} xf} = \frac{\sum xf}{\sum f} .$$

При всеки конкретен случай, когато трябва да се изчисли средна величина, като се избира между средната аритметична и средната хармонична, трябва добре да се прецени връзката между осредняваните величини и съответните тегла. Погрешния избор на средната ще доведе до неверен резултат.

Примерите и формулите показват, че средната хармонична в статистиката не е друга обобщаваща характеристика, различна от средната аритметична, а само друга форма на изчисляване, т.е. ***преобразувана форма на средната аритметична***.

3.4.4. Средна геометрична величина

Тази средна се прилага обикновено при осредняване на относителни величини и в частност на темпове на растеж. Нейната непретеглена форма е:

(3.13) $$\bar{x}_g = \sqrt[N]{\Pi x} ,$$

където П е знак за произведение.

При претеглена средна геометрична теглата са степенни показатели, тъй като показват колко пъти дадената величина трябва да се

умножи сама на себе си, щом тя се повтаря толкова пъти. Формулата на претеглената средна геометрична в такъв случай приема вида:

$$\bar{x}_g = \sqrt[\Sigma f]{\Pi x^f} \qquad (3.14)$$

Логаритмичният вид на непретеглената средна геометрична е

$$log\bar{x}_g = \frac{logx_1 + logx_2 + \ldots + logx_N}{N} = \frac{\sum logx}{N}, \qquad (3.15)$$

а на претеглената

$$log\bar{x}_g = \frac{f_1 logx_1 + f_2 logx_2 + \ldots + f_N logx_N}{f_1 + f_2 + \ldots + f_N} = \frac{\sum(f\ logx)}{\sum f}. \qquad (3.16)$$

Както се вижда, логаритъмът на средната геометрична е средна аритметична от логаритмите на осредняваните величини. Практическото използване на средната геометрична с някои нейни модификации се разглеждат в гл. 10.

3.4.5. Средна квадратична величина

Средната квадратична величина няма самостоятелно приложение като обобщаваща характеристика на разпределенията. Тя се използва като форма за изчисляване на характеристики на вариацията (вж. точка 3.5).

Непретеглената средна квадратична има следната формула:

$$\bar{x}_q = \sqrt{\frac{\sum x^2}{N}}. \qquad (3.17)$$

Претеглената средна квадратична:

$$\bar{x}_q = \sqrt{\frac{\sum x^2 f}{\sum f}}. \qquad (3.18)$$

По подобие на средната квадратична могат да се конструират средни от по-висока степен: средна кубична ($\bar{x}_{cub} = \sqrt[3]{\frac{\sum x^3 f}{\sum f}}$, средна

биквадратична ($\bar{x}_{biq} = \sqrt[4]{\frac{\sum x^4 f}{\sum f}}$) и т.н. Те обаче почти нямат практическо приложение при статистическия анализ.

Ако алгебричните средни се разглеждат формално като различни математически изрази и се изчисляват от едни и същи осреднявани величини, те ще се различават по числова стойност, но винаги ще се подреждат в следния ред:

$$\bar{x}_h < \bar{x}_g < \bar{x}_a < \bar{x}_q < \bar{x}_{cub} \text{ и т.н.}$$

Това подреждане е известно като ***мажорантност на алгебричните средни величини***.

3.4.6. Медиана

Медианата е позиционна средна. Тя е онова значение на признака, което има единицата, заемаща централно положение в статистическия ред, ранжиран възходящо или низходящо по значенията на признака. Броят на единиците (случаите), които имат значения на признака, по-малки от медианата, е равен на броя на единиците (случаите), които имат значения на признака, по-големи от медианата. Медианата следователно разделя съвкупността от единици на две равни части.

Ако всяка единица е дадена с конкретното си значение на признака, медианата се определя непосредствено, като се установи значението на признака на единицата, стояща в средата на реда. Ако например 7 младежи са подредени във възходящ ред по ръст в сантиметри:163, 165, 166, 168, 171, 176, 178, медианата (*Me*) ще бъде 168 см, тъй като младежът с ръст 168 см. заема централно положение и преди него има толкова младежи с по-нисък ръст, колкото са и младежите с по-висок ръст.

Ако единиците са четно число, в средата се намират две единици, а не една. В такъв случай медианата се получава като полусбор от значенията на признака на двете единици. Ако в примера прибавим още един младеж с ръст 182 см.,

$$Me = \frac{168+171}{2} = 169{,}5 \text{ см.}$$

При интервален вариационен ред намирането на медианата се усложнява, тъй като значението на признака на единицата, намираща се в средата, трябва да се намери чрез интерполиране в рамките на интервала, в който се намира медианата.

Ще поясним изчисляването на медианата при интервален вариационен ред по данните от примера, от който беше изчислена средната аритметична (табл. 3.3.).

Необходимо е преди всичко да се определи поредния номер на единицата (заетото лице), която се намира в средата на реда. Той се намира като към общия брой на всички единици (заети) се прибави единица и сумата се раздели на 2.

В примера: $\frac{\sum f+1}{2} = \frac{211+1}{2} = 106$.

За да се установи в коя група се намира 106-ият от всички заети, се проследяват кумулативните честоти.

Таблица 3.3

Разпределение на заетите лица във фирма "Н" по размер на месечната работна заплата през м. октомври 2008 г.

Групови интервали – лв.	Брой на заетите (честоти)	Кумулативни честоти
	f	C
500 - 540	10	10
540 - 580	19	29
580 - 620	27	56
620 - 660	38	94
660 - 700	46	140
700 - 740	35	175
740 - 780	24	199
780 - 820	12	211
	211	x

Очевидно е, че той се намира в групата заети със заплати в интервала 660-700 лв. Тази група, в която се намира медианата, се нарича ***медианна група***, а груповият интервал - ***медианен интервал***. В този интервал са попаднали 46 души. Трябва да се установи кой по ред от тях е търсеният 106-ти. Щом до медианната група има общо 94 души, в медианната група търсеният е 12-ият (106 - 94 = 12). Трябва да се установи каква е неговата заплата. Приема се, че намиращите се в тази медианна група се разпределят в интервала 660-700 лв така, че заплата на всеки в групата е по-голяма спрямо предходния с едно постоянно число, представляващо част от ширината на интервала, падаща се средно на един зает в медианната група. Затова е необходимо да се раздели ширината на интервала (40 лв) на броя на лицата в медианната група: 40 / 46 = 0,87 лв. Следва, че 12-ият поред в медианата група, т.е. 106-ият в целия ред, има с 10,44 лв (12 **.** 0,87 = 10,44 лв) по-голяма заплата, отколкото последният в предмедианната група. Предполага се, че последният преди групата на медианата има заплата, съвпадаща с долната граница на медианния интервал, т.е. 660 лв. И щом търсеният 106-ти има с 10,44 лв по-голяма заплата, значи че неговата заплата е 660 + 10,44 = 670,44 лв. Следователно медианата е Me = 670,44 лв..

Изложеният път за намиране на медианата може да се синтезира в следната формула:

$$Me = L_{Me} + \left(\frac{\sum f + 1}{2} - C_{Me-1} \right) \cdot \frac{h}{f_{Me}} \quad , \tag{3.19}$$

където:

L_{Me} е долната граница на медианния интервал;

C_{Me-1} - кумулативната честота на предмедианния интервал (общ брой на единиците до групата на медианата);

h - ширината на медианния интервал;

f_{Me} - броят на единиците в медианния интервал.

Ако се заместят съответните символи във формулата с конкретните числа от примера, ще се получи

$$Me = 660 + \left(\frac{211 + 1}{2} - 94 \right) \cdot \frac{40}{46} = 670{,}44 \text{ лв.}$$

Медианата не зависи от конкретните значения на признака на всички единици. Тя следователно не се влияе от екстремалните (много големи или много малки) значения на признака, докато средната аритметична зависи от такива екстремални значения.

Медианата има важно ***свойство***: сумата от абсолютните стойности на разликите (отклоненията) между значенията на признака и медианата е минимум, т.е. винаги по-малка от сумата на абсолютните стойности на разликите между значенията на признака и всяко друго число (A), различно от медианата:

$$\sum |x - Me|\, f < \sum |x - A|\, f \,. \qquad (3.20)$$

Това свойство придава на медианата функция на център на разпределението и обуславя нейното използуване в редица случаи при статистическия анализ.

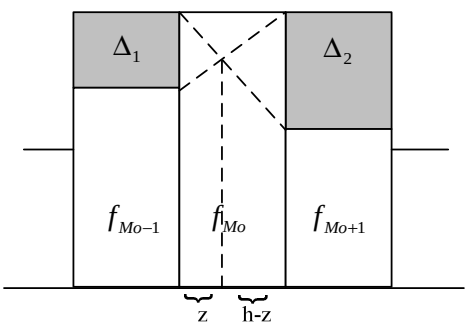
Медианата, както беше посочено, разделя единиците на съвкупността на две равни части. Ако всяка от тези части се третира отделно, за тях също могат да се изчислят медиани. По такъв начин броят на единиците в цялата съвкупност ще се раздели на 4 равни части. Затова тези "медиани" се наричат ***квартили***. По аналогичен начин могат да се намерят ***квинтили,*** които разделят съвкупността на 5 равни части, ***децили*** - на десет равни части, ***центили (перцентили)*** - на 100 равни части и др. Всички те се обединяват под общото наименование ***квантили*** или ***градиенти***. Те се определят принципно по същия начин, както и медианата. При определяне примерно на квартилите трябва да се намерят значенията на признака на единиците, които се намират на 1/4, респ. на 3/4, от реда на всички единици, ранжирани възходящо или низходящо.

3.4.7. Мода

Мода се нарича значението на признака с най-голяма честота, т.е. което се среща най-често в дадена съвкупност. Мода са например: размерът на работната заплата, който се получава от най-голям брой работници в дадена фирма; цената, по която се продава най-голяма част от дадена стока; номерът на мъжките обувки, който се търси най-много от купувачите и др.

При дискретен ред, т.е. когато са дадени конкретните значения на признака на всяка единица, а не в интервали, модата се определя непосредствено без изчисления, тъй като непосредствено се вижда кое значение на признак има най-голяма честота. При интервалните редове тя се намира чрез изчисляване, при което се търси конкретно значение на признака в съответния интервал, имащо най-голяма честота. По броя на случаите в групите може непосредствено да се установи само коя е ***модалната група***, т.е. групата с най-голям брой случаи. Точното значение на признака, което се среща най-често в модалната група, зависи от разпределението на единиците в интервала, а то зависи от общото разпределение в цялата съвкупност и в частност от разликата между броя на случаите в предмодалната и в следмодалната група.

На фиг.3.11 са дадени три стълба от хистограма на разпределението - стълбът на модалната, предмодалната и следмодалната група. Броят на случаите в модалната група се означава с f_{Mo}, в предмодалната - с f_{Mo-1}, а в следмодалната - с f_{Mo+1}. Разликите между броя на единиците в модалната група и предмодалната е $\Delta_1 = f_{Mo} - f_{Mo-1}$, а между модалната и следмодалната е $\Delta_2 = f_{Mo} - f_{Mo+1}$. Щрихованата площ в чертежа представя нагледно тези разлики.



Фиг. 3.11

Очевидно е, че ако разпределението е симетрично, двете разлики биха били еднакви и център на най-голямо натрупване би била средата на интервала на модалната група. Ако предмодалната група има повече случаи от следмодалната, т.е. ако $\Delta_1 < \Delta_2$, натрупването ще бъде най-

голямо при значение на признака, по-малко от средата на итервала. Обратно, ако следмодалната група има повече единици от предмодалната, т.е. ако $\Delta_1 > \Delta_2$, натрупването ще бъде най-голямо при значение на признака, по-голямо от средата на интервала.

Общата ширина на интервала на модалната група (h) се дели от модата на две части - от долната граница на интервала до модата (z) и от модата до горната граница на интервала (h-z). Ако се установи съотношение между тези две части, може да се намери разликата между долната граница на модалната група (L_{Mo}) и модата. Ако тази разлика се прибави към долната граница, ще се намери модата. Това съотношение зависи от разликите между броя на единиците в модалната и в двете съседни групи (Δ_1 и Δ_2). Затова може да се състави следната пропорция:

$$\frac{z}{h-z} = \frac{\Delta_1}{\Delta_2}$$

Неизвестна и търсена величина е z. Тя е

$$\Delta_2 z = \Delta_1 h - \Delta_1 z;$$
$$\Delta_1 z + \Delta_2 z = \Delta_1 h;$$
$$z(\Delta_1 + \Delta_2) = \Delta_1 h;$$
$$z = \frac{\Delta_1 h}{\Delta_1 + \Delta_2}.$$

Щом е намерено z, модата ще се намери, като се прибави то към долната граница на модалната група:

(3.21) $$Mo = L_{Mo} + z; \qquad Mo = L_{Mo} + \frac{\Delta_1 h}{\Delta_1 + \Delta_2}.$$

Тъй като $\Delta_1 = f_{Mo} - f_{Mo-1}$ и $\Delta_2 = f_{Mo} - f_{Mo+1}$ формулата на модата може да се запише още по следния начин:

(3.22) $$Mo = L_{Mo} + \frac{(f_{Mo} - f_{Mo-1})h}{(f_{Mo} - f_{Mo-1}) + (f_{Mo} - f_{Mo+1})}.$$

Ще илюстрираме изчисляването на модата по тази формула, като използуваме ***примера***, от който бяха изчислени претеглената средна аритметична и медианата (вж. табл.3.3)

Най-много заети има с размер на заплатата между 660 и 700 лв. Това е модалната група. Долната граница на интервала е $L_{Mo} = 660$, $f_{Mo} = 46$, $f_{Mo-1} = 38$, $f_{Mo+1} = 35$, $h = 40$ лв.

$$Mo = 660 + \frac{(46-38).40}{(46-38)+(46-35)} = 676{,}84 \text{ лв.}$$

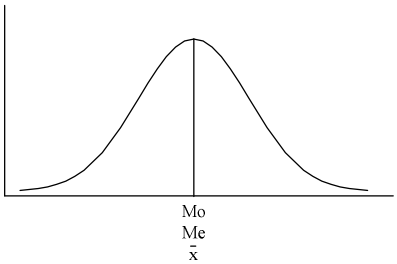
Изложеният начин за намиране на модата е напълно издържан при интервален ред с еднакви ширини на интервалите. Ако интервалите са с различни ширини, може да се допусне грешки при определяне на модалния интервал въз основа на броя на случаите (честотите). При такива положения модалният интервал трябва да се определи по ***плътността на разпределението***, т.е. модата е в интервала с най-голяма плътност (най-голяма честота на единица интервал).

Модата, както средната аритметична и медианата, е характеристика на центъра на разпределението.

3.4.8. Съотношение между средната аритметична, медианата и модата

Между средната аритметична, медианата и модата, като центрове на разпределението, съществува определено съотношение, което зависи от формата на разпределението.

При напълно симетрично разпределение средната аритметична, медианата и модата съвпадат: $\bar{x} = Mo = Me$, т.е. има един център на разпределението (фиг. 3.12).

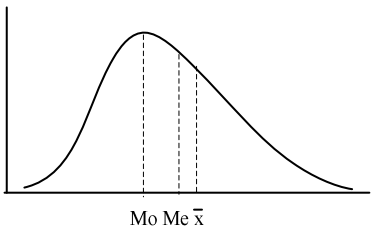


Фиг. 3.12

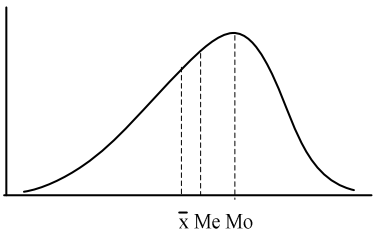
При несиметрично разпределение трите средни се различават. Колкото по-голяма е асиметрията, толкова по-голямо е и различието между тях.

Когато разпределението е несиметрично с дясно изтеглен полигон (фиг. 3.13.), модата е по-малка от медианата, а медианата от своя страна е по-малка от средната аритметична: $Mo < Me < \bar{x}$.

Когато разпределението е несиметрично, с ляво изтеглен полигон (фиг. 3.14.), подреждането е обратно: $\bar{x} < Me < Mo$.



Фиг. 3.13



Фиг. 3.14

Установено е емпирично, че при умерено асиметрично разпределение средната аритметична, медианата и модата се намират в определено ***приблизително съотношение***: разликата между средната аритметична и модата е приблизително равна на три пъти разликата между средната аритметична и медианата:

$$\bar{x} - Mo \approx 3(\bar{x} - Me). \qquad (3.23)$$

От това съотношение се извежда следната ***емпирична формула на модата***:

(3.24) $$Mo \approx \bar{x} - 3(\bar{x} - Me).$$

От тази формула може да се изведат емпирични формули за намиране на медианата и средната аритметична, ако са известни съответно другите две средни:

(3.25) $$Me \approx \frac{2\bar{x} + Mo}{3}.$$

(3.26) $$\bar{x} \approx \frac{3Me - Mo}{2}.$$

Посоченото съотношение е приблизително и важи при умерено асиметрично разпределение.

От решения пример се вижда, че средната аритметична е най-малка ($\bar{x} = 666{,}73$ лв), следвана от медианата ($Me = 670{,}44$ лв) и модата ($Mo = 676{,}84$ лв). Следователно разпределението е несиметрично с изтеглено ляво рамо на кривата.

3.5. Статистическа вариация (разсейване)

3.5.1. Обща характеристика на статистическата вариация

Беше установено, че при изследване на емпиричните разпределения, чрез съответни средни величини се измерва центърът на разпределенията. Но при статистическия анализ често е необходимо да се измери степента на различията между статистическите единици по значенията на даден признак. Възможно е две съвкупности да имат еднакви средни величини (еднакви центрове на разпределението), но различията между значенията на признаците на единиците на съвкупностите да са твърде големи. Ако в два отрасъла на промишлеността средните работни заплати са еднакви, но разликите между индивидуалните заплати в единия са малки, а в другия са много големи, необходимо е да се проучат причините, които пораждат по-голямата диференциация в заплащането във втория отрасъл.

При този и в редица други подобни случаи самата средна добива по-определен смисъл, ако се допълни с характеристика за различията

(вариацията) между осредняваните величини. Нуждата от такава характеристика възниква и при извадковите изследвания, за да се прецени точността на получените резултати, както и да се определи предварително необходимият обем на извадката. Някои други методи на статистическия анализ се основават на изследване на вариацията на значенията на признаците на единиците на съвкупностите.

Тези различия между значенията на признака на отделните единици, както и различията въобще между всякакви величини, които могат да се осреднят, се наричат ***статистическа вариация*** или ***статистическо разсейване.***

За измерване на статистическото вариацията са разработени различни методи.

3.5.2. Размах на вариацията (разсейването)

Най-елементарният измерител е ***размахът на вариацията (d)***. Той е разликата между максималното (x_{max}) и минималното (x_{min}) значение на признака (в литературата се нарича още вариационен размах, размах на колебанията, ранг и др.):

(3.27) $$d = x_{max} - x_{min} .$$

Това е абсолютният размер на размаха на вариацията. За да може той да се сравнява за две и повече съвкупности, трябва да се приведе в относителна форма, като се отнесе към средната аритметична. Това отношение се нарича относителна форма на размаха или ***коефициент на вариацията*** по размаха. То обикновено се умножава по 100, за да се представи в процент:

(3.28) $$V_d = \frac{x_{max} - x_{min}}{\bar{x}} \cdot 100 .$$

Размахът на вариацията е твърде неточна мярка, тъй като зависи само от двете крайни величини. Този недостатък на размаха ограничава неговото приложение само за най-общо ориентиране върху амплитудата на варирането. Само когато разпределението е симетрично или малко се отклонява от симетричното, размахът придобива по-определен смисъл и

се намира в известно приблизително съотношение с останалите измерители.

Специфична област на приложение на размаха е статистическият контрол на процесите, когато е необходимо да се получи бързо информация за вариацията въз основа на малки последователни извадки от наблюдаваните единици (за приложение на статистическите контролни карти).

3.5.3 Средно аритметично отклонение

Сравнително по-точна характеристика на вариацията (разсейването) може да се получи, ако тя се основава върху отклоненията на значенията на признака от тяхната средна аритметична като център на разпределението. Колкото по-малка е вариацията, толкова отделните значения на признака ще бъдат по-близо до средната и обратно, колкото е по-голяма, толкова отделните значения на признака ще са по-отдалечени от средната величина.

Ако се намери средната величина на отклоненията от центъра (в случая средната аритметична величина), тя ще измерва общата вариация. Ако осредняването се извърши аритметично, т.е. да се изчисли средна аритметична от абсолютните стойности на отделните разлики $|x-\bar{x}|$, ще се получи ***средно аритметично (линейно) отклонение*** (δ - делта) . (Съгласно едно свойство на средната аритметична сумата от разликите $(x-\bar{x})$ е равна на 0).

Формулата на ***непретегленото средно аритметично отклонение*** е

$$\delta = \frac{\sum |x-\bar{x}|}{N}, \qquad (3.29)$$

а на претегленото

$$\delta = \frac{\sum |x-\bar{x}|\, f}{\sum f}. \qquad (3.30)$$

Относителната форма на средното аритметично отклонение се получава, като се отнесе абсолютният му размер към средната

аритметична и частното се умножи по 100. Нарича се още ***коефициент на вариацията*** по средното аритметично отклонение (V_δ):

$$V_\delta = \frac{\delta}{\bar{x}} \cdot 100 . \quad (3.31)$$

3.5.4. Средно квадратично (стандартно) отклонение и дисперсия

Най-прецизна и най-често употребявана мярка на вариацията е ***средното квадратично (стандартно) отклонение*** (σ - сигма)**.**

Изходните логически основи на средното квадратично (стандартно) отклонение са същите, както при средното аритметично отклонение - изхожда се от разликите между значенията на признака на отделните единици и тяхната средна аритметична като център на разпределението. Осредняването обаче става като средна квадратична от посочените разлики.

Непретегленото средно квадратично отклонение се изчислява по формулата:

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{N}} . \quad (3.32)$$

Формулата на ***претегленото средно квадратично отклонение*** е:

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2 \, f}{\sum f}} . \quad (3.33)$$

Относителната форма на средното квадратично отклонение се получава като отношение на абсолютния му размер към средната аритметична величина. Представя се обикновено в процент. Нарича се ***коефициент на вариацията*** по средното квадратично отклонение (V_σ).

$$V_\sigma = \frac{\sigma}{\bar{x}} . 100 . \quad (3.34)$$

В редица случаи вместо средното квадратично отклонение се използува неговият квадрат, наречен ***дисперсия*** (σ^2).

Като обобщаваща характеристика (параметър) на емпиричното разпределение средното квадратично отклонение, респективно дисперсията, съответства на аналогичен параметър на теоретичните разпределения (вж. гл. 4) и в частност на нормалното разпределение. Това го прави особено подходящо при решаването на редица въпроси на методологичния апарат на статистическия анализ.[1]

Ще илюстрираме изчисляването на средното аритметично и средното квадратично отклонение по данните от ***примера***, съдържащ се в табл. 3.1. За целта е съставена табл. 3.4.

Таблица 3.4

Изчисляване на средното аритметично отклонение, средното квадратично отклонение и дисперсията на заплатите във фирма "Н" през м. октомври 2008 г.

x	f	$x-\bar{x}$ $(\bar{x}=667)$	$\lvert x-\bar{x}\rvert f$	$(x-\bar{x})^2$	$(x-\bar{x})^2 f$
520	10	-147	1470	21609	216090
560	19	-107	2033	11449	217531
600	27	-67	1809	4489	121203
640	38	-27	1026	729	27702
680	46	13	598	169	7774
720	35	53	1855	2809	98315
760	24	93	2231	8649	207576
800	12	133	1596	17689	212268
	211	x	12619	x	1108459

[1] Относно някои особености при изчисляването на средното квадратично (стандартно) отклонение и на дисперсията въз основа на извадки като оценки на аналогичните параметри на генералните съвкупности вж. гл. 5.

$$\delta = \frac{\sum |x - \bar{x}| f}{\sum f} = \frac{12619}{211} = 59{,}81 \text{ лв};$$

$$V_\delta = \frac{\delta}{\bar{x}} 100 = \frac{59{,}81}{367} 100 = 16{,}3\%.$$

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2 f}{\sum f}} = \sqrt{\frac{1108459}{211}} = \sqrt{5253{,}36} = 72{,}48 \text{ лв};$$

$$V_\sigma = \frac{\sigma}{\bar{x}} 100 = \frac{72{,}48}{667} 100 = 19{,}7\%; \qquad \sigma^2 = 5253{,}36.$$

Дисперсията има някои важни ***свойства***, които могат да се дефинират и като свойства на средното квадратично отклонение.

Първо. Ако към отделните значения на *x* се прибави или от тях се извади постоянно, произволно избрано число *A*, дисперсията (а следователно и средното квадратично отклонение) не се изменя.

Второ. Ако отделните значения на *x* се умножат или се разделят на постоянно, произволно избрано число *A*, дисперсията се увеличава или съответно се намалява A^2 пъти.

Трето. Дисперсията е по-малка от средния квадрат на значенията на *x* с квадрата на тяхната средна аритметична.

$$\sigma^2 = \frac{\sum x^2 f}{\sum f} - \bar{x}^2. \tag{3.35}$$

Четвърто. Дисперсията е по-малка от средния квадрат на разликите между значенията на *x* и някое произволно избрано постоянно число *A* с квадрата на разликата между средната аритметична и това постоянно число.

$$\sigma^2 = \frac{\sum (x - A)^2 f}{\sum f} - (\bar{x} - A)^2. \tag{3.36}$$

Пето. Ако една статистическа съвкупност е разделена на групи (подсъвкупности) общата дисперсия (σ_o^2) е равна на сумата от: 1/ средната аритметична от дисперсиите на отделните групи($\overline{\sigma}_i^2$), наречена вътрешногрупова дисперсия; 2/ дисперсията на средните аритметични на групите (σ_m^2), наречена междугрупова дисперсия.

$$\sigma_o^2 = \overline{\sigma}_i^2 + \sigma_m^2 .$$

Това свойство е известно като правило за събиране и разлагане на дисперсии.

3.5.5. Средно квадратично отклонение при алтернативни категорийни признаци

При статистическия анализ се налага изчисляването на ***средното квадратично отклонение при алтернативни (бинарни, дихотомни) признаци.***

Известно е вече, че за разпределение по такива признаци се прилага дихотомна (бинарна) скала, като двете алтернативни значения на признака се означават с 1 и 0.

Относителният дял на единиците, които имат едното значение, се означава с p, а относителният дял на другото - с q. В такъв случай се получават две числови значения на признака (1 и 0), на които съответстват тегла p и q, при което $p + q = 1$, следователно $p = 1 - q$, а $q = 1 - p$.

Може да се състави формула на средната аритметична, като се има предвид нейната основна формула ($\bar{x} = \frac{\sum xf}{\sum f}$). Чрез заместване на x с числовите значения на алтернативния признак - 0 и 1, а теглата f - с p и q, ще се получи:

$$\frac{1 \cdot p + 0 \cdot q}{p + q} = p . \qquad (3.37)$$

Като се заместят x, f и $\bar{x}$ във формулата на средното квадратично отклонение ($\sigma = \sqrt{\frac{\sum (x - \bar{x})^2 f}{\sum f}}$) с 1 и 0, p и q, ще се получи:

$$\sigma_p = \sqrt{\frac{(1-p)^2 p + (0-p)^2 q}{p+q}} . \qquad (3.38)$$

Тъй като $1 - p = q$ и $p + q = 1$, то

(3.39) $$\sigma_p = \sqrt{q^2 p + p^2 q} = \sqrt{pq(q+p)} = \sqrt{pq}\,.$$

Ако q се замести с $1 - p$,

(3.40) $$\sigma_p = \sqrt{p(1-p)}\,.$$

Във връзка с интерпретирането на средното квадратично (стандартно) отклонение при алтернативни категорийни признаци трябва да се има предвид, че неговата максимална числова стойност може да бъде 0,5, тъй като произведението pq може да има максимална стойност 0,25 (когато двете групи единици, обособени по двете алтернативни значения, имат еднакъв относителен дял, т.е. когато $p = 0{,}5$ и $q = 0{,}5$).

3.5.6. Квартилно отклонение

Известно е, че медианата и квартилите разделят единиците на съвкупността на четири равни части. Когато вариацията е малка, квартилите се разполагат по-близо до медианата, и, обратно, при по-голяма вариация отдалечеността им от медианата е по-голяма. Това означава, че разликите между медианата и квартилите са чувствителни на изменението на вариацията. Това дава основание и тези разлики да се използуват за измерване на вариацията. Тъй като двете разлики са равни само при симетрично разпределение, а при асиметрично разпределение малко или много се различават една от друга, като мярка на вариацията се използва полусборът от двете разлики - разликата между медианата и първия квартил (Q_1) и между третия квартил (Q_3) и медианата. Тази мярка на вариацията се нарича ***квартилно отклонение* (Q),** известно в литературата още като метод на ***Артур Боули (1869-1957)***. Това всъщност е половината от разликата между двата квартила:

(3.41) $$Q = \frac{(Q_3 - M_e) + (M_e - Q_1)}{2} = \frac{Q_3 - Q_1}{2}\,.$$

Разликата между двата квартила се нарича в литературата още ***интерквартилен размах,*** а квартилното отклонение – ***полуинтер-квартилен размах.***

Относителната форма на квартилното отклонение се получава, като се отнесе абсолютният му размер към полусбора от двата квартила и се умножи по 100, за да се представи в процент. Нарича се още ***коефициент на вариацията по квартилното отклонение***. Той може да се запише така:

$$V_Q = \frac{\frac{Q_3 - Q_1}{2}}{\frac{Q_1 + Q_3}{2}} \cdot 100 = \frac{Q_3 - Q_1}{Q_1 + Q_3} \cdot 100 . \qquad (3.42)$$

По аналогия на квартилното отклонение могат да се конструират измерители въз основа на други квантили - децили и центили. Затова измерителите, конструирани на основата на квантили, могат да се нарекат общо ***квантилни отклонения***. От тях обаче най-често се използува квартилното отклонение.

Квартилното отклонение се изчислява лесно и дава обща представа за степента на разсейването (вариацията). Трябва обаче да се има предвид, че то не характеризира разсейването около някакъв център, а се основава върху размаха между квартилите. Тъй като при изчисляването му не се вземат всички значения на признака, възможно е в някои случаи квартилите в два реда да съвпаднат даже и когато вариацията около средната величина е различна.

3.5.7. Средна разлика

Вариацията може да се разглежда не само като степен на отдалеченост на значенията на признака на отделните единици от средната аритметична като център на разпределение, или като диапазон, в който се разполагат единиците, или определени характеристики на разпределението (например квартилите). Тя може да се разглежда и като различие между значенията на признака на всяка единица спрямо всички други единици на съвкупността. Върху такава изходна постановка е конструирана ***средната разлика*** като мярка на вариацията, известна в литературата като средна разлика на ***Корадо Джини (1884-1965)***. Тя представлява отношение на сумата на всички разлики към техния брой.

Без да навлизаме в подробен коментар и в математическо извеждане, ще посочим една удобна формула за изчисляване на средната разлика (*G*):

$$G = \frac{\sum x_i f_i \left(C_i^{(+)} - C_i^{(-)}\right)}{\frac{\sum f_i \left(\sum f_i - 1\right)}{2}}, \qquad (3.43)$$

където:

$C_i^{(+)}$ са прогресивно-кумулативните честоти, образувани чрез последователно сумиране с натрупване на честотите отгоре надолу по групите във вариационния ред;

$C_i^{(-)}$- регресивно-кумулативните честоти, образувани чрез последователно сумиране с натрупване отдолу нагоре по групите във вариационния ред.

Средната разлика се използва сравнително рядко като мярка на статистическото разсейване (вариацията).

3.5.8. Емпирично съотношение между измерителите на вариацията

Според числовата им стойност измерителите на вариацията се подреждат винаги в определен ред. Това се нарича ***мажорантност*** на тези измерители:

$$d > G > \sigma > \delta > Q \qquad (3.44)$$

Доказано е, че когато разпределението е нормално, между тях съществува определено приблизително съотношение. То може да се представи така:

$$d \approx 6\sigma \approx 7{,}5\delta \approx 9Q;\ G \approx 1{,}128\sigma \approx 1{,}414\delta \approx 1{,}647Q;$$

$$\sigma \approx 1{,}253\delta \approx 1{,}486Q;\ \delta \approx 0{,}978\sigma \approx 1{,}185Q;\ Q \approx 0{,}674\sigma \text{ и т.н.}$$

Изразът $d \approx 6\sigma$ е известен като ***правилото на 6-те сигми***. То се използва практически например при статистическия контрол на качеството за намиране на приблизителната стойност на σ чрез размаха на вариацията (*d*).

Посочените съотношения могат да се използват за намирането на приблизителната стойност на който и да е измерител, ако е даден друг. Това е възможно, разбира се, ако има достатъчно основание да се смята, че разпределението е нормално или незначително се отклонява от него.

3.6. Асиметрия и ексцес

3.6.1. Моменти на разпределението

При изследване на статистическите разпределения се използуват такива обобщаващи характеристики, които имат общото наименование ***моменти на разпределението***. Някои от разгледаните средни величини и измерители на вариацията са по форма и по същество моменти на разпределението.

Общо под момент на разпределението се разбира средната аритметична от *k*-тите степени на разликите между значенията на признака (*x*) и някоя постоянна величина. Ако за знак на момент се приеме *M*, а постоянната величина се означи с *A*, общата формула на моментите би имала следния вид:

$$M_k = \frac{\sum (x - A)^k \ f}{\sum f}. \qquad (3.45)$$

Като се придава на *k* значение 1, 2, 3 и т.н., получават се моменти от различен порядък - първи момент, втори момент, трети момент и т.н.

Ако за постоянна величина се приеме средната аритметична, моментите се наричат ***централни*** и се означават с *m*.

$$m_k = \frac{\sum (x - \bar{x})^k \ f}{\sum f}. \qquad (3.46)$$

Очевидно е, че ***първият централен момент*** е равен на 0, тъй като $(x - \bar{x}) = 0$ съгласно едно от свойствата на средната аритметична.

Вторият централен момент е дисперсията:

$$m_2 = \frac{\sum (x - \bar{x})^2 \ f}{\sum f} = \sigma^2. \qquad (3.47)$$

Съществено значение при статистическия анализ имат още ***третият*** и ***четвъртият централни моменти***.

$$(3.48) \qquad m_3 = \frac{\sum (x - \bar{x})^3 f}{\sum f};$$

$$(3.49) \qquad m_4 = \frac{\sum (x - \bar{x})^4 f}{\sum f}.$$

Когато централните моменти от съответен порядък се отнесат към средното квадратично отклонение, повдигнато на съответна степен, се получават ***нормирани централни моменти*** от съответен порядък. Някои от тези нормирани централни моменти се използват като характеристики относно формата на разпределенията.

3.6.2. Измерване на асиметрията

Асиметрията може да се измери по различни методи, при които се получават ненаименовани величини, наречени ***коефициенти на асиметрията*.**

1. Един от възможните коефициенти е ***моментният коефициент*** на асиметрията (γ_1). Той е нормиран трети централен момент. Изчислява се като отношение на третия централен момент към средното квадратично отклонение, повдигнато на трета степен:

$$(3.50) \qquad \gamma_1 = \frac{m_3}{\sigma^3}.$$

При симетрично разпределение $\gamma_1 = 0$, тъй като третият централен момент (m_3) в този случай е равен на нула. Счита се, че когато γ_1 превишава по абсолютна стойност 0,5 , асиметрията е значителна.

Моментният коефициент е добре обоснован измерител на асиметрията.

2. Известно е, че ако разпределението е симетрично, средната аритметична, медианата и модата съвпадат, а при асиметрично разпределение се различават. Ако даденото емпирично разпределение се отклонява малко от симетричното, разликата между посочените средни е малка. Ако асиметрията е по-голяма, по-голяма е и разликата между

средните. Средната аритметична реагира най-чувствително на асиметрията и се отклонява от модата към полегатото рамо повече, отколкото медианата. Това дава основание разликата между средната аритметична и модата да се използува за измерване на асиметрията. Тъй като тази разлика е абсолютна величина, изразена в съответна мярка, в каквато са изразени значенията на *x*, тя се привежда в коефициент (Sk_1), като се раздели на средното квадратично отклонение:

$$Sk_1 = \frac{\bar{x} - M_o}{\sigma}. \qquad (3.51)$$

При симетрично разпределение $\bar{x} = Mo$ и следователно $S_{k_1} = 0$. Когато разпределението е асиметрично с дясно изтеглена крива, асиметрията е положителна, т.е. $S_{k_1} > 0$, тъй като $\bar{x} > Mo$. Обратно, ако разпределението е с ляво изтеглена крива, асиметрията е отрицателна, т.е. $S_{k_1} < 0$, тъй като $\bar{x} < Mo$.

Известно е, че при умерено асиметрично разпределение съществува определено приблизително съотношение между средната аритметична, медианата и модата, съгласно което $(\bar{x} - Mo) \approx 3(\bar{x} - Me)$. В такъв случай може вместо $(\bar{x} - Mo)$ да се използува разликата $(\bar{x} - Me)$, взета три пъти. Тогава коефициентът на асиметрията ще има следната формула:

$$S_{k_2} = \frac{3(\bar{x} - Me)}{\sigma}. \qquad (3.52)$$

Очевидно е, че когато разпределението е симетрично, $S_{k_2} = 0$, тъй като $\bar{x} = Me$. При асиметрично разпределение с дясно изтеглена крива $\bar{x} > Me$ и следователно $S_{k_2} > 0$. При ляво изтеглена крива $\bar{x} < Me$ и затова $S_{k_2} < 0$.

Коефициентите S_{k_1} и S_{k_2} могат да приемат стойност в границите от -3 до +3. Счита се, че когато са различни от 0 и по абсолютна стойност не превишават 1, разпределението може да се приеме за умерено асиметрично.

3. За измерване на асиметрията може да се използват квантилите, въз основа на които се изчисляват ***квантилни коефициенти***.

Известно е например, че при симетрично разпределение двата квартила (Q_1 и Q_3) се намират на еднакво разстояние от медианата, т.е. (Q_3-Me) = (Me-Q_1). При асиметрично разпределение двете разлики не са еднакви. Колкото асиметрията е по-голяма, толкова по-голяма е и разликата между двете разлики. Това дава основание разликата (Q_3-Me) - (Me-Q_1) да се приеме като абсолютна мярка на асиметрията. За да се изрази в коефициент, тя се разделя на разликата между двата квартила, т.е. на интерквартилния размах. Така се получава ***квартилният коефициент*** на асиметрията (K_Q):

$$K_Q = \frac{(Q_3 - Me) - (Me - Q_1)}{Q_3 - Q_1}. \qquad (3.53)$$

Ако се разкрият скобите и се направят възможните съкращения, формулата ще приеме вида:

$$K_Q = \frac{Q_1 + Q_3 - 2Me}{Q_3 - Q_1}. \qquad (3.54)$$

При симетрично разпределение $K_Q = 0$. При асиметрично разпределение с дясно изтеглена крива $K_Q > 0$, тъй като $(Q_3 - Me) > (Me - Q_1)$. При асиметрично разпределение с ляво изтеглена крива $K_Q < 0$, тъй като $(Q_3 - Me) < (Me - Q_1)$. Квартилният коефициент може да приема стойности от -1 до +1. Доказано е, че когато K_Q е до ± 0,1, асиметрията е умерена.

По аналогия на квартилния коефициент могат да се конструират ***децилен*** и ***центилен*** коефициент на асиметрията.

3.6.3. Измерване на ексцеса

Ексцесът също може да се измери чрез коефициенти, изчислени по различни методи.

1. Най-точният измерител е ***моментният коефициент*** на ексцеса (γ_2), който се получава като частно от четвъртия централен момент и средното квадратично отклонение на четвъртата степен, т.е. това е нормираният четвърти централен момент на разпределението:

(3.55) $$\gamma_2 = \frac{m_4}{\sigma^4}.$$

При нормален ексцес $\gamma_2 = 3$, при положителен (наднормален) - $\gamma_2 > 3$, а при отрицателен (поднормален) - $\gamma_2 < 3$. Тези числови стойности на γ_2 се получават поради това, че при нормално разпределение съществува следното съотношение между четвъртия централен момент и втория централен момент (дисперсията):

$$m_4 = 3m_2^2, \text{ т.е. } m_4 = 3\sigma^4.$$

Ако от коефициента γ_2 се извади 3, ще се получи:

(3.56) $$\gamma_3 = \frac{m_4}{\sigma^4} - 3.$$

Изчисленият по тази формула коефициент ще бъде 0 при нормален ексцес. При наднормален ексцес $\gamma_3 > 0$, а при поднормален - $\gamma_3 < 0$.

2. Установено е, че съществува определена връзка между ексцеса и квантилите, въз основа на която могат да се изведат ***квантилни коефициенти*** на ексцеса. Те обаче намират ограничено приложение, тъй като моментният коефициент е по-прецизен.

Квартилно-децилният коефициент на ексцеса ($E_{Q/D}$) е:

(3.57) $$E_{Q/D} = \frac{\frac{1}{2}(Q_3 - Q_1)}{D_9 - D_1} = \frac{Q_3 - Q_1}{2(D_9 - D_1)}.$$

Тъй като $D_1 = C_{10}$, а $D_9 = C_{90}$ ***квартилно-центилният*** коефициент ($E_{Q/C}$) е:

(3.58) $$E_{Q/C} = \frac{Q_3 - Q_1}{2(C_{90} - C_{10})}.$$

И в двата варианта на формулата коефициентът на ексцеса има при наднормален ексцес максимална стойност 0,5. При поднормален (отрицателен) ексцес пределната му стойност е 0. За нормален се счита ексцесът, когато коефициентът е 0,263.

3.7. Практикум

3.7.1. Въпроси за самопроверка

1. Кои разпределения се наричат честотни?
2. Как се получава абсолютната и относителната плътност на разпределението?
3. Кои са основните обобщаващи характеристики на едномерното емпирично разпределение?
4. Какво е полигон на разпределението и хистограма?
5. Как се дефинират средните величини в статистиката?
6. Какво е общото правило за определяне на адекватните тегла при изчисляването на претеглената средна аритметична?
7. В какви случаи се прилага средната хармонична?
8. Каква средна величина е медианата?
9. Каква средна величина е модата?
10. Какво е съотношението между средната аритметична, медианата и модата и каква е връзката му с формата на разпределението?
11. Какво се разбира под вариация или статистическо разсейване?
12. Какви свойства притежава дисперсията?
13. Какво е основанието да се използва интерквартилният размах за измерване на вариацията?
14. Какво означава правилото на 6-те сигми?
15. Какво представляват централните моменти на разпределението?
16. Какво се разбира под асиметрия и ексцес?
17. Как се изчисляват моментните коефициенти на асиметрията и ексцеса?
18. Какво е логическото основание да се използва разликата между средната аритметична и модата за измерване на асиметрията?

3.7.2 Задачи за упражнение

Задача 1. Дадени са данни за 95 фирми в следващата таблица.

Таблица 3.5
Разпределение на група фирми по размер на реализираната печалба през 2007 г.

Групови интервали (хил. лв.)	Брой на фирмите
200-400	2
400-600	5
600-800	10
800-1000	18
1000-1200	30
1200-1400	17
1400-1600	8
1600-1800	4
1800-2000	1

Необходимо е:

1. Да се изчислят относителните и кумулативните честоти.
2. Да се изчисли абсолютната и относителната плътност на разпределението.
3. Да се начертаят полигонът на разпределението и хистограмата.
4. Да се изчисли средната аритметична (печалбата средно на една фирма).
5. Да се изчисли медианата.
6. Да се изчисли модата.
7. Да се коментира съотношението между средната аритметична, медианата и модата във връзка с формата на разпределението.
8. Да се изчислят средната аритметична, медианата и модата по емпиричните формули и да се сравнят с изчислените по основните формули.

Отговори:

а) По основните формули (в хил. лв.):

$\bar{x} = 1074{,}7;\ \ Me = 1086{,}7;\ \ Mo = 1096{,}0$

б) По емпиричните формули (в хил. лв.):

$\bar{x} = 1082{,}0\ \ Me = 1081{,}9;\ \ Mo = 1100{,}5$

Задача 2. Дадени са данни в следващата таблица.

Таблица 3.6
Продадени количества от стока "А" и продажни цени на три пазара през април 2008 г.

Пазари	Цена за 1 кг (лв)	Стойност на продадените количества (лв)
А	2,0	2400
Б	2,5	2500
В	1,8	3240
		8140

Необходимо е да се изчисли средната цена за цялото количество стоки, продадени на трите пазара.

Отговор:

$\bar{y} = 2{,}04$ лв.

Задача 3. По данните от задача 1 да се изчислят:

1. Дисперсията и средното квадратично отклонение;
2. Коефициентът на вариацията (по средното квадратично отклонение);
3. Квартилното отклонение.

Отговори: $\sigma^2 = 98730{,}26;\ \sigma = 314{,}21;\ V_\sigma = 29{,}23\%;\ Q = 19\%$.

Задача 4. При контрол на качеството на произведени плодови консерви в едно предприятие е установено чрез извадка, че 5 % от проверените кутии не отговарят на съществуващия стандарт (95 % са стандартни). Необходимо е по тези данни да се изчисли средното квадратично отклонение, като по дихотомната скала за стандартните консерви се записва 1, а за нестандартните - 0.

Отговор: $\sigma = 0{,}2179$

Задача 5. По данните от задача 1 да се изчислят:

1. Моментният коефициент на асиметрията (γ_1);
2. Квартилният коефициент на асиметрията (K_Q);
3. Коефициента на асиметрията чрез средната аритметична, модата и стандартното отклонение.

Отговори: $\gamma_1 = -0{,}005;\ K_Q = -0{,}03;\ S_{k_1} = -0{,}068$.

3.7.3. Абсурдните средни величини

Средните величини са обобщаващи статистически характеристики с много голямо познавателно значение. Но не винаги. Има случаи, при които те просто са абсурдни. Това може да се илюстрира със следните забавни примери.

1. ***Дж. Глас и Дж. Стенли***[1] разказват, че на пейка в парка случайно се събрали 5 мъже. Двамата били безработни бездомници с по 25 цента в джобовете. Третият бил работник със спестени 2000 долара, четвъртият - чиновник с 15000 долара по банкова сметка и петият - милионер, притежател на 5000000 долара. От тези данни могат да се изчислят средни величини: Модата е $Mo = 25$ цента, медианата - $Me = 2000$ долара и средната аритметична - $\bar{x} = 1003400{,}10$ долара.

Очевидно е, че модата в случая характеризира състоянието на безработните, медината - само на работника. Средната аритметична също няма смисъл. В случая няма център на разпределението като обобщаваща характеристика на съвкупността.

2. Известният руски писател ***Глеб Успенски (1843 – 1902)*** пише в разказа си ***"Четвърт кон"***[2]:

"В село Присухино училището има 30 ученика, в село Засухино - 20, а в село Оплеухино - само 2 ученика. Следва, че средният брой ученици на едно училище е 17,333. Това е все едно, ако взема милионера Колотушкин, в джоба на който има милион, и към него прибавя Кукушкин, който има само грош, и тогава всеки от тях ще има средно по половин милион.

3. В книгата си ***"Животът по Мисисипи" Марк Твен (1835 – 1910)*** прави следните изчисления и заключения.

"За 176 години Мисисипи е скъсила с 242 мили долното си течение. Това прави средно малко повече от миля и една трета на година. Следователно, всеки уравновесен, нормален човек, който не е нито сляп, нито идиот, може да се убеди, че през древния силурски период, от който

[1] **Гласс, Дж. И Дж. Стенли.** Статистические методы в педагогике и психологии. М., 1976, с. 70.
[2] **Успенский, Г. И.** Избранные сочинения. Москва-Ленинград, 1949, с. 428.

идният ноември се навършват точно един милион години, долното течение на Мисисипи е било над 1300000 мили. По същия начин всеки може да се убеди, че след 742 години тя ще бъде само 1,75 мили... Науката е същинска магия. И най-незначителният факт може да наплоди какви ли не догадки.”[1]

[1] **Твен, М.** Животът по Мисисипи. С., 1985, с. 139.

4. ТЕОРЕТИЧНИ РАЗПРЕДЕЛЕНИЯ

> "Статистиката е съвкупност от методи, които дават възможност за вземане на оптимални решения в условията на неопределеност."
>
> А. Уолд

Тази глава предлага кратък преглед на основни положения от теорията на вероятностите и характеристика на най-важните за статистиката теоретични (вероятностни) разпределения. Тук се обръща внимание не толкова на математическия апарат (формулите), колкото на познавателната същност на понятията и съответните параметри. Изяснява се практическият смисъл на закона за големите числа. Насочва се вниманието към специфичните особености на отделните теоретични разпределения и в частност на извадковите разпределения, имащи пряка връзка със статистическите изследвания. Усвояването на съдържанието на тази глава е абсолютно необходимо условие за разбиране същността и прилагането на статистическите методи, разглеждани в следващите глави.

4.1. Същност на теоретичните разпределения и основни понятия

Теоретичните разпределения са предмет на изследване в теорията на вероятностите, но имат особено голямо значение за теорията и практиката на статистиката. Във връзка с разглеждането им по-нататък е необходимо да се дефинират и изяснят някои основни понятия на теорията на вероятностите: опит (изпитване, наблюдение), събитие, случайна величина, вероятност и др.

Под ***опит*** (изпитване, наблюдение) се разбира най-общо действие, осигуряващо условията за появата на определено събитие. Например подхвърляне на монета или зар, избиране случайно на произведени изделия, за да се провери тяхното качество и др.

Събитието е резултат от опита. При подхвърлянето на монета събитието е получаването на ези или тура, при случайно избраните

изделия събитието е отделното изделие да се окаже стандартно или нестандартно. Събитията могат да бъдат сигурни, невъзможни и случайни (вероятни).

Сигурно събитие е това, което безусловно настъпва. Ако приемем, че всички произведени изделия са стандартни, то провереното едно изделия да е стандартно е сигурно събитие.

Невъзможно събитие е това, което изобщо не може да се случи.

Случайно (вероятно) събитие е това, което при опита може да се случи (да се сбъдне), но може и да не се случи. Ако в партидата произведени изделия има определен брой стандартни и нестандартни, при случайно проверено изделие то може да се окаже или стандартно, или нестандартно.

Следователно, докато при сигурните и невъзможните събития има един възможен изход от опита (наблюдението), при случайните събития възможностите са повече.

Случайната величина е такава величина, която в резултат на различни случайни обстоятелства може да приема различни стойности, които като правило предварително не са известни.

Случайни величини са например: резултатът, който се получава при тираж на спортния тотализатор или на държавната лотария; броят на рекламациите, които една фирма получава за определено време; броят на дефектните изделия при направена извадка за проверка на качеството на произведена продукция; броят на поръчките за доставка на продукти в къщи, постъпили в един магазин за определено време и т.н.

Случайната величина е ***прекъсната (дискретна)***, ако може да приема само отделни, изолирани една от друга стойности. Тя е ***непрекъсната (индискретна)***, когато може да приема произволни стойности в определен интервал. Това деление на случайните величини е напълно аналогично на разгледаните вече прекъснати и непрекъснати вариационни статистически признаци.

Вероятността може да се дефинира най-общо като мярка на обективната възможност да настъпи дадено събитие. Има и други дефиниции.[1]

Всяка възможна стойност на случайната величина е свързана с определена ***вероятност*** (p_i). Вероятността на сигурно събитие е 1, на невъзможно събитие - 0, а на случайно събитие - между 0 и 1. Отделните възможни стойности на случайната величина са несъвместими и сумата на техните вероятности е $\sum_{i=1}^{n} p_i = 1$.

Да си представим, че от една партида произведени изделия се прави случайна извадка от n изделия и се установява, че от тях m изделия са дефектни. Отношението $\frac{m}{n}$ е относителна честота на дефектните. Нейната стойност предварително (преди опита) не е известна. Това е случайна величина, която при отделните извадки като правило ще има различни стойности. Ако обаче се правят много на брой извадки (много опити), относителните честоти $\frac{m}{n}$ ще се групират около някаква стойност p, която е ***статистическата вероятност***, наричана още ***апостериорна вероятност*** (от лат. *a posteriori* - след опита).

Ако се подхвърля зар, предварително са известни възможните резултати (изходи) - 1 точка, 2 точки и т.н. до 6 точки. Вероятността за всеки отделен възможен резултат е $\frac{1}{6}$. И тъй като тази вероятност е известна преди опита (преди подхвърляне на зара), тя се нарича ***априорна вероятност*** (от лат. *a priori* - до опита, преди опита).

От много опити, като посочения за проверка на качеството на произведени изделия чрез извадки, се установява, че между вероятностите и относителните (релативните) честоти, които се получават по опитен път, има връзка. Установява се по-конкретно, че при достатъчно голям брой опити (голям брой наблюдавани случаи), по силата на закона за големите числа, относителните честоти, с вероятност , клоняща към единица, се доближават до съответните единични вероятности.

[1] Вж. **Сугарев, З**. и **С. Каменаров**, Теория на вероятностите, С., 1974.

Както честотите при емпиричните изследвания се разпределят по съответните значения на вариационните признаци (по интервална скала), така и вероятностите се разпределят по възможните стойности на случайните величини. Тези разпределения се наричат ***вероятностни*** или ***теоретични разпределения*** (за разлика от емпиричните честотни разпределения).

Такова разпределение при прекъсната (дискретна) случайната величина може да се представи най-общо в табл. 4.1.

Таблица 4.1

x_1	x_2	x_3	$\cdots$	$\cdots$	x_{n-1}	x_n
p_1	p_2	p_3	$\cdots$	$\cdots$	p_{n-1}	p_n

Когато разпределенията се отнасят за прекъснати (дискретни) случайни величини, те се наричат също ***прекъснати (дискретни) теоретични разпределения***, а когато се отнасят за непрекъснати случайни величини - ***непрекъснати (индискретни) теоретични разпределения***.

От посочената връзка между вероятностите и относителните (релативните) честоти произтича връзката между теоретичните (вероятностните) разпределения и емпиричните (честотните) разпределения. При решаването на редица методологични въпроси на теорията на статистиката, теоретичните разпределения се използуват като модели за изследване на емпиричните разпределения.

Както емпиричните така и теоретичните разпределения биват ***едномерни, двумерни и многомерни*** и имат различна форма.

4.2. Закон на разпределението

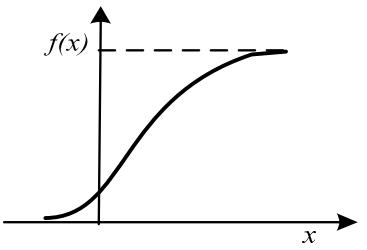
Връзката между възможните стойности на случайната величина и техните вероятности се нарича ***закон на разпределението***. Той може да се дефинира чрез две основни функции.

1. ***Функция на разпределението*** на вероятностите, наричана още интегрален закон на разпределението или интегрална функция на

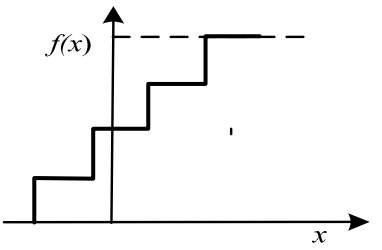
разпределението. Тя се отнася както за непрекъснати, така и за прекъснати случайни величини.

Тази функция $F(x)$ задава вероятността случайната величина да приема стойност до *x*. Тя е неотрицателна и ненамаляваща. За непрекъснати случайни величини $F(x)$ е представена графично на фиг. 4.1., а за прекъснати - на фиг. 4.2. В математическа форма функцията е:

(4.1) $$F(x) = P(X) \leq x .$$



Фиг. 4.1



Фиг. 4.2

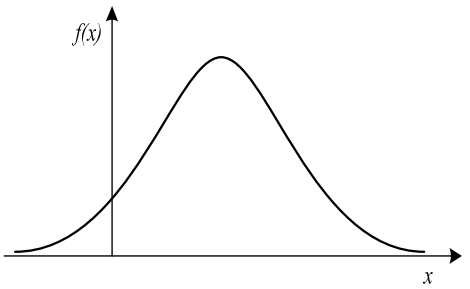
2. ***Функция на плътността*** на вероятностите, наричана още диференциален закон на разпределението или диференциална функция на разпределението. Тя задава вероятността случайната величина да приема стойност, намираща се в даден интервал.

Ако се фиксира какъвто и да е интервал от *a* до *b*, вероятността случайната величина да има стойност, попадаща в този интервал, е:

(4.2) $$P(a < X < b) = \int_a^b f(x)dx .$$

Графично функцията на плътността (диференциалният закон) на разпределението е показан на фиг. 4.3.

Функцията на плътността на вероятностите е неотрицателна ($f(x) \geq 0$) и интегралът в границите от - ∞ до +∞ е равен на единица - $\int_{-\infty}^{\infty} f(x)dx = 1$, аналогично на $\sum_{i=1}^{n} P_i = 1$.



Фиг. 4.3

4.3. Математическо очакване и дисперсия на случайна величина

Графичното или чрез математическа функция представяне на закона на разпределението е полезно, но недостатъчно. Необходими са определени параметри на теоретичните разпределения, аналогично на емпиричните разпределения.

Основните параметри на теоретичните разпределения са ***математическото очакване*** и ***дисперсията***, респ. ***средното квадратично отклонение***.

Математическото очакване *E(X)* е средната стойност на случайната величина, изчислена от всички нейни възможни стойности (x_i), претеглени с техните вероятности (p_i). То характеризира центъра на теоретичното разпределение, аналогично на средната аритметична като център на емпиричното разпределение.

При прекъсната (дискретна) случайна величина се изразява с формулата:

$$E(X) = \bar{x} = \sum_{i=1}^{n} x_i \; p_i \,. \tag{4.3}$$

Математическото очакване на непрекъсната (индискретна) случайна величина е

$$E(X) = \bar{x} = \int_{-\infty}^{\infty} x f(x) dx \,. \tag{4.4}$$

Математическото очакване притежава характерни свойства:

1. Математическото очакване на сума от случайни величини е равно на сумата от техните математически очаквания:

(4.5) $E(X_1 + X_2 + ... + X_n) = E(X_1) + E(X_2) + ... + E(X_n).$

2. Математическото очакване на произведение от независими в съвкупност случайни величини е равно на произведението от техните математически очаквания:

(4.6) $E(X_1, X_2, ..., X_n) = E(X_1).E(X_2) ... E(X_n).$

Дисперсията, респ. стандартното отклонение, на случайна величина измерва вариацията на възможните ѝ стойности около математическото очакване.

Дисперсията на прекъсната (дискретна) случайна величина е

(4.7) $$\sigma^2 = \sum_{i=1}^{n} \left(x_i - E(X)\right)^2 p_i \ ,$$

а средното квадратично отклонение е положителен квадратен корен на дисперсията:

(4.8) $$\sigma = \sqrt{\sum_{i=1}^{n} \left(x_i - E(X)\right)^2 p_i} \ .$$

Дисперсията на непрекъсната (индискретна) случайна величина е

(4.9) $$\sigma^2 = \int_{-\infty}^{\infty} \left(x_i - E(X)\right)^2 f(x)dx \ ,$$

а средното квадратично (стандартно) отклонение -

(4.10) $$\sigma = \sqrt{\int_{-\infty}^{\infty} \left(x_i - E(X)\right)^2 f(x)dx} \ .$$

Дисперсията има някои важни свойства, които имат значение за статистическия анализ.

1. Дисперсията на сума от независими случайни величини е равна на сумата от техните дисперсии:

(4.11) $$\sigma^2_{(X_1+X_2+...+X_N)} = \sigma^2_{X_1} + \sigma^2_{X_2} + ... + \sigma^2_{X_n}.$$

2. Дисперсията на разлика между две независими случайни величини е равна на сумата от техните дисперсии:

(4.12) $$\sigma^2_{(X_1-X_2)} = \sigma^2_{X_1} + \sigma^2_{X_2}.$$

3. Ако X_1, X_2,..., X_n са еднакво разпределени независими случайни величини, дисперсията на всяка от които е σ^2, тогава:

а) дисперсията на тяхната сума е

(4.13) $$n\sigma^2;$$

б) дисперсията на техните средни аритметични е

(4.14) $$\frac{\sigma^2}{n}.$$

Ако от стойностите (x_i) на една случайна величина (X) се извади математическото очакване (средната величина на тези стойности) и разликите се разделят на стандартното отклонение, ще се получи друга случайна величина (Z) със стойности z_i. Тя се нарича ***стандартизирана случайна величина***:

(4.15) $$z_i = \frac{x_i - \bar{x}}{\sigma}.$$

Характерно и особено важно за такава случайна величина е това, че тя има математическо очакване 0 и дисперсия 1. Както ще стане ясно по-нататък, това има съществено значение при анализа на разпределенията.

4.4. Закон за големите числа

В гл. 1 беше дадена най-обща характеристика на логическата същност на закона за големите числа. В теорията на вероятностите той се дефинира по-строго.

В най-обща формулировка законът за големите числа гласи, че съвкупното действие на голям брой случайни фактори води при определени условия до резултат, почти независещ от отделния случай. В

съгласие с тази обща формулировка чрез теоретичен анализ и многократни експериментални проверки са изведени и доказани съответни теореми, които са опора при статистическото изучаване на явленията, имащи вероятностен характер.

Първият значителен принос в тази област е направен от знаменития швейцарски математик ***Якоб Бернули (1654-1705)***. Той е дефинирал и доказал теорема, известна като ***теорема на Бернули***, която гласи, че ако при независими опити (изпитвания) вероятността на някакво събитие е *p*, вероятността, относителната (релативната) честота $\frac{m}{n}$ на появяване на събитието, удовлетворяваща неравенството $\left(\frac{m}{n} - p\right) < \varepsilon$, където ε е произволно малко положително число, става произволно близка до единица при достатъчно голям брой на опитите (изпитванията) - *n*. По друг начин казано: при достатъчно голям брой (*n*) независими опити може с вероятност произволно близка до единица да се твърди, че относителната честота ще се различава произволно малко от вероятността:

$$P\left(\left|\frac{m}{n} - p\right| < \varepsilon\right) \to 1. \qquad (4.16)$$

Бележитият френски математик и физик ***Симон Дени Поасон (1781-1840)***, който за пръв път въвежда термина закон за големите числа, е доказал по-обща теорема за случая, когато при независими опити вероятността на събитието А приема стойности $p_1, p_2, ..., p_n$ (по реда на опитите). Тя гласи, че в този случай при достатъчно голямо *n* относителната (релативната) честота $\frac{m}{n}$ ще се различава произволно малко от средната аритметична на вероятностите ($\bar{p}$), т.е.

$$P\left(\left|\frac{m}{n} - \bar{p}\right| < \varepsilon\right) \to 1. \qquad (4.17)$$

По-генералното математическо обобщение на закона за големите числа принадлежи на знаменития руски математик ***Пафнутий Львович Чебишев (1821-1894)***. Той е доказал теорема (***теоремата на Чебишев***), от която се извежда важно за статистиката следствие.

При определени условия (те могат точно да се установят) може да се твърди с вероятност, близка до единица, че средната аритметична на голям брой независими случайни величини ще се различава произволно малко от математическото им очакване:

$$P\left(\left|\bar{x} - E(x)\right| < \varepsilon\right) \to 1. \qquad (4.18)$$

Въз основа на теоремата на Чебишев е изведена ***централната пределна теорема***. От нея произтичат съществени методологични постановки в областта на теорията на репрезентативните статистически изследвания.

Представената най-общо логическа същност на закона за големите числа, както и свързаните с него математически теореми, дават значителен тласък в развитието на теорията на вероятностите и на статистическата методология.

4.5. Основни теоретични разпределения

В теорията на вероятностите са открити и изследвани различни теоретични разпределения. Всяко от тях има свои особености и е подчинено на определен закон на разпределение. В теорията на статистиката имат по-широко приложение разпределенията: биномно, поасоново, нормално, хипергеометрично, и известните като извадкови, *t*-разпределение, χ^2-разпределение, *F*-разпределение.

4.5.1. Биномно разпределение

Биномното разпределение е разпределение на прекъсната (дискретна) случайна величина. Нарича се още ***разпределение на Бернули***, тъй като за пръв път е открито и описано от ***Якоб Бернули***. Първоначално то е било дефинирано, изследвано и проверено опитно във връзка с хазартните игри. След това обаче е развито за общ случай, когато

при опити (наблюдения) възможните изходи са два (алтернативни събития): положителен или отрицателен, успех или неуспех, "да" или "не" и т.н. При емпиричните изследвания, както е известно, в такива случаи се прилага дихотомна скала.

Най-достъпно е обяснението му с примера за подхвърляне на монета. При всеки опит, т.е. при всяко подхвърляне на една монета, са възможни два еднакво вероятни резултата - ези и тура с вероятности ½ и ½. За двата възможни резултата вероятностите могат да се означат с p и q, където $p+q=1$. Такива опити, независими един от друг, които се повтарят многократно и имат само два възможни изхода при непроменящи се вероятности, еднакви за всички опити, се наричат ***опити на Бернули***. Когато при конкретни изследвания задачата може да се формулира чрез условията при опитите на Бернули, може да се използува биномното разпределение.

И така, при подхвърляне на една монета възможните резултати (изходи) са ези (Е) и тура (Т). При подхвърляне на две монети (или една монета два пъти) отделните последователни резултати са: ЕЕ, ЕТ, ТЕ, ТТ. При подхвърляне на три монети (или една монета три пъти) - ЕЕТ, ЕТЕ, ТТЕ, ЕТТ, ТЕТ, ТЕЕ, ТТТ, ЕЕЕ и т.н.

Броят на различните резултати, независимо от подреждането на Е и Т, като продължава увеличаването на броя на монетите в сериите (или броя на подхвърлянията на една монета), ще изглежда като подреждането, посочено в табл. 4.2.

Таблица 4.2

Брой на единиците в серията	Брой на отделните еднакви резултати	Брой на всички възможни резултати
1	1 1	$2 = 2$
2	1 2 1	$4 = 2^2$
3	1 3 3 1	$8 = 2^3$
4	1 4 6 4 1	$16 = 2^4$
5	1 5 10 5 1	$32 = 2^5$
	и т.н.	

Ако броят на отделните еднакви резултати при съответен броя единици в серията се раздели на броя на всички възможни резултати, ще се получат вероятностите

1/2 1/2
1/4 2/4 1/4
1/8 3/8 3/8 1/8
1/16 4/16 6/16 4/16 1/16
1/32 5/32 10/32 10/32 5/32 1/32
и т.н.

Вижда се че, броят на отделните еднакви резултати (табл. 4.2), както и вероятностите, се подреждат във формата на триъгълник. Той е известен в литературата като ***триъгълник на Блез Паскал (1623-1662)***.

Числата, които се съдържат в табл. 4.2, показващи отделните еднакви резултати, всъщност са биномните коефициенти в развитието на ***Нютоновия бином***.

При приетите означения с p и q на вероятностите по опитите на Бернули биномът $(q+p)^n$ ще има развитие:

$$(4.19) \qquad (q+p)^n = q^n + C_n^1 q^{n-1} p + C_n^2 q^{n-2} p^2 + ... + C_n^{n-1} q p^{n-1} + p^n .$$

Вероятността на всеки възможен резултат следователно е

$$(4.20) \qquad P_{n(k)} = C_n^k p^k q^{n-k} ,$$

където

$P_{n(k)}$ е вероятността на всеки възможен резултат при n единици в серията;

n - броят на единиците в серията;

k - броят на сбъдванията на едното събитие /в примера - едната страна на монетата/;

n-k - броят на сбъдванията на противоположното събитие /в примера - другата страна на монетата/;

p - вероятността за сбъдването на едното събитие при отделен опит;

q - вероятността за сбъдването на противоположното събитие при отделен опит;

C_n^k - комбинациите на n елементи от k -ти клас.

Известно е от математиката, че

(4.21) $$C_n^k = \frac{n!}{k!(n-k)!}.$$ [1]

Като се замести този израз на C_n^k във формула 4.20 ще се получи

(4.22) $$P_{n(k)} = \frac{n!}{k!(n-k)!} p^k q^{n-k}.$$

Тъй като $q = 1 - p$, формулата може да се запише и така:

(4.23) $$P_{n(k)} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}.$$

Формулата за изчисляване на вероятността на всеки възможен резултат при биномното разпределение се нарича ***формула на Бернули***, а разпределението се нарича ***биномно разпределение***.

За удобство са съставени таблици, по които се намират вероятностите за дадени стойности на k, n и p. (Вж. приложение 1).

Ще илюстрираме изчисляването на биномните вероятности и използването на таблиците с ***пример***.

В машиностроителен завод се произвеждат два вида лагери. При опаковането е трябвало в 40 амбалажни каси да се поставят лагери от вида А. В 10 от касите обаче погрешно са поставени лагери от вида Б. Касите са затворени, означени с етикети за лагери А и наредени произволно в склада. Каква е вероятността в 5 случайно взети каси да се окаже, че броят на тези с лагери от вида Б е : 0, 1, 2, 3, 4,5 ?

Вероятността се съдържа в приложение 1, в колоната за $p = 0{,}25$, тъй като 10/40 = 0,25. За $n = 5$ в колоната намираме следните вероятности (табл.4.3):

[1] Известният от математиката знак (!), наречен ***факториел*** означава, че числото пред него се умножава последователно на всички предходни числа в обратен ред до 1. Например 6! Значи 6.5.4.3.2.1 = 720.

Таблица 4.3

k	0	1	2	3	4	5
$P_{n(k)}$	0,2373	0,3955	0,2637	0,0879	0,0146	0,0010

Таблицата в посоченото приложение съдържа стойности за p от 0,05 до 0,50. Може обаче да се наложи намирането на разпределението на вероятностите при $p > 0,5$. Ако в примера приемам, че p не е 0,25, а 0,60, може да се интересуваме от вероятността примерно 3 каси да съдържат лагери от вида Б ($k = 3$) при $n = 5$.

Комбинациите имат едно свойство, според което $C_n^k = C_n^{n-k}$.

Освен това $p^k \cdot q^{n-1} = p^{n-k} \cdot q^k$. Може да се докаже, че вероятността едно събитие да се сбъдне k пъти от n опита при вероятност за отделен опит p е равна на вероятността то да се сбъдне n-k пъти от n опита при вероятност за отделен опит 1-p:

$$Pn(k;p) = P(n-k;1-p).$$

При условието в примера вместо $n = 5$, $k = 3$ и p = 0,6 записваме $n = 5$, $k = 5 - 3 = 2$ и $p = 1-0,6 = 0,4$. Търсената вероятност се намира в таблицата в колоната за $p = 0,4$ и тя е 0,3456.

Биномното разпределение, както и всяко друго теоретично разпределение, има параметрите математическо очакване и дисперсия. Така за всяко едно от двете алтернативни събития (изходи от опита) се приемат условни значения 1 и 0. Тогава за отделен единичен опит математическото очакване ще бъде

$$E(\mathrm{k}) = 1p + 0q = p, \qquad (4.24)$$

а дисперсията

$$\sigma^2 = (1 - p)^2 p + (0 - p)^2 q = pq \qquad (4.25)$$

и средното квадратично отклонение -

$$\sigma = \sqrt{pq} = \sqrt{p(1-p)}. \qquad (4.26)$$

При n независими един от друг опити, при които вероятността остава неизменна, математическото очакване е

(4.27) $E(k;n) = p + p + p + ... + p = np.$

Дисперсията в този случай е

(4.28) $\sigma^2 = npq,$

а средното квадратично отклонение -

(4.29) $\sigma = \sqrt{npq}$.

Ако вместо случайната величина k (броят на сбъдванията на едно събитие при n опита) се разглежда $\frac{k}{n}$ като относителна честота на сбъдването на събитието, математическото очакване ще бъде

(4.30) $E\left(\frac{k}{n}\right) = \frac{np}{n} = p$,

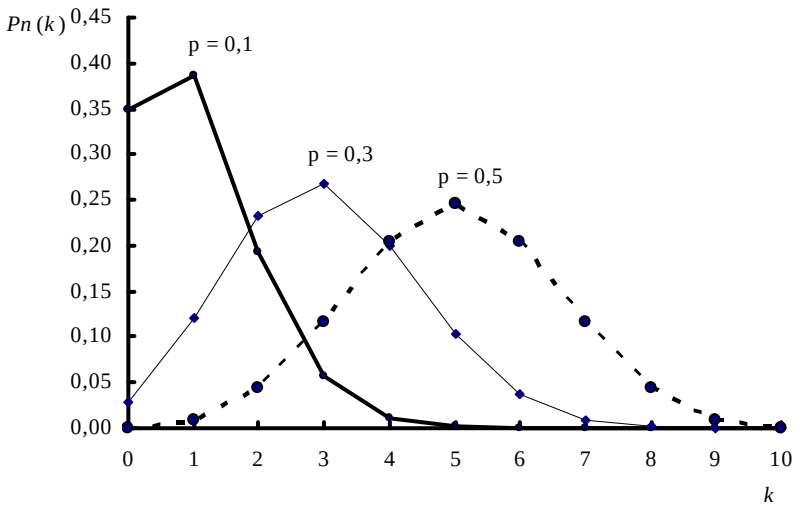
дисперсията -

(4.31) $\sigma^2 = \frac{npq}{n^2} = \frac{pq}{n}$

и средното квадратично отклонение -

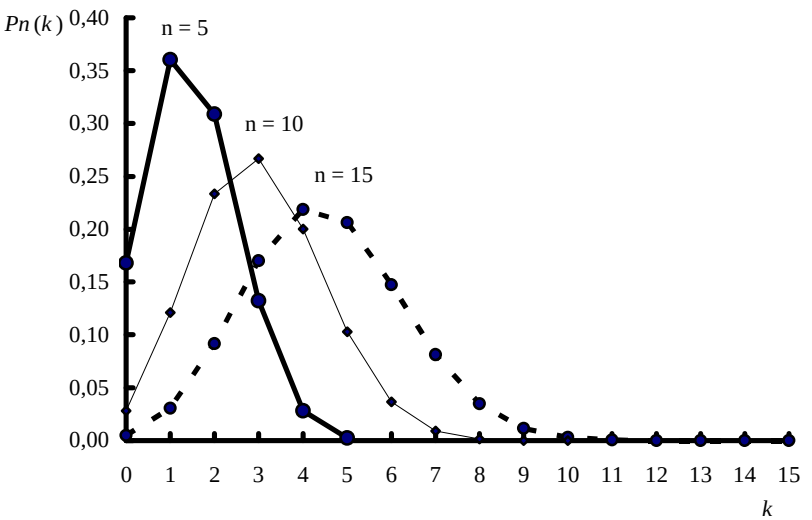
(4.32) $\sigma = \sqrt{\frac{pq}{n}}$.

Биномно разпределение (n = 10, p = 0,1; p = 0,3; p = 0,5)



Фиг. 4.4

Биномно разпределение (p = 0,3, n = 5; n = 10; n = 15)



Фиг. 4.5

Формата на биномното разпределение зависи от p и n. Ако $p = q$, т.е. ако $p = 0,5$, биномното разпределение е симетрично. Ако $p \neq q$, но те остават неизменни, а n (винаги цяло положително число) се увеличава, асиметрията намалява. Когато $p < 0,5$, асиметрията е положителна, т.е. кривата е с полегато дясно рамо, а когато $p > 0,5$ - асиметрията е отрицателна, т.е. кривата е с полегато ляво рамо. На фиг. 4.4 е показано как се изменя формата на разпределението при еднакво n и различно p,а на фиг. 4.5 - при еднакво p и различно n.

4.5.2. Хипергеометрично разпределение

Хипергеометричното разпределение също е разпределение на прекъсната (дискретна) случайна величина, както и биномното. Но докато при биномното вероятностите за двете алтернативни събития (двете стойности на случайната величина) са постоянни, при хипер-геометричното те се променят.

Да приемем теоретично, че в една партида от N произведени изделия (генерална съвкупност) има M стандартни и $N - M$ нестандартни.

Прави се извадка от n изделия, като последователно попадащите в извадката не се връщат отново в генералната съвкупност.

В извадката ще попаднат k стандартни изделия и $n - k$ нестандартни.

Вероятността на стандартните преди да се направи извадката е $p = \frac{M}{N}$, а вероятността на нестандартните е $q = 1 - p = 1 - \frac{M}{N}$. Но след изтеглянето на всяка единица за образуване на извадката тя не се връща в генералната съвкупност и вероятността за останалите се променя. А последователното изтегляне на n изделия без връщане от гледна точка на резултата е равносилно на изтеглянето на n изделия предварително.

Вероятността при извадка от n изделия k от тях да бъдат стандартни е

$$P_{n(k)} = \frac{C_M^k . C_{N-M}^{n-k}}{C_N^n}. \qquad (4.33)$$

Когато на k се придават стойности 0, 1, 2, 3 и т.н. и по формулата се изчисляват вероятностите на случайната величина k, получава се ***хипергеометрично разпределение***. Това разпределение е подходящ теоретичен модел при извадкови изследвания, когато извадката се прави по схема без връщане (чрез безвъзвратен подбор). Теоретично погледнато, то се приближава асимптотично към биномното при увеличаване на N - обема на генералната съвкупност, от която се прави извадката (n).

Както всяко разпределение, така и хипергеометричното има своите параметри. Математическото очакване е

$$E(k) = np, \quad (4.34)$$

а дисперсията

$$\sigma^2 = \frac{N-n}{N-1}.npq. \quad (4.35)$$

Както се вижда дисперсията при хипергеометричното разпределение се различава спрямо биномното с множител $\frac{N-n}{N-1}$, който е коректив за промяната във вероятностите за единиците в генералната съвкупност да попаднат в извадката след изваждането на всяка единица без връщането й в генералната съвкупност (към този въпрос ще се върнем в гл. 5).

4.5.3. Поасоново разпределение

Поасоновото разпределение може да се разглежда, от една страна, като пределна форма на биномното (когато n е достатъчно голямо, а p е много малко, клонящо към 0), а от друга страна, като разпределение, което описва т.нар. ***Поасонов процес, на името на Симон Поасон***.

Ако при опитите на Бернули n се увеличава неограничено, а p намалява и се стреми към 0, като np остава постоянно число, равно на λ (ламбда), която е от порядъка на няколко единици, тогава вероятността даденото събитие да се сбъдне k пъти (брой на появяванията на събитието в серия от еднакви независими опити) е

$$(4.36) \qquad P_{n(k)} = \frac{(np)^k}{k!} e^{-(np)} ,$$

или като се замести np с λ, се получава

$$(4.37) \qquad P_{n(k)} = \frac{(\lambda)^k}{k!} e^{-(\lambda)} .$$

Разпределението на случайната величина k, изразено чрез формулата, е известно като ***Поасоново разпределение***. То е ***дискретно***, т.е. разпределение на прекъсната (дискретна) случайна величина.

За вероятностите $P_{n(k)}$ е съставена таблица, по която те могат да се намерят за всяка стойност на λ. (вж. приложение 2).

Характерно за Поасоновото разпределение е това, че математическото му очакване и дисперсията са равни помежду си и се изразяват с $\lambda = np$, която е единствен параметър. Това е лесно обяснимо, като се има предвид, че в биномното разпределение математическото очакване е np, а дисперсията - npq. При Поасоновото разпределение $p \to 0$ и следователно $q = (1-p) \to 1$. От това следва, че $npq = np$.

В общия случай, когато $np \leq 5$, Поасоновото разпределение може да служи като апроксимация на биномното. То е подходящо за практически цели именно, когато $\lambda = np$ е малко число (няколко единици). Прилага се при изследване на редица процеси в областта на статистическия контрол на качеството, в областта на масовото обслужване и др., при които се наблюдава многократно повтаряне на определено събитие през някакъв интервал. Такъв процес се дефинира като Поасонов процес.

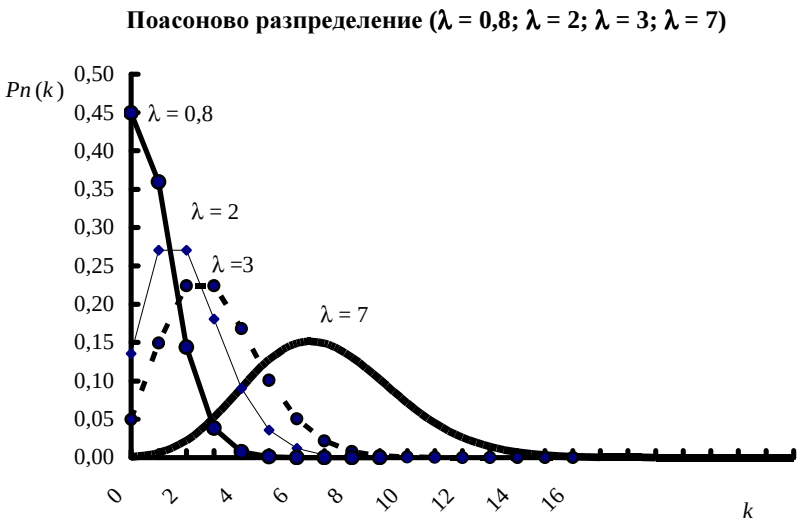
Да приемем ***например***, че в дадена фирма постъпват поръчки по телефона за доставка на стоки. В определен интервал, примерно 30 минути, разделен на по-малки интервали (субинтервали) - примерно 5 минути, се получават определен брой поръчки по телефона. Ако си представим, че увеличаваме неограничено броя на субинтервалите, които следователно стават много малки, ще стигнем до положение, при което в даден субинтервал или има или няма поръчка. Това може да се твърди с вероятност почти равна на единица (сигурност), тъй като вероятността за повече от една поръчка в един субинтервал е нищожно малка. Такъв

именно процес се нарича Поасонов. Може да се установи каква е вероятността да се получат k поръчки в даден интервал с продължителност T. Доказано е, че вероятността е

$$P_{n(k)} = \frac{(\lambda T)^k}{k!} e^{-(\lambda T)} \quad . \tag{4.38}$$

Тъй като λ е постоянно число и в случая означава очакваният брой поръчки за единица интервал (плътност на поръчките), T означава очакваният брой поръчки за даденото време.

Много процеси, като контролът върху качеството на даден вид изделие на поточна линия, масовото обслужване с чакане (ремонт на уреди, отстраняване на аварии в електроснабдяването и др.), кацането на самолети на дадено летище, трудовите злополуки в производството и т.н. могат да се разглеждат като Поасонов процес и да се изследват в определен статистически аспект с помощта на Поасоновото разпределение.



Фиг. 4.6.

Формата на кривата на Поасоновото разпределение зависи от стойността на λ. Това е показано на фиг. 4.6.

4.5.4. Нормално разпределение

Нормалното разпределение е непрекъснато, т.е. разпределение на непрекъсната (индискретна) случайна величина.

Първооткривател на нормалното разпределение е английският математик от френски произход ***Абрахам дьо Моавър (1667-1754)*** и е определено от него като непрекъсната форма на биномното разпределение. В началото на XIX в. независимо от Моавър нормалното разпределение е било открито, изследвано и описано от видния немски математик ***Карл Фридрих Гаус (1777-1855)*** и френския математик, физик и астроном ***Пиер Симон Лаплас (1749-1827)***, които са го приложили в теорията на случайните грешки. Затова често се нарича ***Гаусово*** или ***Гаус - Лапласово разпределение***.

Кривата, описваща нормалното разпределение, която е идеално симетрична, се нарича ***нормална крива***.

Това разпределение има изключително голямо значение за разработването и прилагането на редица методологични положения в теорията на статистиката. В много случаи се разглежда като теоретичен модел на закономерности в масовите явления, които се намират под действието на множество независими случайни причини.

Може най-общо да се каже, че когато една случайна величина има множество стойности и те са резултат на много и независими помежду си фактори, тя има нормално разпределение.

Нормалното разпределение може да се представи аналитично чрез неговата функция на плътността на вероятностите (диференциална функция на разпределението):

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\bar{x})^2}{2\sigma^2}}, \qquad (4.39)$$

където π и e са известни математически константи (π = 3,141593; e = 2,718282).

Функцията на разпределението на вероятностите (интегралната функция) на нормалното разпределение има вида:

$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}}dx \quad . \tag{4.40}$$

Нормалното разпределение има два параметъра - математическо очакване $E(X) = \bar{x}$ и дисперсия, респ. средно квадратично отклонение.

Съгласно известната обща формула на математическото очакване на индискретна случайна величина, ***математическото очакване*** при нормално разпределение е

$$E(X)=\int_{-\infty}^{\infty} xf(x)dx \quad . \tag{4.41}$$

Дисперсията е

$$\sigma^2=\int_{-\infty}^{\infty}(x-E(X))^2\, f(x)dx \quad , \tag{4.42}$$

а стандартното отклонение -

$$\sigma=\sqrt{\int_{-\infty}^{\infty}(x-E(X))^2\, f(x)dx} \quad . \tag{4.43}$$

Практически са възможни различни конкретни стойности на математическото очакване и на стандартното отклонение и следователно са възможни много нормални разпределения. Това създава неудобство при статистическите изследвания. Неудобството се преодолява чрез формирането на нормално разпределение на стандартизирана случайна величина. Това ще рече да се намери функцията на разпределението на стандартизираните (нормираните) отклонения $z_i=\frac{x_i-\bar{x}}{\sigma}$.

Така ще се получи нормирано или ***стандартно нормално разпределение***, което има математическо очакване 0 и дисперсия 1. Неговата функция на плътността на вероятностите (диференциална функция на разпределението) ще има вида

$$(4.44) \qquad f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}.$$

Интегралната функция на разпределението съответно има вида:

$$(4.45) \qquad F(z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}e^{-\frac{z^2}{2}}dz \ .$$

Функцията по формула 4.45 се отнася само за половината от площта под нормалната крива. За да се обхване цялата площ, т.е. вероятностите за всички стойности на случайната величина, е удобно за практически цели да се използва функцията

$$(4.46) \qquad F(z)=\frac{2}{\sqrt{2\pi}}\int_{0}^{z}e^{-\frac{z^2}{2}}dz \ .$$

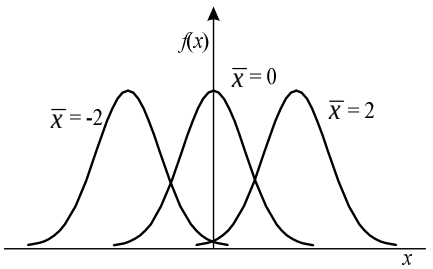
Интегралната функция дава възможност при всяка нормално разпределена случайна величина да се определя вероятността нейна стойност да се намира в даден интервал или да се определя относителния дял на площта под нормалната крива, съответстващ на даден интервал.

За интегралната и диференциалната функция на нормалното разпределение има съставени стандартни таблици, които се използват при статистическите изследвания без да се налага всеки път да се прилагат формулите (вж. приложение 3 и 4)

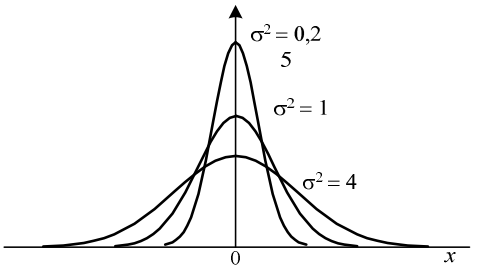
Нормалното разпределение, респективно нормалната крива, имат някои характерни ***свойства***.

1. Нормалното разпределение се определя напълно от математическото очакване (средната) и стандартното отклонение. Математическото очакване определя центъра на разпределението, а σ - формата на кривата. При по-малко σ кривата е по-стръмна, а при по-голямо σ - по-полегата. На фиг. 4.7 са показани три криви на нормалното разпределение с три различни стойности на средната аритметична (математическото очакване), а с еднакви дисперсии. На фиг. 4.8 са показани три криви на нормално разпределение с различни дисперсии, а с еднакви средни аритметични величини.

2. Кривата на нормалното разпределение е напълно симетрична по отношение на вертикалната линия, пресичаща абсцисната ос в точка $\overline{x}$.



Фиг. 4.7



Фиг. 4.8

3. Максимумът на функцията (максималната ордината на кривата) е

$$y_{max} = \frac{1}{\sigma\sqrt{2\pi}}. \qquad (4.47)$$

При $\sigma = 1$, $y_{max} = 0{,}398942 \approx 0{,}4$. За $x \to \infty$, y се приближава асимптотично към абсцисната ос, т.е. стреми се към 0.

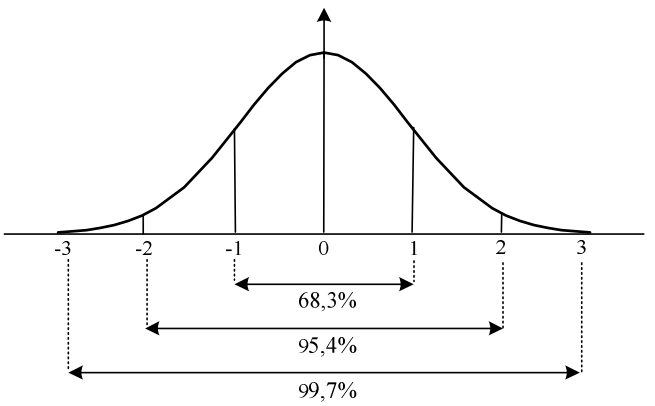
4. Стандартното (средното квадратично) отклонение определя абсцисите на инфлексните точки на кривата, т.е. точките, в които кривата преминава от изпъкнала в длъбната и обратно. Ординатите на инфлексните точки са равни приблизително на $0{,}6y_{max}$. За стандартно нормално разпределение инфлексните точки се намират при $z = \pm 1$.

5. Между ординатите $-z\sigma$ и $+z\sigma$ винаги се обхваща еднаква част от площта под нормалната крива, каквато и да е стойността на σ. Това означава, че делът от площта под нормалната крива, затворена в тези граници, зависи само от z, а не и от σ.

В границите от $-z\sigma$ до $+z\sigma$ се съдържа 0,6827 (68,27%) от площта под нормалната крива, в границите от -2σ до $+2\sigma$ - 0,9545 (95,45%), а в границите от -3σ до $+3\sigma$ - 0,9973 (99,73 %), т.е почти цялата площ.

Това свойство е особено важно. Делът от площта под нормалната крива съответствува на вероятността дадена стойност на случайната величина да се намира в определен интервал. Следователно може с

определена вероятност да се твърди, че дадена характеристика, която има нормално разпределение, няма да се намира зад пределите на $\pm z\sigma$.



Фиг. 4.9

На фиг. 4.9 е показано разпределение на площта под нормалната крива.

При конкретния анализ за практически цели се използват обикновено интервалите от $-1{,}96\sigma$ до $+1{,}96\sigma$, който обхваща 95 % от площта под нормалната крива и от $-2{,}58\sigma$ до $+2{,}58\sigma$, обхващащ 99 %. Това ще рече вероятности 0,95 и 0,99.

За практическото използуване на нормалното разпределение имат съществено значение и някои други негови свойства, изведени в ***теореми***.

1. Ако $X_1, X_2, X_3, ..., X_n$ са нормално разпределени независими случайни величини, тяхната сума е също нормално разпределена. Както беше посочено, математическото очакване и дисперсията на сума от независими случайни величини могат да се намерят като сбор от математическите очаквания и на дисперсиите на отделните случайни величини.

2. Горната теорема може да се разшири: ако $X_1, X_2, X_3, ..., X_n$ са нормално разпределени независими случайни величини, а α, β, γ..., са

някакви константи, тогава сумата от произведенията $\alpha_1 X_1 + \beta_2 X_2 + \gamma_3 X_3 + ...$ също е нормално разпределена.

3. Сумата на достатъчно голям брой независими случайни величини, които не са нормално разпределени, е ***асимптотично нормална величина***, т.е. тя се стреми към нормално разпределение, с увеличаване на броя на случайните величини, които се сумират. При решаването на голям брой задачи в различни области от действителността случайните величини се разглеждат като резултат от много независими фактори и условно могат да се приемат като сума от много независими случайни величини. Общата случайна величина в този случай се стреми към нормално разпределение.

От централната пределна теорема и от други теореми, свързани със закона за големите числа, следва, че каквото и да е разпределението в дадена генерална съвкупност, при достатъчно голям брой извадки с обем *n*, техните средни величини имат нормално разпределение с обща средна аритметична, равна на средната аритметична на генералната съвкупност, и дисперсия ($\sigma_{\bar{x}}^2$) *n* пъти по-малка от дисперсията на генералната съвкупност (σ_o^2):

$$\sigma_{\bar{x}}^2 = \frac{\sigma_o^2}{n}. \qquad (4.48)$$

Стандартното отклонение ($\sigma_{\bar{x}}$) съответно е

$$\sigma_{\bar{x}} = \frac{\sigma_o}{\sqrt{n}}. \qquad (4.49)$$

Това положение създава големи възможности за анализ и особено за оценяване на параметрите на генералните съвкупности по характеристиките на извадките.

Посочените теоретични положения придобиват по-конкретен практически смисъл при емпиричните изследвания.[1]

[1] Повече за нормалното разпределение, за логаритмичната му форма, както и за други теоретични разпределения, вж. **Сугарев, З. И С. Каменаров**, цит. съч.

Нормалното разпределение намира широко приложение при статистическите изследвания като теоретичен модел. Това важи и при извадковите изследвания, но само когато извадките са достатъчно големи. При сравнително малки извадки не се получават нормални разпределения. Следователно в такива случаи използването на свойствата на нормалното разпределение е некоректно. Понятието малка извадка не трябва да се разбира като определено постоянно число. То зависи от търсената обобщаваща характеристика. При оценяване на средна аритметична например, извадката (n) е малка, когато е до 30 единици, при оценяване на дисперсия - до 100 единици и т.н. Изобщо извадката е малка, когато при нейния обем не се получава нормално разпределение на съответната характеристика.

Тъй като при много случаи се налага изследванията да се правят чрез малки извадки, в теорията на статистиката са открити и по съответен начин са дефинирани ***извадкови разпределения***. Те не са нормални в разглеждания смисъл, но са свързани с нормалното разпределение. Това са t-разпределението (на Стюдент), χ^2-разпределението (на Пирсън) и F-разпределението (на Фишер).

4.5.5. *t*-разпределение

Да допуснем теоретично, че от генерална съвкупност с неизвестно разпределение се правят извадки и се изчисляват средни аритметични величини ($\bar{x}_i$). Те ще варират около общата им средна ($\bar{x}$). Разликите ($\bar{x}_i - \bar{x}$) могат да се стандартизират, като се разделят на стандартното отклонение на средните, което по формула 4.49 е $\sigma_{\bar{x}} = \frac{\sigma_o}{\sqrt{n}}$. Така ще се получи стандартизирана случайна величина (t):

$$t_i = \frac{\bar{x}_i - \bar{x}}{\frac{\sigma_o}{\sqrt{n}}} \quad . \tag{4.50}$$

При малки извадки тази стандартизирана случайна величина няма нормално разпределение. Английският статистик ***Уилям Госет (1876-***

1937), с псевдоним ***Стюдент***, е изследвал нейното разпределение и е дефинирал съответната функция на плътността. Това разпределение се нарича ***t-разпределение*** или ***разпределение на Стюдент*** със степени на свобода $\phi = n - 1$.

Степените на свобода **(**ϕ**)** могат да се дефинират най-общо като число, показващо броя на стойностите на случайната величина, които могат свободно да варират, без с това да се изменя дадена обща характеристика. Да приемем например, че са дадени 5 значения на признака - 4, 8, 12, 9, 2. Тяхната обща сума е $\sum x = 35$, а средната аритметична е $\bar{x} = \frac{35}{5} = 7$. Да приемем, че признакът на отделните единици може да има други значения, но при условие че се запазва непроменена общата средна ($\bar{x}$). В такъв случай броят на свободно вариращите значения на признака е n - 1 = 5 - 1 = 4, тъй като петото значение е ограничено от поставеното условие (да се запази същата средна) и трябва да допълни общата сума до 35. Ако съвкупността е разделена на три групи с брой на единиците n_1, n_2, n_3, и със съответни средни аритметични $\bar{x}_1, \bar{x}_2, \bar{x}_3$ за всяка отделна група степените на свобода ще бъдат $(n_1 - 1), (n_2 - 1), (n_3 - 1)$. Или общо степените на свобода могат да се означат с $\phi = n - 1$. За трите групи, взети заедно, степените на свобода ще бъдат $(n_1 - 1) + (n_2 - 1) + (n_3 - 1) = n - 3$. В общия случай, ако означим броя на групите с *k*, степените на свобода ще бъдат $\phi = n - k$. За груповите средни $\bar{x}_i$, които имат обща средна $\bar{x}$, степените на свобода ще бъдат $\phi = k - 1$.

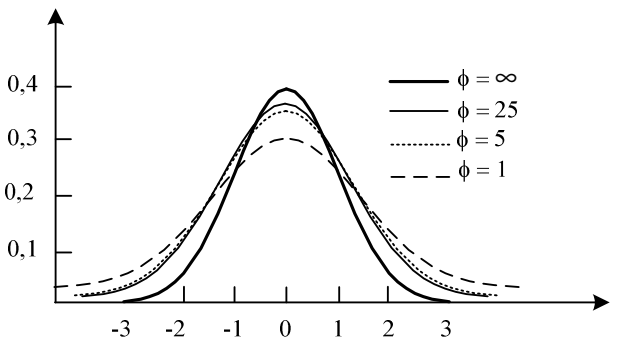
При практическото използуване на *t*-разпределението не се налага функцията му да се изчислява, тъй като има съставени стандартни таблици, по които стойностите й се намират при дадените степени на свобода (вж. приложение 5).

Характерно за *t*-разпределението е, че то, както нормалното разпределение, има симетрична, камбановидна крива.

Формата на кривата обаче зависи само от един единствен параметър - степените на свобода (ϕ). Колкото ϕ е по-малко, толкова *t*-разпределението повече се отдалечава от нормалното, тъй като неговата

крива е по-широка в основата. При увеличаване на ϕ, което означава увеличаване на *n*, *t*-разпределението се доближава до нормалното, а при достатъчно голямо ϕ - би се покрило с нормалното.

На фиг. 4.10. са показани криви на *t*-разпределението при различни степени на свобода, както и нормалната крива за сравнение.



Фиг. 4.10

Известно е, че при нормално разпределение 5 % от площта под кривата лежи зад пределите $\pm 1{,}96\sigma$, а 1 % е зад пределите $\pm 2{,}58\sigma$. При *t*-разпределението, ако ϕ = 5, тези граници са съответно ±2,57 и ±4,03. Ако обаче ϕ = 10, те са ±1,98 и ±2,6, т.е. твърде много се доближават до границите при нормално разпределение. Иначе казано, ако използваме информация от малки извадки и правим вероятностни заключения, опирайки се на нормалното разпределение, можем да допуснем съществени грешки. В такива случаи трябва да се опрем на *t*-разпределението, или на друго извадково разпределение според познавателната задача и интересуващите ни параметри.

t-разпределението има широко приложение в статистическите изследвания и в частност при проверката на хипотези, която се разглежда в гл. 6.

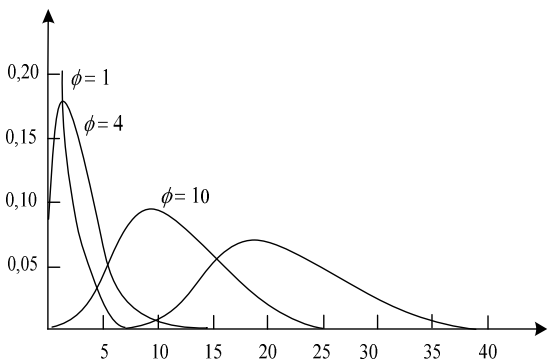
4.5.6. χ^2 - разпределение

Изследвайки "поведението" на сумата от квадратите на нормално разпределени случайни величини, английският математик, биолог и статистик ***Карл Пирсън (1857-1936)*** описва функцията на едно ново разпределение с оглед практическото й използване при извадкови изследвания и нарича разпределението ***χ^2-разпределение*** (от гр. буква хи). В съвременната литература то се нарича още разпределение на Пирсън.

Може да се дефинира: случайната величина, представляваща сума от квадратите на n независими случайни величини със стандартно нормално разпределение, се нарича случайна величина с χ^2-разпределение и със степени на свобода $\phi = n$.

Пирсън е дефинирал функцията на плътността на това разпределение, за което също има съставена стандартна таблица (вж. приложение 6).

Кривата му е асиметрична с положителната асиметрия, но с увеличаване на $\phi = n$ се стреми към нормалната крива, т.е. формата на χ^2-разпределението, както и на t-разпределението, зависи само от степените на свобода. На фиг. 4.11. са показани криви на χ^2-разпределение при различно число степени на свобода.



Фиг. 4.11

χ^2-разпределението се използува при статистическата проверка на хипотези и в частност при проверката на съответствието между

емпирични и теоретични разпределения (вж. гл. 6), при проверка на зависимостта между определени признаци, при проверка на принадлежността на две извадки към една обща съвкупност и др.

4.5.7. *F*-разпределение

Ако две независими случайни величини U_1 и U_2 имат χ^2-разпределение със степени на свобода ϕ_1 и ϕ_2, отношението

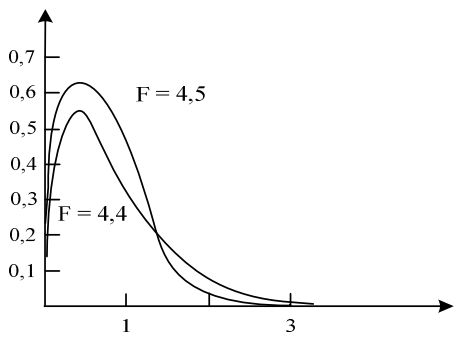
$$F = \frac{U_1}{\phi_1} : \frac{U_2}{\phi_2} \qquad (4.51)$$

има разпределение, наречено ***F-разпределение*** или ***разпределение на Фишер*** в чест на неговия откривател ***Роналд Фишер (1890-1962)***. То има степени на свобода

$$\phi_1 = (n_1 - 1) \text{ и } \phi_2 = (n_2 - 1) .$$

По форма *F*-разпределението е асиметрично с положителна асиметрия и област на изменение от 0 до ∞. Формата на кривата му зависи от степените на свобода ϕ_1 и ϕ_2. С увеличаване едновременно на степените на свобода ϕ_1 и ϕ_2 се стреми към нормалното разпределение. На фиг. 4.12 са показани две криви на разпределението, от които едната е при ϕ_1 = 4 и ϕ_2 = 5, а другата - при ϕ_1 = 4 и ϕ_2 = 4.

За *F*-разпределението също има съставени стандартни таблици (вж. приложение 7 и 8).



Фиг. 4.12

Както ще стане ясно в следващите глави, *F*-разпределението разкрива твърде големи възможности пред статистическия анализ. По-специално то стои в основата на дисперсионния анализ, който се прилага в различни области.

4.6. Практикум

4.6.1. Въпроси за самопроверка

1. Кои величини се наричат случайни? Примери?
2. Каква е връзката между вероятностите и относителните (релативните) честоти?
3. В какво се състои разликата между теоретичните и емпиричните разпределения?
4. Какво се разбира под закон на разпределението и по какъв начин може да се представи?
5. Какъв е смисълът на функцията на разпределението на вероятностите (интегралния закон) и функцията на плътността на вероятностите (диференциалния закон)?
6. Кои са основните параметри на теоретичните разпределения?
7. Каква е разликата между биномното, хипергеометричното и Поасоновото разпределение?
8. Кои са параметрите на нормалното разпределение?
9. Какво е математическото очакване и дисперсията на стандартното нормално разпределение?
10. Кои разпределения се наричат извадкови?
11. Какво се разбира под степени на свобода?
12. Каква е формата на *t*-разпределението и от какво зависи тя?
13. Каква е формата на кривата на χ^2-разпределението?
14. От какво зависи формата на кривата на *F*-разпределението?

4.6.2. Задачи за упражнение

Задача 1. В камион за международни превози (TIR) са натоварени 200 кашона, от които 150 са с консерви, а в 50 с етикет на консерви са поставени цигари с цел контрабанден износ.

Какви са вероятностите при митническа проверка на 10 случайно избрани кашони да се попадне на (k) 0, 1, 2, 3, 4, 5, 6, 10 кашона с цигари?

Решение: Търсените вероятности се намират в приложение 1 за функцията на биномното разпределение. По условие проверените кашони (извадката) са $n = 10$, а относителният дял на кашоните с цигари в камиона е $p = \frac{50}{200} = 0{,}25$ или 25 % (с консерви са $q = 1 - p = 0{,}75$ или 75 %).

В таблицата за вероятностите на биномното разпределение в колоната за $p = 0{,}25$ на редовете $n = 10$ се намират търсените вероятности:

$P_{10(0)} = 0{,}0563$ $\qquad$ $P_{10(4)} = 0{,}1460$

$P_{10(1)} = 0{,}1877$ $\qquad$ $P_{10(5)} = 0{,}0584$

$P_{10(2)} = 0{,}2816$ $\qquad$ $P_{10(6)} = 0{,}0162$

$P_{10(3)} = 0{,}2503$ $\qquad$ $P_{10(10)} = 0{,}0000$

Най-голяма е вероятността в извадката (проверените кашони) да попаднат 2 кашона с цигари.

Задача 2. В център за спешна медицинска помощ са постъпвали средно по 4 души на час.

Каква е вероятността за определен ден да постъпят за час 2, 3, 4, 5, 6, 7 пациенти?

Решение: Броят на постъпващите пациенти по спешност (k) има Поасоново разпределение с вероятности, намиращи се в стандартна таблица (приложение 2). В тази таблица за $\lambda = 4$ и за $k = 2{,}3{,}4{,}5{,}6{,}7$ вероятностите са:

$P_{4(2)} = 0{,}1465$ $\qquad$ $P_{4(5)} = 0{,}1563$

$P_{4(3)} = 0{,}1954$ $P_{4(6)} = 0{,}1042$

$P_{4(4)} = 0{,}1954$ $P_{4(7)} = 0{,}0595$

Задача 3. Разпределението на младежите донаборници, явили се пред комисията през 2005 г. е нормално по признака "ръст". Средният им ръст е 172 см, а стандартното отклонение - $\sigma = 5$ см.

Каква е вероятността случайно избран донаборник да има ръст между 172 см и 174 см?

Решение: Намира се стандартизираната разлика (нормираното отклонение)

$$z = \frac{x - \bar{x}}{\sigma} = \frac{174 - 172}{5} = \frac{2}{5} = 0{,}4$$

В стандартната таблица за функцията $F(z)$, т.е. за площта под нормалната крива (приложение 4) се намира за $z = 0{,}4$ вероятност 0,3108. Тъй като стойностите в таблицата се отнасят за цялата площ под нормалната крива за дадена стойност на z, намерената вероятност трябва да се раздели на 2, за да се намери търсената вероятност по условието на задачата, т.е. 0,3108:2 = 0,1554.

4.6.3. Из света на вероятностите

Известният английски учен статистик **Роналд Фишер** разказва следната забавна случка.[1]

В компания една дама твърдяла, че ако й поднесат чаша чай с мляко, тя може да познае дали първо е налят чаят и след това млякото, или обратно.

Въпросът е как да се провери дали дамата наистина има такава способност?

Един възможен вариант е да се предложат 2 чаши чай с различна последователност на наливането - първо чай, после мляко (ЧМ) и първо мляко, после чай (МЧ). От дамата да се поиска да избере една от чашите и

[1] Цит. по **Закс, Л.** Статистическое оценивание. Перевод с англ., М., 1976, с. 116.

да обяви последователността. Внимание! Дамата има ½ вероятност (50 % шанс) да познае съвсем случайно.

По-подходящ е друг вариант. Предлагат се 8 чаши, от които 4 с последователност ЧМ и 4 - с последователност МЧ, разположени на масата произволно. Дамата трябва да се покани да дегустира и да отдели 4 чаши с последователност ЧМ. От 8 чаши тя може да отдели 4 в 70 комбинации, от които само една е верен отговор. Вероятността да попадне случайно на верния е извънредно малка - 0,0143 или 1,43 %. Тази вероятност може да се намали още при увеличаване на броя на чашите.

5. СТАТИСТИЧЕСКИ ЗАКЛЮЧЕНИЯ. СТАТИСТИЧЕСКО ОЦЕНЯВАНЕ

> "Когато статистикът ... се основава на принципите на интервалното оценяване ... поставя на карта научната си репутация."
>
> Е. Кейн

Тази глава предлага изключително важни знания в една широка и бързо развиваща се област както на теорията, така и на практиката. Те са необходими за всеки, който прави извадкови изследвания или използва резултати от такива изследвания в сферата на икономиката, социологията, експерименталното дело, контрола върху процесите, бизнеса и др. Усвоявайки тези знания читателят ще разбере и ще спазва изискванията и условията за формирането на представителни извадки. Ще знае как чрез информацията, която извадките предлагат, може да прави оценки и заключения за неизвестни параметри на генералните съвкупности, които го интересуват, какви методи за оценки да избира, как да контролира максимално допустимите статистически грешки и др.

5.1. Същност на статистическите заключения

Много масови явления, които са обект на статистическо изучаване, не допускат изчерпателно обхващане на съвкупностите. Но и когато изчерпателното изучаване е възможно, то често изисква много време и средства и поради това се оказва нецелесъобразно. В такива случаи се прилага репрезентативният метод, т. е. наблюдават се непосредствено ***извадки*** от генералните съвкупности. Когато са направени така, че да представят добре генералните съвкупности, от които произлизат, те се наричат ***представителни (репрезентативни) извадки***. Когато се правят представителни (репрезентативни) изучавания, целта е чрез характеристиките на извадките да се направят заключения за неизвестните параметри на генералните съвкупности. За да бъдат извадките представителни и да се правят обосновани и верни заключения за генералните съвкупности, трябва да се спазват и прилагат определени принципи, теоретико-

методологични положения и процедури. Тези принципи, правила и процедури са предмет на разглеждане във важен дял на статистиката - ***теорията на статистическите заключения***.

Три основни положения характеризират статистическите заключения.

Първо, това е техният ***вероятностен характер***. Макар и да е достатъчно представителна една извадка, по своето съдържание, със своите характеристики не може да възпроизведе напълно и абсолютно точно генералната съвкупност. Тя я представя с определено приближение. Казано по друг начин, характеристиките на извадките се проявяват като случайни величини, свързани с определени вероятности. Затова статистическите заключения се опират върху основни положения на теорията на вероятностите, включително и върху теореми, свързани със закона за големите числа.

Второ, статистическите заключения се правят само въз основа на информация, получена от представителни (репрезентативни) извадки.

Трето, статистическите заключения се опират на съответни теоретични разпределения. Като случайни величини статистическите характеристики на извадките имат определени емпирични разпределения. Но изследването им предполага да се използват адекватни теоретични разпределения като модели.

В теорията на статистическите заключения са се формирали две направления (два клона): статистическо оценяване и статистическа проверка на хипотези.

Същността и целта на ***статистическото оценяване*** се състои в получаването от извадките на такива обобщаващи характеристики, които могат да се приемат като обосновани ***оценки*** на неизвестните параметри на генералните съвкупности. Теорията на статистическите заключения съдържа изискванията, на които трябва да отговарят оценките и методите, посредством които те се получават.

При ***статистическата проверка на хипотези*** предварително се формулират определени предположения (хипотези) относно параметрите на генералната съвкупност, нейните функции на разпределения и др., а

след това по информацията от извадките и дадени общи теоретични положения се проверява дали хипотезите се потвърждават или не.

По-нататък в тази глава се разглежда теорията на статистическото оценяване, а в следващата - статистическата проверка на хипотези.

5.2. Репрезентативни извадки

Беше вече подчертано, че статистическите заключения в двете им направления (правенето на оценки и проверката на хипотези) се основават върху представителни (репрезентативни) извадки. Ето защо е необходимо да се дефинират по-конкретно и по-определено принципите за формиране на такива извадки.

За да бъде една извадка представителна (репрезентативна) и да осигурява надеждни заключения за генералната съвкупност от която е излъчена, тя трябва да бъде направена по начин, гарантиращ еднаква възможност (еднаква вероятност) за всяка единица от генералната съвкупност да попадне в извадката. Казано по друг начин, извадката трябва да се формира чрез случаен подбор на единиците, които ще попаднат в нея. При случаен подбор отклоненията на характеристиките на извадките от съответния параметър на генералната съвкупност ще имат случаен характер, ще се подчиняват на "поведението" и на разпределенията на случайните величини и може по съответен начин и в определена степен да бъдат контролирани. Тези отклонения (разлики) между характеристиките на извадките и оценяваните неизвестни параметри на генералните съвкупности са конкретните ***стохастични (вероятностни) грешки***, които произтичат от обстоятелството, че характеристиките на извадките са получени от сравнително малък брой единици, а не от всички единици на генералната съвкупност. Ако извадките не са излъчени по принципа на случайния подбор, а е допусната каквато и да е преднамереност и изобщо при нарушаване на случайността, ще се получат ***систематични грешки***, които не се подчиняват на законите на разпределение на случайните величини и не може да бъдат измерени и контролирани.

Когато извадката е случайна и има достатъчно голям обем е в сила разглежданото вече важно теоретично положение за връзката между относителните честоти и вероятностите. В случая то означава, че относителните /релативните/ честоти са достатъчно точни оценки на съответните вероятности и тяхното разпределение може да се изследва въз основа на теоретичните /вероятностните/ разпределения.

Практически представителна (репрезентативна) извадка може да се образува по различни начини, чрез различни механизми.[1]

1. Един от начините за получаване на случайна представителна извадка е ***простият случаен подбор***. Той се прилага, когато генералната съвкупност е сравнително малка. За да се приложи такъв подбор, единиците на генералната съвкупност предварително се номерират в съставен за целта списък. По-нататък обикновено се използува ***таблица на случайните числа***. Като изходен пункт се избират произволна страница, ред и колона на таблицата и числата се четат хоризонтално или вертикално. Вземат се тези от тях, които не превишават най-големия номер по списъка на единиците на генералната съвкупност. В извадката попадат единиците, чиито номера съвпадат с намерените числа в таблицата.

Таблиците на случайните числа се генерират с помощта на компютър и случайността на подбора чрез тях е осигурена.

2. Друг възможен подход е ***механичният подбор*** (наричан още систематичен). И при него се използва списък с номерирани единици на генералната съвкупност. След като е определен обемът на извадката, се установява на колко единици от генералната съвкупност се пада една единица в извадката. Така се определя по списъка ***крачка на подбора***. Ако например извадката е 10-процентова, всяка десета единица по списъка ще се включва в извадката. Началото по списъка, от което се тръгва, се определя произволно, обикновено с помощта на таблицата на случайните числа.

[1] По-подробно относно качествата на отделните видове извадки, условия за използуване и пр. вж. **Цонев, В**., Основи на репрезентативното изучаване, С., 1958; **Йетс, Ф.,** Выборочный метод в переписях и обследованиях, М., 1965; **Кокрен, У**., Методы выборочного исследования, М.,1976.

Задължително изискване на този начин е списъкът да е съставен съвършено случайно. Ако подреждането на единиците на генералната съвкупност в него е преднамерено и се получава някакво циклично изменение в значенията на интересуващите ни признаци, извадката няма да бъде случайна и ще съдържа систематична грешка.

3. В много случаи е необходимо да се формира извадка, която да е представителна не само за цялата генерална съвкупност, а и за обособени нейни части, подсъвкупности. Тогава подборът се нарича ***районiран*** (или още стратифициран, разслоен). Когато се избират единиците от отделната подсъвкупност, броят им може да бъде пропорционален на частта (дела) на всяка подсъвкупност в генералната съвкупност (пропорционален подбор). Когато обаче ни интересува някакъв особено важен вариационен признак, добре е отделните подсъвкупности на генералната съвкупност да се представят пропорционално на дисперсиите или стандартните отклонения, ако те са известни.

4. Когато генералните съвкупности са много големи и обикновено са разположени върху обширни територии, често се прилага ***гнездови подбор***, наричан още ***сериен*** или ***степенен***. При този подбор образуването на извадките протича на две или повече степени. На първата степен се прави случайна извадка от по-големи съставни единици, наречени гнезда. На втората степен от гнездата се избират пак случайно статистическите единици, които се включват в извадката. Ако е необходимо, подборът може да протече на три и повече степени.

Различните видове извадки пораждат някои особености при по-нататъшната работа по изчисляването на стохастичните грешки и статистическите оценки.

Какъвто и да е механизмът за формирането на извадката, тя може да се прави по ***схема с връщане (възвратен подбор)***, или по ***схема без връщане (безвъзвратен подбор)***. В първия случай всяка единица, попаднала в извадката, отново се връща в генералната съвкупност. Във втория случай попадналите в извадката единици не се връщат и не участват по-нататък при излъчването на следващите.

При възвратния подбор се осигурява спазване на изискването всички единици на генералната съвкупност да имат еднаква вероятност да попаднат в извадката. При безвъзвратен подбор, след включване на всяка единица в извадката, вероятността за следващите се променя и то чувствително, ако генералната съвкупност не е много голяма. Това не значи, че извадките трябва да се правят само чрез възвратен подбор. Но когато подборът е безвъзвратен, трябва това да намери отражение по съответен начин във формулите за изчисляването на стандартните грешки при статистическото оценяване.

5.3. Точкови оценки

Когато по характеристиките на извадките се правят оценки на неизвестни параметри на генералните съвкупности, тези оценки се изразяват в конкретни числа. Това може да бъде например средна аритметична, дисперсия и др. Такива оценки се наричат ***точкови***. Положението им на абсцисната ос при графично представяне се определя с точки.

Най-общо постановката е следната. Генералната съвкупност има параметър θ_o (тхета), който не е известен. Необходимо е по характеристика на извадката да се направи оценка $\hat{\theta}$, която да се доближава достатъчно плътно до параметъра θ_o. Ако например трябва да се установи какво е средното потребление на месо на лице от населението чрез оценка, получена от извадка, трябва по данните от извадката да намерим такава средна, която да възпроизвежда най-добре (която е най-добра оценка) неизвестното средно потребление на цялото население.

В други случаи интересът е насочен към интервалите, в които се намират параметрите на генералните съвкупности. Такива оценки се наричат ***интервални***. При графично представяне те се изобразяват като интервали на абсцисната ос, които съдържат множество точки, за които се предполага (с определена вероятност), че някоя от тях е интересуващият ни параметър на генералната съвкупност. Този интервал, в който се предполага, че се намира параметърът, се нарича ***доверителен***

интервал. Но и тогава, когато се правят интервални оценки, трябва предварително да са направени точковите.

5.3.1. Свойства на точковите оценки

Оценките имат вероятностен характер. Поради това, не може да се твърди, че дадена оценка напълно съвпада с оценявания параметър. Но винаги, когато са възможни различни оценки, съвършено логичен е стремежът да се намери най-добрата. Но това е само желание, тъй като няма еднозначен критерий и механизъм да се избере най-добрата оценка. Затова е необходимо да се знае на какви общи изисквания трябва да отговаря една точкова оценка, или какви свойства да притежава. Без да се имат предвид тези свойства, не може да се прецени коя оценка е добра, дали изобщо дадена характеристика на извадката може да се приеме за оценка на интересуващия ни параметър на генералната съвкупност.

Отговорът на този въпрос трябва да се търси в закономерностите на разпределението на характеристиките на извадките. От тези закономерности са изведени необходимите ***свойства на оценките***, а те са неизместеност, ефективност, състоятелност и достатъчност.

1. ***Неизместена оценка*** е тази, която има математическо очакване, равно на оценявания параметър:

$$E\left(\hat{\theta}\right)=\theta_o \qquad (5.1)$$

Това означава, че при достатъчно голям брой извадки, разпределението на оценките трябва да е концентрирано около θ_o. Ако оценяваният параметър е например средната аритметична на генералната съвкупност, при голям брой извадки разпределението на средните на извадките ще бъде съсредоточено около $\bar{x}_0$. Съгласно централната пределна теорема $E(\bar{x})=\bar{x}_o$. Следователно средната аритметична на извадката е неизместена оценка на средната на генералната съвкупност.

Ако се приеме друга характеристика на извадката като оценка (θ'), която примерно е по-голяма от $\bar{x}$ с константа c, математическото й очакване ще бъде

(5.2) $$E(\theta') = E(\bar{x} + c) = E(\bar{x}) + E(c) = \bar{x}_0 + c\,.$$

Очевидно е, че в такъв случай центърът на разпределението не е в точка $\bar{x}_0$, а е ***изместен*** в точка $\bar{x}_0 + c$. Затова в този случай θ' е изместена оценка. Тя е положително изместена, тъй като $E(\theta') > \bar{x}_0$. Ако $E(\theta') < \bar{x}_0$, оценката е отрицателно (ляво) изместена оценка.

Установено е например, че дисперсията на извадката (σ^2) е изместена оценка на дисперсията в генералната съвкупност (σ_0^2) тъй като

(5.3) $$E(\sigma^2) = \sigma_0^2 \frac{n-1}{n}.$$

Щом изместеността се изразява с $\frac{n-1}{n}$, за да се получи неизместена оценка, трябва дисперсията на извадката да се умножи по $\frac{n}{n-1}$, т.е.

(5.4) $$\hat{\sigma}^2 = \sigma^2 \frac{n}{n-1}.$$

И тъй като дисперсията на извадката е $\sigma^2 = \frac{\sum (x - \bar{x})^2}{n}$, формула 5.4 ще приеме вида:

(5.5) $$\hat{\sigma}^2 = \frac{\sum (x - \bar{x})^2}{n} . \frac{n}{n-1} = \frac{\sum (x - \bar{x})^2}{n-1}.$$

Неизместеността е важно качество, защото неизместената оценка не съдържа систематична грешка, докато изместената недооценява или надценява параметъра на генералната съвкупност. Тъй като неизместеността не е единственият критерий за избора на оценката, не може да се каже, че щом тя не е изместена е най-добрата. Ето защо е необходимо да се преценява и съобразно другите критерии.

2. ***Ефективността*** на оценките зависи от техните дисперсии. Една оценка е толкова по-ефективна, колкото дисперсията ѝ е по-малка. И това е напълно логично. Ако една оценка има по-малка дисперсия от

друга възможна, може да се очаква, че възпроизвежда по-точно оценявания параметър. Разбира се, това е вярно, ако оценката е неизместена.

Ако например се вземат като центрове на разпределението средната аритметична и медианата на извадка, то и двете са неизместени оценки, но дисперсията на средната аритметична е

$$\sigma^2_{\bar{x}} = \frac{\sigma_0^2}{n}, \qquad (5.6)$$

а на медианата -

$$\sigma^2_{Me} = \frac{\sigma_0^2}{n}.\frac{\pi}{2} \ . \qquad (\pi=3{,}14) \qquad (5.7)$$

Очевидно е, че дисперсията на медианата е 1,57 пъти ($\frac{3{,}14}{2}$) по-голяма от дисперсията на средната аритметична. Затова медианата не е ефективна оценка на средната на генералната съвкупност.

Следователно ако са дадени две възможни оценки на параметъра θ_0 и са известни техните дисперсии, може да се прецени коя от тях е по-ефективна. Практически обаче възникват трудности, защото обикновено не се разполага с техните дисперсии. Известни възможности разкриват специално разработени формули за изчисляване на минималната граница на дисперсията, известна като неравенство на ***Крамер - Рао***.[1]

3. ***Състоятелността*** (наричана още сходимост или плътност) на оценките се определя според това, как те реагират при увеличаване обема на извадките. Ако при увеличаване обема на извадката оценката се стреми към оценявания параметър, тя е състоятелна. Може да се каже, че много точкови оценки притежават това свойство.

4. ***Достатъчността*** (наричана още изчерпателност) обикновено се определя така: оценката е достатъчна (изчерпателна), ако при нейното изчисляване се изчерпва цялата информация, която се съдържа в извадката. Средната аритметична например притежава това свойство,

[1] Вж. **Кендал, М**. и **А. Стюарт**, Статистические выводы и связи, М., 1973, с.23 и с.1.

защото няма друга оценка, която би могла да включи повече информация за средната на генералната съвкупност.

Има разработени методи за по-детайлно изследване на изчерпателността на оценките при различни ситуации.[1]

5.3.2. Методи за точково оценяване

Намирането на точкови оценки, като се имат предвид изложените свойства, е възможно по различни методи.

1. ***Метод на аналогията*** (метод на моментите). Чрез този метод обикновено се намират оценки на параметри, които по характера си са моментни от някакъв порядък - средна аритметична, дисперсия и др.

В основата на този метод стои принципът на аналогията: моментите в генералната съвкупност се оценяват чрез аналогичните моменти на извадката. Например за оценка на средната аритметична, дисперсията и др. служат средната аритметична, дисперсията и др. на извадката.

Тъй като много параметри, за които се търсят оценки, имат формата на моменти на разпределението, методът намира широко приложение.

2. ***Метод на максималното правдоподобие***. Той се основава на принципа, оценката на неизвестния параметър на генералната съвкупност да се намери по начин, при който е максимално правдоподобно предположението, че възпроизвежда параметъра на генералната съвкупност. За целта се търси максимум на специална функция, която се нарича ***функция на правдоподобието***.

Приложението на метода е свързано с конкретния вид на разпределението на генералната съвкупност, т.е. трябва да е известна функцията на разпределението. Това ограничава в голяма степен възможностите за практическото му използване.

[1] Вж. **Кендал, М**. и **А. Стюарт**, Цит. съч., с.40 и с.1.

3. ***Метод на най-малките квадрати***. Това е един от най-широко прилаганите методи. Може успешно да се използува за получаване на неизместени и ефективни оценки, когато те са линейни функции на наблюдаваните стойности на случайната величина. Ако например между две случайни величини съществува линейна зависимост, тя може да се изследва чрез линеен модел, тъй като се изразява с функцията

$$Y = \alpha + \beta x + \varepsilon , \qquad (5.8)$$

където ε е конкретна грешка на оценката. Оценките на самите параметри на модела α и β се намират по метода на най-малките квадрати, който се основава на изискването, $\sum \left(Y - \hat{Y}\right)^2 = minimum$.

Методът на най-малките квадрати също не е универсален. Когато обаче са осигурени посочените условия (линейност на функцията и някои други), той е предпочитан, защото е много удобен. При нормално разпределени съвкупности методът на най-малките квадрати води до резултати, които се получават и по метода на максималното правдоподобие.

Методът на най-малките квадрати се прилага широко в областта на регресионния анализ и за установяване на трайната тенденция (тренда) на развитие при анализа на динамичните редове (вж. гл. 8 и 10).

При избора на метод за получаване на оценки, при всеки конкретен случай се има предвид характерът на оценявания параметър, обстоятелството, че неговата оценка е случайна величина с някакво разпределение и необходимостта да се осигурят желаните свойства на оценките.

При избора на метода имат значение и такива съображения, които не се включват в разгледаните свойства на оценките и в конструкцията на методите. Те са чисто практически и се отнасят до преценка на разходите на труд, средства и време. Съвсем не е без значение на каква "цена" може да се получи желаната оценка и какви потребности тя ще задоволява. Както във всички други области на статистическия анализ, не е достатъчно само добре да е овладян методологичният апарат. Необходимо е специалистът, оперирайщ с този апарат, добре да познава

явленията, които ще изследва, целите, за които ще служат резултатите, реалните възможности да се постигнат желаните резултати и очакваните последици от едно или друго заключение.

5.4. Интервални оценки

Както беше посочено, точковата оценка фиксира определено число като най-приемлив израз на интересуващия ни параметър. Казано по друг начин, от множество възможни оценки избираме една, приемайки (въз основа на определени критерии), че тя е най-приемлива. В редица случаи обаче се предпочита не конкретно число (точкова оценка) като вероятна стойност на параметъра, а ***определен интервал***, за който може да се твърди (с определена вероятност), че съдържа параметъра. Такава оценка се нарича ***интервална***. Тя също има вероятностен характер и съдържа точковата, но и вероятната грешка.

Заключението, което се прави въз основа на интервалната оценка съдържа някаква неопределеност. Но съществува възможност да се контролира степента на увереност относно достоверността на заключенията.

5.4.1. Максимална грешка и доверителен интервал

Интервалът, за който се твърди, че съдържа оценявания параметър на генералната съвкупност, се нарича ***доверителен интервал***, или интервал на доверителност. Вероятността, с която се гарантира заключението се нарича ***доверителна*** или ***гаранционна*** вероятност.

За да се разбере смисъла и начина за съставяне на доверителния интервал (интервалната оценка) и свързаната с него максимална грешка, трябва да се имат предвид някои разгледани вече основни положения, свързани с теоретичните разпределения.

Беше установено, че ако от дадена генерална съвкупност със средна аритметична $\bar{x}_0$ и дисперсия σ_0^2 се правят случайни извадки с обем *n*, то ако извадките се увеличават, разпределението на средните на извадките ще се стреми към нормалното и тяхната обща средна ще бъде

равна на средната на генералната съвкупност, дисперсията им ще бъде равна на $\frac{\sigma_0^2}{n}$. Следователно стандартното отклонение е:

$$\sigma_{\bar{x}} = \frac{\sigma_0}{\sqrt{n}}. \qquad (5.9)$$

То се нарича ***стандартна грешка на оценката***. Защо $\sigma_{\bar{x}}$ измерва грешката?

Ако се направи една извадка, нейната средна обикновено ще се отклонява от средната на генералната съвкупност. Това отклонение $\varepsilon_1 = (\bar{x}_1 - \bar{x}_0)$ е конкретната стохастична грешка на тази единствена извадка. Тя показва каква (в абсолютни единици) ще бъде грешката, ако се приеме $\bar{x}$ за оценка на $\bar{x}_0$. При следващата извадка също ще се получи някаква грешка- $\varepsilon_2 = (\bar{x}_2 - \bar{x}_0)$ и т.н. Ако са направени *k* извадки, ще се получат *k* средни и *k* конкретни грешки. От тези *k* на брой грешки може да се намери средна грешка. Тя следва да се изчисли като средноквадратично (стандартно) отклонение:

$$\bar{\varepsilon} = \sigma_{\bar{x}} = \sqrt{\frac{\sum (\bar{x}_i - \bar{x}_0)^2}{k}}. \qquad (5.10)$$

Това обаче е една абстрактна, теоретична постановка. Практически не е възможно да се постъпи по този начин, защото средната на генералната съвкупност не е известна и като правило се прави само една извадка. Известно е обаче от формула 5.9, че стандартното отклонение на средните на извадките е $\sqrt{n}$ пъти по-малка от стандартното отклонение в генералната съвкупност (σ_0), или то е $\sigma_{\bar{x}} = \frac{\sigma_0}{\sqrt{n}}$. Тази формула съдържа стандартното отклонение на генералната съвкупност, което не е известно. То обаче може да се замести с неговата оценка, изчислена по данните от извадката - $\hat{\sigma}$. Ако се приеме по-нататък стандартната грешка (стандартното отклонение), изчислена от една единствена извадка, да се означава с μ, ще се получи:

$$(5.11) \qquad \mu_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}}.$$

За да се състави доверителния интервал, необходимо е стандартната грешка да се свърже с определена вероятност (доверителна вероятност) и по този начин да се получи ***максимално допустимата грешка*** в условията на една извадка.

Известно е, че средните на извадките са случайни величини и че с увеличаване на *n* те се стремят към нормално разпределение. Затова могат да се използуват някои известни вече свойства на нормалното разпределение.

Беше установено, че в границите $\pm z\sigma$ се обхваща определен дял от площта под нормалната крива (който отговаря на определена вероятност). Тук в ролята на σ влиза μ, а зад *z*, наричано в случая доверителен коефициент, стои определена вероятност. При това положение ***максималната грешка*** ($\Delta_{\bar{x}}$) ще бъде:

$$(5.12) \qquad \Delta_{\bar{x}} = z\mu_{\bar{x}}.$$

Ако от средната на извадката се извади и към тази средна се прибави максималната грешка, получава се доверителния интервал, или интервалната оценка на средната на генералната съвкупност ($\bar{x}_0$):

$$(5.13) \qquad \left(\bar{x} - \Delta_{\bar{x}}\right) \le \bar{x}_0 \le \left(\bar{x} + \Delta_{\bar{x}}\right).$$

Това означава, че средната на генералната съвкупност не е по-малка от $\left(\bar{x} - \Delta_{\bar{x}}\right)$ и не е по-голяма от $\left(\bar{x} + \Delta_{\bar{x}}\right)$ и това твърдение се прави с определена вероятност, изразена чрез коефициента *z*. Той се намира в таблицата за площите под нормалната крива (приложение 4).

Обикновено заключенията при интервалното оценяване се правят с вероятност 0,95 (z = 1,96) и вероятност 0,99 (z = 2,58).

Максималната грешка, изчислена по посочената формула, е абсолютна величина, изразена в съответна мярка, в каквато са изразени оценяваните параметри. В някои случаи е целесъобразно да се приведе в относителна величина. При оценката на средната аритметична например относителният размер на максималната грешка ще бъде:

(5.14) $$\Delta_{\bar{x}(\%)} = \frac{\Delta_{\bar{x}}}{\bar{x}}.100\,.$$

Посочената формула (5.11) на стандартната грешка е валидна в този й вид при извадка, формирана с възвратен подбор. Тогава генералната съвкупност е хипотетична, безкрайно голяма, неизчерпаема. Когато подборът е безвъзвратен, както беше посочено, вероятностите за единиците на генералната съвкупност да попаднат в извадката се променят след изваждането на всяка единица. Ето защо е необходимо да се направи корекция на $\mu_{\bar{x}}$, наричана обикновено ***корекция за крайност на генералната съвкупност***. Доказано е, че тази корекция може да стане с коефициента $\sqrt{\frac{N-n}{N-1}}$, където N е обемът на генералната съвкупност, а n - обемът на извадката. Тъй като

$$\sqrt{\frac{N-n}{N-1}} \approx \sqrt{1-\frac{n}{N}}\,, \quad \text{обикновено се използва } \sqrt{1-\frac{n}{N}}\,.$$

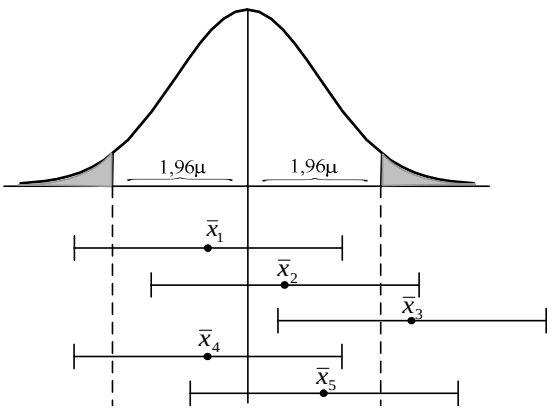
С тази корекция стандартната грешка при безвъзвратен подбор ще има следната формула:

(5.15) $$\mu_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}}\sqrt{1-\frac{n}{N}}\,.$$

Очевидно е, че когато n е сравнително малко, а N е много голямо, $\sqrt{1-\frac{n}{N}}$ ще клони към единица и следователно ще е без значение за стойността на $\mu_{\bar{x}}$. Затова е възприето, ако $\frac{n}{N} < 0{,}05$, т.е. ако извадката (n) е по-малка от 5 % от генералната съвкупност, да не се прави корекция с множителя $\sqrt{1-\frac{n}{N}}$.

Беше вече посочено, че заключението, което следва от интервалната оценка, винаги е свързано с определена вероятност - 0,95, 0,99 или друга. Ако теоретично си представим, че се правят много

случайни извадки със съответни средни - $\bar{x}_i$, могат да се съставят толкова доверителни интервали, колкото са извадките. Доверителните интервали имат някакво разпределение. Ако се знае средната на генералната съвкупност, може да се окаже, че тя се съдържа в някои интервали, а в други не се съдържа. Това е показано примерно на фиг. 5.1. Вижда се, че интервалите, построени по $\bar{x}_1, \bar{x}_2, \bar{x}_4, \bar{x}_5$ съдържат оценявания параметър ($\bar{x}_0$), но интервалът, построен по $\bar{x}_3$ (една извадка) не го съдържа.



Фиг. 5.1.

Доверителен интервал (интервална оценка), построен при доверителна вероятност 0,95, означава, че от общо 100 интервала, 95 ще съдържат оценявания параметър и само 5 може да не го съдържат. Ако доверителният интервал е определен при вероятност 0,99, означава, че от 100 доверителни интервала е възможно един да не съдържа параметъра.

Трябва да се има предвид, че когато се увеличава доверителната вероятност, доверителният интервал се разширява. Има следователно "противоречие" между желанието вероятността да е по-голяма и доверителният интервал да е по-тесен. Няма и не може да има общовалидно правило, или формула за решаване на това "противоречие". Въпросът е в ръцете на специалиста, който прави конкретното изследване. Той трябва да го реши съобразно характера и целите на изследването, като съзнава добре последиците от своето заключение.

Изчисляването на максималната грешка и на доверителния интервал при интервална оценка на средна величина е илюстрирано с пример в точка 10.8.

Изложеният подход за намиране на максималната грешка и на доверителния интервал важи по принцип за различни параметри. Доверителният интервал при интервално оценяване на какъвто и да е параметър (θ_0) е

(5.16) $$\hat{\theta} \pm \Delta_\theta$$

или $$\left(\hat{\theta} - z\,\mu_\theta\right) \le \theta_0 \le \left(\hat{\theta} + z\,\mu_\theta\right),$$

където: $\Delta_\theta = z\,\mu_\theta$, а $\mu_\theta = \dfrac{\hat{\sigma}_\theta}{\sqrt{n}}$ при възвратен подбор, съответно

$\mu_\theta = \dfrac{\hat{\sigma}_\theta}{\sqrt{n}}\sqrt{1-\dfrac{n}{N}}$ при безвъзвратен подбор.

Единствената разлика е в изчисляването на $\hat{\sigma}_\theta$ за различните параметри.

Ако например е необходимо да се определи доверителен интервал на относителен дял на единиците на съвкупността, които имат дадено качество, ще се приложи познатата формула за изчисляване на средно квадратично отклонение при алтернативни (дихотомни) признаци:

$$\sigma_p = \sqrt{pq} = \sqrt{p(1-p)},$$

където:

p е относителният дял на единиците в извадката, които имат интересуващото ни качество;

q - относителният дял на единиците, които нямат това качество.

Тъй като σ_p за генералната съвкупност не е известно, за негова оценка се приема σ_p на извадката.[1] Следователно при възвратен подбор стандартната грешка ще се изчисли по формулата:

[1] Трябва да се има предвид, че p в извадката е неизместена оценка на p в генералната съвкупност.

$$\mu_p = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}} \quad , \tag{5.17}$$

а при безвъзвратен подбор -

$$\mu_p = \sqrt{\frac{p(1-p)}{n}\left(1-\frac{n}{N}\right)} \quad . \tag{5.18}$$

Максималната грешка е

$$\Delta_p = z\,\mu_p \, . \tag{5.19}$$

Следователно доверителният интервал ще бъде

$$p \pm \Delta_p \tag{5.20}$$

или:

$$\left(p - z\mu_p\right) \le p_0 \le \left(p + z\mu_p\right). \tag{5.21}$$

Изчисляването на максималната грешка и на интервалната оценка на относителен дял може да се илюстрира със следващия ***пример***.

Контролните органи по качеството, стандартизацията и метрологията са направили проверка в едно захародобивно предприятие. От намиращите се в склада 10000 (N) пакета със захар е направена случайна извадка с безвъзвратен подбор от 100 (n) пакета и при измерването им е установено, че 8 от тях, или 8% (p = 0,08) имат по-малко тегло от обявеното на опаковката.

Необходимо е да се направи интервална оценка на относителния дял на пакетите с по-малко тегло във всичките 10000 пакета (генералната съвкупност). Поради важността на заключението и с оглед на бъдещите практически мерки, интервалната оценка ще се направи с доверителна вероятност 0,99.

Знае се, че относителният дял на единиците в извадката (n), които имат или нямат дадено качество, е неизместена точкова оценка на относителния дял в генералната съвкупност (p_0). Знае се също, че стандартното отклонение при алтернативни (бинарни, дихотомни)

признаци е $\sigma_p = \sqrt{pq} = \sqrt{p(1-p)}$. Следователно в примера то е $\sigma_p = \sqrt{0{,}08.0.92} = 0{,}271$.

Стандартната грешка е

$$\mu_p = \sqrt{\frac{p(1-p)}{n}\left(1-\frac{n}{N}\right)} = \sqrt{\frac{0{,}08.0{,}92}{100}\left(1-\frac{100}{10000}\right)} = 0{,}027 .$$

Максималната грешка при доверителна вероятност 0,99 е:

$\Delta_p = 2{,}58.0{,}027 = 0{,}07$ или 7,0 %.

Следователно доверителният интервал е

$(8-7) \le p_0 \le (8+7)$ или $1\,\% \le p_0 \le 15\,\%$.

Следва заключението: с вероятност 0,99 (с 1 % риск за грешка) може да се твърди, че относителният дял на пакетите с тегло под обявеното във всичките 10000 пакета не е по-малък от 1 % и не е по-голям от 15 %. Получава се много широк интервал. Той поражда съмнение относно обема на извадката, т.е. дали тя е била предварително добре обоснована.

Описаният начин за намиране на доверителния интервал за относителен дял се основава върху нормалното разпределение. При достатъчно големи извадки, ако p не е много малко, така че да се изпълни условието np и $n(1 - p) > 5$, той дава практически добри резултати. Има разработени формули, които се основават на по-строги изисквания и се прилагат при различни условия.[1]

5.5. Оценки, основани на обединени извадки

Ако от една съвкупност са направени две или повече извадки, оценките, изчислени от тях, биха били различни. Те обаче могат да се обединят и да се получи една обща оценка.

Ако например се оценява средна величина на генерална съвкупност и по две извадки (n_1 и n_2) са получени средните $\bar{x}_1$ и $\bar{x}_2$ като

[1] Вж. **Закс, Л.**, Статистиическое оценивание. М., 1976, с.304 и сл.

точкови оценки, обединената оценка следва да се изчисли като средна от средните на двете извадки, претеглени с обемите на извадките:

$$\bar{x} = \frac{\bar{x}_1 n_1 + \bar{x}_2 n_2}{n_1 + n_2}. \qquad (5.22)$$

Ако се оценява дисперсията по две извадки, получени са двете оценки $\hat{\sigma}_1^2$ и $\hat{\sigma}_2^2$ със степени на свобода $(n_1 - 1)$ и $(n_2 - 1)$. Тези степени на свобода произлизат от познатата формула $\hat{\sigma}^2 = \frac{\sum (x - \bar{x})^2}{n-1}$, от която следва, че $\hat{\sigma}_1^2 = \frac{\sum (x_i - \bar{x}_1)^2}{n_1 - 1}$ и $\hat{\sigma}_2^2 = \frac{\sum (x_i - \bar{x}_2)^2}{n_2 - 1}$. Обединената оценка също може да се изчисли като средна от двете оценки с тегла $(n_1 - 1)$ и $(n_2 - 1)$.

$$\hat{\sigma}^2 = \frac{\hat{\sigma}_1^2 (n_1 - 1) + \hat{\sigma}_2^2 (n_2 - 1)}{(n_1 - 1) + (n_2 - 1)} = \frac{\hat{\sigma}_1^2 (n_1 - 1) + \hat{\sigma}_2^2 (n_2 - 1)}{n_1 + n_2 - 2}. \qquad (5.23)$$

Известно е обаче, че $\hat{\sigma}^2 = \sigma^2 \frac{n}{n-1}$, следователно $\hat{\sigma}_1^2 = \sigma_1^2 \frac{n_1}{n_1 - 1}$ и $\hat{\sigma}_2^2 = \sigma_2^2 \frac{n_2}{n_2 - 1}$ Затова формула 5.23 ще се развие по-нататък:

$$\hat{\sigma}^2 = \frac{\sigma_1^2 \frac{n_1}{n_1 - 1}(n_1 - 1) + \sigma_2^2 \frac{n_2}{n_2 - 1}(n_2 - 1)}{n_1 + n_2 - 2}.$$

След възможните съкращения в числителя:

$$\hat{\sigma}^2 = \frac{\sigma_1^2 n_1 + \sigma_2^2 n_2}{n_1 + n_2 - 2}. \qquad (5.24)$$

Следва, че стандартното отклонение като обединена оценка е:

$$\hat{\sigma} = \sqrt{\frac{\sigma_1^2 n_1 + \sigma_2^2 n_2}{n_1 + n_2 - 2}}. \qquad (5.25)$$

Степените на свобода също са обединени:

$$\phi = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2 .$$

По същия начин ще се получи обединена оценка на дисперсията и средното квадратично отклонение на относителни дялове.

Известно е, че $\sigma_p = \sqrt{pq} = \sqrt{p(1-p)}$. При две извадки обединената оценка на σ_p ще бъде:

$$\sigma_p = \sqrt{\frac{p_1 q_1 n_1 + p_2 q_2 n_2}{(n_1 - 1) + (n_2 - 1)}} = \sqrt{\frac{p_1 q_1 n_1 + p_2 q_2 n_2}{n_1 + n_2 - 2}} . \qquad (5.26)$$

Такива обединени оценки са необходими примерно при статистическата проверка на хипотези, която се разглежда в гл. 6.

5.6. Обем на извадката

В цялото предходно изложение относно точковите и интервалните оценки се приемаше, че извадките са вече направени. Практически при емпиричните изследвания трябва първо да се определи ***обемът на извадката*** и след това да се пристъпи към изследването. Но за да се разбере начина, по който се определя обема на извадката, трябва да се познават формулите за стандартната и максималната грешка.

Беше посочено, че максималната грешка на средна аритметична при възвратен подбор и при известно стандартно отклонение в генералната съвкупност е:

$$\Delta_{\bar{x}} = z\,\mu_{\bar{x}} = z\frac{\sigma_0}{\sqrt{n}} . \qquad (5.27)$$

Ако неизвестна величина е n (обемът на извадката), може да се състави формулата:

$$n = \frac{z^2\,\sigma_0^2}{\Delta_{\bar{x}}^2} . \qquad (5.28)$$

Очевидно е, че обемът на извадката зависи право пропорционално от вариацията (σ_0^2) в генералната съвкупност и от вероятността (която

стои зад z), с която ще се прави заключението и обратно пропорционално от максимално допустимата грешка $\Delta_{\bar{x}}^2$.

При безвъзвратен подбор подходът е същият, но изходната формула на максималната грешка е

$$\Delta_{\bar{x}} = \frac{z\sigma_0}{\sqrt{n}}\sqrt{1-\frac{n}{N}},$$

от която се извежда

$$n = \frac{z^2\sigma_0^2 N}{\Delta_{\bar{x}}^2 N + z^2\sigma_0^2}. \qquad (5.29)$$

Очевидно е, че за да се изчисли обемът на извадката, трябва предварително да се зададе по субективна преценка максималната грешка и желаната доверителна вероятност. Ако се предвиди по-малка максимална грешка, извадката ще бъде по-голяма. По-голяма ще бъде тя и при по-голяма вероятност и по-голяма дисперсия. Тук отново решението относно максималната грешка и доверителната вероятност взема специалистът, чийто професионализъм е поставен на изпитание.

Практически проблем възниква по отношение на дисперсията, която във формулите се отнася за генералната съвкупност. Възможни са две решения. Първо, тя може да се вземе от минали наблюдения на същия обект. Второ, ако такова наблюдение не е правено, следва да се направи предварително произволна микроизвадка, за да се получи информация за изчисляване на σ^2 като приблизителна оценка на σ_0^2.

Изложеният начин за намиране на обема на извадката е приложим по принцип при оценка на различни параметри. Разбира се, формулите се променят поради различните конструкции на максималната грешка. Трябва при това да се имат предвид и условията за минималния обем на извадката, при който могат да се използуват съответните теоретични разпределения.

Тук ще посочим допълнително само намирането на обема на извадката при оценка на относителен дял. То става също въз основа на формулата на максималната грешка. При възвратен подбор тя е

$$\Delta_p = z\sqrt{\frac{p(1-p)}{n}}.$$

От нея следва, че

$$n = \frac{z^2 p(1-p)}{\Delta_p^2}. \quad (5.30)$$

При безвъзвратен подбор

$$n = \frac{z^2 p(1-p)N}{\Delta_p^2 N + z^2 p(1-p)}. \quad (5.31)$$

Тази формула може да се запише и така:

$$n = \frac{z^2 N}{z^2 + \frac{\Delta_p^2 N}{p(1-p)}}. \quad (5.32)$$

Въз основа на някои преобразувания на формула 5.31 са съставени таблици, по които може да се намери необходимият обем на извадката при предварително зададена точност.[1] Те създават практически удобства и могат да се използуват при всякакви стойности на p.

Описаните формули и процедури за намирането на доверителния интервал и обема на извадката се отнасят за прост случаен подбор. Принципно те важат и за останалите модели на извадки, но в едно или друго отношение се модифицират. Затова според вида на извадката трябва да се избират съответните работни формули.[2]

Освен това, във всички случаи се предполагаше, че извадките са достатъчно големи - условие, при което е възможно да се прилагат методологичните положения, важащи при нормално разпределение. При малки извадки се налагат някои изменения. За някои параметри бяха посочени минималните граници, под които извадката е малка.

[1] Вж. **Yamane, T**., Statistics, An Introductory Analisis, Third Edition. New York,1973,p.728
[2] Вж. **Цонев, В**., Цит. съч.

Подробното разглеждане на теорията на малките извадки излиза извън рамките на тази книга.

5.7. Практикум

5.7.1. Въпроси за самопроверка

1. Какво се разбира под статистически заключения?
2. Какво означава твърдението, че статистическите заключения имат вероятностен характер?
3. При какви условия една извадка е представителна (репрезентативна)?
4. Какво представлява свойството неизместеност на точковите оценки?
5. Какво се разбира под интервална оценка?
6. Какво представлява максималната грешка?
7. Какво значи доверителна вероятност?
8. Как се изчислява стандартната грешка при оценка на средна величина, ако извадката е получена чрез безвъзвратен подбор?
9. Как се изчислява стандартната грешка на относителен дял при извадка, получена чрез възвратен подбор?
10. Как се определя обемът на извадката за оценяване на средна аритметична величина?

5.7.2. Задачи за упражнение

Задача 1. В една фирма 360 работници произвеждат едно изделие. За да се установи средната часова производителност на труда е направена с безвъзвратен подбор случайна извадка от 36 работници (10%-ова извадка). През времето на наблюдението средната часова производителност на работниците в извадката е била 324 броя, при оценка на стандартното отклонение 24 бр. Необходимо е да се изчисли интервална оценка на средно часовата производителност на труда на всичките 360 работници (генералната съвкупност) с доверителна вероятност 0,95.

Дадени са следователно $N = 360$, $n = 36$, $\bar{x} = 324$, $\hat{\sigma} = 24$, $z = 1{,}96$ (при вероятност 0,95).

Решение:

Стандартната грешка е:

$$\mu_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}}\sqrt{1-\frac{n}{N}} = \frac{24}{\sqrt{36}}\sqrt{1-\frac{36}{360}} = 3{,}8$$

Максималната грешка е:

$$\Delta_{\bar{x}} = z\mu_{\bar{x}} = 1{,}96.3{,}8 = 7 \text{ бр. (кръгло)}$$

Като от средната на извадката ($\bar{x} = 324$) се извади и към нея се прибави получената максимална грешка (7 бр.), ще се получи доверителния интервал (интервалната оценка):

$$317 \le \bar{x}_0 \le 331.$$

Следва заключението: с вероятност 0,95 може да се твърди, че действителната средна часова производителност на труда на всичките 360 работници не е по-малко от 317 бр. и не е по-голяма от 331 бр. Рискът тя да не се намира в тези граници е само 5 %.

Задача 2. Фирма, която продава домашни бойлери се интересува от оценката на потребителите за качеството им. За целта има намерение да направи извадкова анкета с въпрос към потребителите: “Доволни ли сте от уреда и при нужда бихте ли си купили същата марка?” По този начин иска да установи каква част от потребителите не одобряват уреда. Извадката ще бъде направена по схема без връщане чрез прост случаен подбор по номерата на фактурите.

Колко потребители трябва да се анкетират (какъв да бъде обемът на извадката) за да се направи интервална оценка на относителния дял (p) на недоволните при максимална грешка ± 2 процентни пункта и доверителна вероятност 0,95. Общо от фирмата са били продадени 2500 бойлера (N). При минало изследване за подобен тип бойлери е било установено, че недоволните са били 1,5 % (p = 0,015).

Решение:

Трябва да се приложи формулата за определяне на обема на извадката при безвъзвратен подбор:

$$n = \frac{z^2 p(1-p)N}{\Delta_p^2 N + z^2 p(1-p)} = \frac{1{,}96^2.0{,}015.0{,}985.2500}{0{,}02^2.2500 + 1{,}96^2.0{,}015.0{,}985} = 134$$

Може следователно да се приеме, че при случайна извадка от 134 анкетирани лица ще бъдат удовлетворени поставените цели в условието.

Задача 3. В голям универсален магазин са правени наблюдения, за да се установи какво е разпределението на посетителите по времето, което прекарват в магазина. Установено е, че то е много близко до нормалното. При последното наблюдение са анкетирани случайно избрани 50 посетители. Изчислено е, че те средно са престояли в магазина 25 минути ($\bar{x}$) при дисперсия $\sigma^2 = 64$ минути.

Необходимо е да се направи интервална оценка за средното време на престоя в магазина на всички негови посетители при доверителна вероятност 0,95. Трябва да се има предвид, че генералната съвкупност в случая е безкрайно голяма.

Отговор: 22,8 мин. $\leq \bar{x}_0 \leq$ 27,2 мин.

Задача 4. Предстои да се организира извадково изследване на задграничните пътувания на български граждани. Един от признаците е времето на престоя в други страни. Наблюдението ще се проведе по календарен план в продължение на 3 месеца. Установено е, че през определения период зад граница се извършват общо около 1 млн. пътувания. От минали наблюдения е известно, че дисперсията на времето на престоя в други страни $\sigma^2 = 260$ (данните са примерни).

Колко голяма трябва да бъде извадката (броят на анкетираните лица), за да се направи оценка на продължителността на пребиваването в други страни за всички пътуващи (генерална съвкупност) при максимална грешка 2 дни и доверителна вероятност 0,95. Извадката ще се направи по схема без връщане.

Отговор: n = 250

5.7.3. Поуката от несполуката

В теорията на извадковите статистически изследвания е станал почти класически следният случай на пълно фиаско.[1]

Редакцията на американското списание “Literary Digest” е предприела през 1936 г. извадково проучване на общественото мнение преди предстоящите президентски избори относно най-вероятният президент при състезаващите се двама кандидати – Франклин Делано Рузвелт и Алфред Ландън. След получаване и съответно обработване на получените анкетни карти, списанието обявява, че бъдещият президент ще се казва Ландън (57 % от гласовете). В действителност изборите печели с голямо превъзходство Франклин Делано Рузвелт (над 60 % от гласовете). Обществото е скандализирано. Какво беше се случило?

Приемайки твърде наивно, че формират случайна извадка, организаторите на анкетата са определяли (по случаен път) адресите, на които ще изпращат анкетните карти, от телефонния указател, от списъка на притежателите на автомобили и от списъка на абонатите на списанието. Това става по време, когато страната е в продължителна остра депресия, и голяма част от избирателите от по-бедните слоеве, които са симпатизирали на Рузвелт, не са били представени в извадката, защото не са имали телефони и коли, не са били и абонати на списанието. Следователно извадката е съставена при пълно нарушение на един основен принцип - да се даде еднаква възможност на всички единици на генералната съвкупност да попаднат в извадката. Въпреки големия й размер (анкети са изпратени на 10 милиона души, от които над 2 милиона души са отговорили) тя не е възпроизвела правилно съотношението на симпатизантите на двамата кандидати, т.е. тя се оказва силно ***изместена*** извадка.

Неуспехът в това проучване е бил катастрофален удар за списанието, но става сериозен и незабравим поучителен пример.

[1] **Кимбл, Г.** Как правильно пользоваться статистикой, М., 1982, с.145.

6. СТАТИСТИЧЕСКА ПРОВЕРКА НА ХИПОТЕЗИ

"Хипотезата може само да бъде проверена, но никога не може да бъде доказана."

А. Закс

Тази глава обхваща един друг аспект на статистическите заключения – възможностите и подходите за проверка дали определени предположения (хипотези) относно генералните съвкупности, които се правят въз основа на извадки, действително се потвърждават или не. Тук читателят ще научи изключително важни изисквания за коректен подход, за възможни два вида грешки, какво представляват критериите за проверка на различните хипотези и др. Тези знания специалистът, анализаторът може да използва в различни области според професионалната си ориентация – икономиката, социалните процеси, технологиите, експерименталното дело, бизнеса и др.

6.1. Същност на статистическата проверка на хипотези

Статистическата проверка на хипотези като клон на статистическите заключения е свързана със статистическите оценки, но е друг подходът към параметрите на изучаваните съвкупности и решава други познавателни задачи. Различието се състои в това, че не се търсят оценки на определени неизвестни параметри, а предварително се правят предположения (хипотези) относно параметрите и след това се проверява дали емпиричните данни, получени от случайни извадки, потвърждават или не хипотезите. Трябва при това да се има предвид, че статистическото понятие за хипотеза е свързано със случайни величини и че хипотезите могат да се проверяват.

Много примери могат да илюстрират характерът на статистическите хипотези и смисълът на проверката им.

1. Селекционери са създали нов сорт пшеница. Въз основа на проведени опити се твърди, че той е по-високодобивен в сравнение с друг. Предполага се, че това се дължи на култивираните по-добри качества на новия сорт. Но това е все още предположение (хипотеза). Трябва да се провери дали разликата между добивите не е случайна.

2. Във фирма е експериментирана нова технология за производството на даден продукт, за която се смята, че осигурява по-добро качество в сравнение с прилаганата преди експеримента. Установено е въз основа на случайни извадки, че при произведените изделия по новата технология относителният дял на нестандартните е по-малък. И в този случай разликата в относителните дялове на нестандартните изделия при двете технологии трябва да се провери.

3. При изследване на разпределението на домакинствата в Р България по размер на средния им годишен доход се твърди, че то се подчинява на закона на нормалното разпределение. Дали наистина има съгласуване между емпиричното и теоретичното нормално разпределение трябва да се провери като хипотеза.

Могат да се посочат още много примери в областта на икономиката, социалните процеси, опитното дело, медицината и др.

Заключението, което се прави при проверката на всяка хипотеза, има вероятностен характер, винаги се свързва с определена вероятност. Прилаганата методология при проверката осигурява възможност да се контролира риска за грешка и да се свежда той до приемливо равнище.

При всяка проверка се дефинират две хипотези. Едната е проверяваната, наречена ***нулева хипотеза***, означавана с H_0. Другата се противопоставя на нулевата, нарича се ***алтернативна*** и се означава с H_1. Заключението след проверката е или потвърждаване на нулевата хипотеза или нейното отхвърляне (приемане на алтернативната).

Ако например се проверява разликата между средните на две извадки, двете хипотези ще се дефинират по следния начин:

$$H_0 : \bar{x}_1 - \bar{x}_2 = 0$$

$$H_1 : \bar{x}_1 > \bar{x}_2 \ , \quad \text{или} \quad H_1 : \bar{x}_1 < \bar{x}_2 \ , \quad \text{или} \quad H_1 : \bar{x}_1 \neq \bar{x}_2 \ .$$

Когато в резултат на проверката се приеме нулевата хипотеза, това още не означава, че тя е вярна, както и обратно, ако се отхвърли, не означава, че не е вярна. Винаги има риск да се допусне грешка. Всъщност могат да се допуснат два типа грешки - грешка от първи род и от втори род.

Грешка от първи род се допуска, когато се отхвърли вярна нулева хипотеза. Вероятността за такава грешка се означава с α и затова се нарича се още ***α-грешка***.

Грешка от втори род се допуска, когато се приеме невярна нулева хипотеза. Рискът за такава грешка се означава с β и затова се нарича още ***β-грешка***.

Правилно е заключението, ако се приеме вярна или се отхвърли невярна нулева хипотеза.

Четирите положения са представени схематично в табл. 6.1.

Таблица 6.1

Вярна хипотеза / Направено заключение	Нулевата (H_o)	Алтернативната (H_1)
Приета е H_o (Отхвърлена е H_1)	Правилно заключение (вероятност: 1 - α	Грешка от втори род (вероятност: β)
Отхвърлена е H_o (Приета е H_1)	Грешка от първи род (вероятност: α)	Правилно заключение (вероятност: 1 - β)

Теорията за статистическа проверка на хипотези се зароди и оформи през 30-те години на XX век. Тя се разви по-нататък в две направления: ***класическа теория*** (на ***Нейман - Пирсън***) и ***теория на***

последователния анализ (***секвенционен анализ на А. Уолд***). При първото направление проверката на хипотези се опира на направени вече извадки и получена от тях информация. При последователния анализ обемът на извадката се определя в хода на проверката, т.е. самата извадка е функция от проверката и тя се проявява като случайна величина. Този подход се прилага сравнително по-рядко при особени условия и главно в експерименталното дело. В следващото изложение се разглежда класическата теория. По принцип обаче основните положения са общовалидни и за двете направления.

6.2. Критерий за проверка на хипотези

Проверката на всяка хипотеза става по определен ***критерий*** или тест. Може най-общо да се каже, че критерият съдържа условията, при които проверяваната хипотеза се приема или отхвърля. По-конкретно критерият съдържа определена характеристика, наречена характеристика на критерия, равнище на значимост и критична област.

Характеристиката на критерия може да бъде от различен вид в зависимост от прилагания метод за проверка и характера на проверяваната хипотеза. Тя има емпирична и теоретична (критична) стойност.

Емпиричната стойност на характеристиката се изчислява от данните, които се съдържат в извадките.

Теоретичната (критичната) стойност на характеристиката фиксира границите на влиянието на случайни фактори. Тя се определя по валидното за дадената хипотеза теоретично разпределение (по стандартна таблица за теоретичното разпределение).

Равнище на значимост (α) се нарича вероятността (риска) за грешка от първи род, т.е. да се отхвърли вярна нулева хипотеза. Специалистът, който прави проверката, задава тази вероятност. Най-често като равнище на значимост се приема $\alpha = 0{,}01$ и $\alpha = 0{,}05$.

Когато се приеме $\alpha = 0{,}01$, това означава, че се допуска 1 %-ов риск да се отхвърли вярна нулева хипотеза. При $\alpha = 0{,}05$ рискът е 5 %-ов.

Логично е при проверка на конкретни хипотези да се проявява стремеж към по-малък риск за грешка от първи род (α). Ако обаче се

намали този риск, увеличава се риска от β-грешка. Отново специалистът разрешава това "противоречие" в желанията. Той преценява последиците от погрешно приемане или погрешно отхвърляне на проверяваната хипотеза.

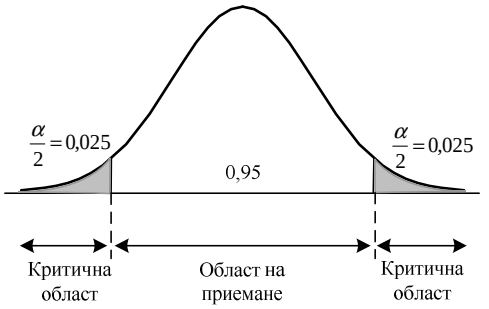
Вероятността да се отхвърли невярна нулева хипотеза се нарича ***мощност на критерия*** и се означава с 1 - β. Казва се, че един критерий е по-мощен, когато е по-голяма вероятността да се отхвърли нулевата хипотеза, ако тя не е вярна.

Критична област се нарича областта в дадено разпределение, която се намира отвъд теоретичната (критичната) стойност на характеристиката и в която нулевата хипотеза се отхвърля. Това важно положение се нуждае от по-широко изясняване.

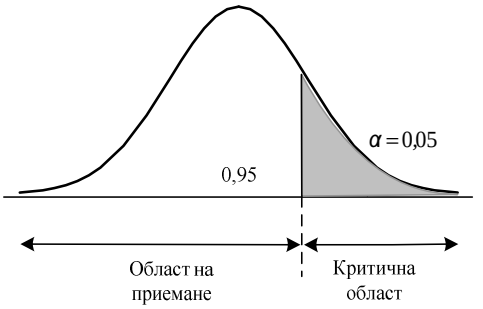
Нека да се проверява разлика между средни аритметични на две извадки. Емпиричната стойност на прилаганата в случая характеристика може да бъде различна. Тя се подчинява на закона на определено разпределение. С теоретичната стойност на характеристиката се определят границите на допустимите случайни колебания. Ако се вземе площта под кривата на нормалното разпределение, тези граници определят ***областта на приемането***. Това означава, че ако емпиричната стойност на характеристиката се намира в тези граници нулевата хипотеза се приема. Останалата област, която се намира извън областта на приемането е критичната област. Ако емпиричната стойност на характеристиката е попаднала в нея, нулевата хипотеза се отхвърля.

Критичната област може да бъде ***двустранна*** или ***едностранна***.

Ако например се проверява разликата между средните на две извадки и алтернативната гласи, че $H_1 : \bar{x}_1 \neq \bar{x}_2$, това значи, че критичната област е двустранна (фиг. 6.1). Ако обаче алтернативната хипотеза е дефинирана $H_1 : \bar{x}_1 > \bar{x}_2$ или $H_1 : \bar{x}_1 < \bar{x}_2$, критичната област е едностранна – дясностранна или лявостранна (фиг. 6.2).



Фиг. 6.1



Фиг. 6.2

Дали чрез алтернативната хипотеза ще се зададе двустранна или едностранна критична област зависи от това дали има информация за възможността разликата да бъде само в едната или в двете посоки, както и от значението, което има за практическите изводи проверката при двустранна или едностранна критична област. Във всеки случай този въпрос се решава с начина, по който се дефинира алтернативната хипотеза.

Когато критичната област е едностранна, равнището на значимост (α), т.е. рискът да се отхвърли вярна нулева хипотеза, се намира изцяло в едната страна (фиг. 6.2). Когато е двустранна, то се разпределя в двете страни по ½ (фиг. 6.1).

Освен това, при едностранна критична област критерият е по-мощен, отколкото при двустранна критична област.

Заключението при проверката на всяка хипотеза се прави като се сравняват емпиричната и теоретичната стойности на съответната характеристика. Ако емпиричната стойност е по-малка или равна на теоретичната, нулевата хипотеза се приема. В такъв случай се казва, че проверяваната разлика (например $\bar{x}_1 - \bar{x}_2$) е ***статистически незначима***, т.е. тя е случайна. Обратно, ако емпиричната стойност на характеристиката е по-голяма от теоретичната, това означава, че проверяваната

разлика е ***статистически значима*** (неслучайна), при което нулевата хипотеза се отхвърля и се приема алтернативната.

Конкретната характеристика (емпирична и теоретична) зависи от това каква хипотеза се проверява, за какви параметри се отнася и съобразно това, какъв метод за проверка се прилага.

Прилаганите методи (критерии) общо се делят на параметрични и непараметрични.

Параметричните методи са конструирани върху опознати разпределения и приложението им изисква обосновани предположения за вида на разпределенията.

Методите (критериите), които не изискват предварително да се знае видът на разпределението, се наричат ***непараметрични.***

По принцип параметричните методи (критерии) са по-точни и когато е възможно, те се предпочитат. По-конкретно при параметричните критерии рискът за грешка от втори род може да бъде по-малък (критерият е по-мощен).

Обстоятелството, че непараметричните критерии не изискват да се познава функцията на разпределението разширява сферата на възможното им приложение. Тя се разширява и от възможността да се прилагат при неинтервални скали (номинална, рангова и ординална скала). Приложението им е свързано с по-малко изчислителна работа и затова често се наричат ***"бързи" критерии***. Те имат по-малка мощност и затова често се казва, че те са ***по-слаби критерии***. Когато обаче извадката е достатъчно голяма, въз основа на непараметричните критерии могат да се получат обосновани резултати.

Непараметричните критерии се считат за ***по-консервативни***, по-слабо селективни. Това означава, че имат по-малка способност да разграничават нулевата хипотеза от алтернативната.

Не е възможно в рамките на тази книга да се разгледат всички възможни методи и техните особености.

Основните методи, които намират най-широко приложение, се основават на разгледани вече теоретични разпределения.

Те могат да се систематизират най-общо в три групи според изходните принципи, върху които са изградени.

1. Метод, който се основава на *t*-разпределението и има характеристика, която е отношение на проверяваната величина (разлика между средни, между относителни дялове и др.) и съответната й стандартна грешка.

$$t = \frac{|\theta_1 - \theta_2|}{\sigma_{(\theta_1 - \theta_2)}}. \qquad (6.1)$$

Според конкретния вид на хипотезата тази характеристика съответно се модифицира.

2. Метод, основаващ се на *F*-разпределението с характеристика дисперсионното отношение, т.е. отношение на две оценки на дисперсията ($\hat{\sigma}_1^2$ и $\hat{\sigma}_2^2$).

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}. \qquad (6.2)$$

3. Метод, основаващ се на χ^2-разпределението с характеристика:

$$\chi^2 = \sum_{i=1}^{k} \left[\frac{(f_i - \hat{f}_i)^2}{\hat{f}_i} \right], \qquad (6.3)$$

където f_i са честотите на даденото емпирично разпределение, а $\hat{f}_i$ - съответстващите им честоти на теоретичното разпределение (например нормалното).

Общата схема (алгоритъмът) на проверката се състои от последователни стъпки.

Първо. Дефинират се нулевата (проверяваната) хипотеза (H_0) и алтернативната хипотеза (H_1). Чрез начина на дефиниране на H_1 се задава критичната област - едностранна или двустранна.

Второ. Определя се методът за проверка. Той зависи от параметрите или свойствата на съвкупността, за които се отнася

хипотезата, от условията за използването на един или друг метод, от стремежа към по-мощен критерий (при равни други условия).

Трето. Изчислява се емпиричната характеристика на критерия по данните от извадките.

Четвърто. Намира се теоретичната стойност на характеристиката по стандартната таблица на съответното теоретично разпределение при предварително зададено равнище на значимост и определените степени на свобода.

Пето. Сравняват се емпиричната и теоретичната стойности на характеристиката и се прави заключение.

Ако емпиричната е по-малка или равна на теоретичната, нулевата хипотеза (H_0) се приема. Ако емпиричната е по-голяма от теоретичната, нулевата хипотеза (H_0) се отхвърля и се приема алтернативната (H_1).

Не бива да се забравя, че заключението е вероятностно. Ако в резултат на проверката се приеме нулевата хипотеза не значи, че сигурно тя е вярна и ако се отхвърли - че не е вярна. Така е, защото винаги има риск за грешка, макар той да е сведен до желания минимален размер.

Приемането на нулевата или алтернативната хипотеза означава само, че наличните емпирични данни са в нейна полза и то винаги със зададена вероятност (риск) за грешка. Както подчертава ***Ф. Милс***, "Същността на статистическата проверка е такава, че на фактите се дава възможност да докажат погрешността на хипотезата, но фактите не доказват хипотезата."[1]

6.3. Проверка на хипотези относно разлика между средни величини

При конкретни емпирични изследвания не рядко се налага да се проверяват хипотези относно разликата между две средни величини. Постановката на познавателната задача може да бъде различна.

Възможно е да разполагаме със средната на дадена генерална съвкупност и със средната на извадка от същата съвкупност. Между тези две средни ще има разлика. Проверката ще даде отговор на въпроса:

[1] **Миллс, Ф.** Статистические методы. М., 1958, с. 241.

случайна ли е тази разлика и може да се приеме, че извадката е представителна за генералната съвкупност, или тази разлика е статистически значима (неслучайна). В друг случай може да е необходима проверка на разлика между средни на две извадки и т.н.

6.3.1. Проверка на хипотеза относно разлика между средни на генерална съвкупност и на извадка

Възможно е средната аритметична на генералната съвкупност да е зададена като определен стандарт (задължителен размер на детайл, срок на годност на препарат и др.). Например един автомат е настроен да произвежда детайл с определен размер. По същество това е център на разпределението ($\bar{x}_0$) с допустими граници на отклонение, т.е. с определена вариация, изразена чрез дисперсията или стандартното отклонение (σ_0). Направена е извадка и е изчислен средният размер на детайлите - $\bar{x}$. Задачата е да се провери дали разликата между $\bar{x}_0$ (средната на генералната съвкупност) и $\bar{x}$ (средната на извадката) е в границите на допустимите случайни колебания, или е значима (неслучайна) и говори за евентуално разстройване на автомата.

Нулевата хипотеза ще гласи, че разликата е случайна, т.е.

$$H_0 : \bar{x} - \bar{x}_0 = 0 .$$

На нулевата хипотеза се противопоставя алтернативната, която може да гласи, че има значима разлика:

$$H_1 : \bar{x} \neq \bar{x}_0 .$$

Този начин на дефиниране на H_1 означава, че критичната област е двустранна.

Ако извадката е достатъчно голяма и имаме основание да се опрем на нормалното разпределение, критерият за проверка на разликата между двете средни се основава върху характеристиката

$$z = \frac{\left|\bar{x} - \bar{x}_0\right|}{\frac{\sigma_0}{\sqrt{n}}} \quad , \tag{6.4}$$

която по същество е позната от описанието на нормалното разпределение. Както се вижда, разликата между двете средни е отнесена към стандартната грешка. Това съответства на принципно положение, представено с формула 6.1.

Ако извадката е сравнително малка и не може да се използва нормалното разпределение, трябва да се използва характеристиката

$$t = \frac{\left|\bar{x} - \bar{x}_0\right|}{\frac{\hat{\sigma}}{\sqrt{n}}}. \tag{6.5}$$

Теоретичната стойност на тази характеристика ($t_т$) се намира в стандартната таблица на *t*-разпределението на Стюдент при предварително зададеното равнище на значимост (α). По общото правило, ако $t \le t_т$, нулевата хипотеза се приема, т.е. не се доказва статистически значима (неслучайна) разлика между $\bar{x}$ и $\bar{x}_0$. Обратно, ако $t > t_т$, нулевата хипотеза се отхвърля.

6.3.2. Проверка на хипотеза относно разлика между средни на две извадки

Ако от дадена генерална съвкупност са направени две извадки с обеми n_1 и n_2, техните средни аритметични $\bar{x}_1$ и $\bar{x}_2$ ще се различават. Предполага се, че разликата може да се дължи на определен фактор. Но тя може да е случайна. Смисълът на проверката и в този случай е да се установи дали е случайна или статистически значима (неслучайна).

Нулевата хипотеза гласи:

$H_0 : \bar{x}_1 - \bar{x}_2 = 0$ или $H_0 : \bar{x}_1 = \bar{x}_2$.

Според конкретния случай и очакването за неслучайност на извадката, алтернативната хипотеза ще се дефинира с едностранна или двустранна критична област:

$$H_1 : \bar{x}_1 > \bar{x}_2 \ ; \ H_1 : \bar{x}_1 < \bar{x}_2 \ ; \ H_1 : \bar{x}_1 \neq \bar{x}_2 \ .$$

При проверка на такава хипотеза характеристиката на критерия е *t*, която е отношение на проверяваната разлика към стандартната грешка (стандартното отклонение).

$$(6.6) \qquad t = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sigma_{(\bar{x}_1 - \bar{x}_2)}} .$$

Известно е, че дисперсията на средните на извадките е $\frac{\sigma_o^2}{n}$. Известно е също, че дисперсията на разлика между две независими случайни величини е:

$$(6.7) \qquad \sigma^2_{(x_1 - x_2)} = \sigma^2_{x_1} + \sigma^2_{x_2} .$$

Следва, че дисперсията на средните на двете извадки е:

$$(6.8) \qquad \frac{\sigma_o^2}{n_1} + \frac{\sigma_o^2}{n_2} = \sigma_o^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right).$$

И тъй като σ_o^2 не е известно, то трябва да се замести с неговата оценка. Но тя трябва да бъде ***обединена*** оценка от оценките, получени от двете извадки ($\hat{\sigma}_1^2$ и $\hat{\sigma}_2^2$). От формула 5.24 в глава 5 е известно, че обединената оценка на дисперсията е:

$$(6.9) \qquad \hat{\sigma}^2 = \frac{\sigma_1^2 n_1 + \sigma_2^2 n_2}{n_1 + n_2 - 2} .$$

Ако с този израз се замести σ_0^2 във формула (6.8), ще се получи стандартната грешка (стандартното отклонение) на разликата между средните на двете извадки:

(6.10) $$\sqrt{\frac{\sigma_1^2 n_1 + \sigma_2^2 n_2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

Следователно формула 6.6 ще приеме вида:

(6.11) $$t = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{\frac{\sigma_1^2 n_1 + \sigma_2^2 n_2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

или още:

(6.12) $$t = \frac{\left|\bar{x}_1 - \bar{x}_2\right|\sqrt{n_1 + n_2 - 2}}{\sqrt{\left(\sigma_1^2 n_1 + \sigma_2^2 n_2\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

(Формулата се модифицира в известна степен, когато средните са с еднакъв обем или са излъчени от различни съвкупности с различни дисперсии).

След като е изчислена емпиричната стойност на *t*, теоретичната ($t_т$) се намира в таблицата за *t*-разпределението на Стюдент (приложение 5) при зададените равнище на значимост (α) и степени на свобода, които в случая са $\phi = n_1 + n_2 - 2$.

При сравняване на *t* с $t_т$ по общото правило нулевата хипотеза се приема или отхвърля.

Проверката на хипотеза относно разликата между средните на две извадки може да се илюстрира със следния ***пример***.

Селекциониран е нов сорт пшеница, за който се твърди, че е по-високодобивен от друг - базов, контролен. Експериментално са засети в опитно поле 30 парцели (n_1) с новия сорт и 22 парцели (n_2) с контролния сорт. Получен е среден добив от новия сорт 650 кг от дка ($\bar{x}_1$) при дисперсия $\sigma_1^2 = 260$, а от контролния - среден добив 600 кг от дка ($\bar{x}_2$) при дисперсия $\sigma_2^2 = 240$. Необходимо е да се провери хипотеза относно разликата между средните добиви от двата сорта. Поради важността на експеримента се предвижда заключението да се направи с равнище на значимост α = 0,01.

Съгласно нулевата хипотеза разликата е случайна (статистически незначима) и се дефинира:

$$H_0 : \bar{x}_1 - \bar{x}_2 = 0 .$$

В случая по алтернативната хипотеза се предполага (и това примерно твърдят селекционерите), че средният добив от новия сорт е значимо (неслучайно) по-висок и че това се дължи на новите му качества. Затова алтернативната хипотеза се задава така:

$$H_1 : \bar{x}_1 > \bar{x}_2 ,$$

което означава още, че критичната област е едностранна.

Емпиричната характеристика е

$$t = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{\frac{\sigma_1^2 n_1 + \sigma_2^2 n_2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\left|650 - 600\right|}{\sqrt{\frac{(260.30) + (240.22)}{30 + 22 - 2}\left(\frac{1}{30} + \frac{1}{22}\right)}} = 11{,}04 .$$

В таблицата за *t*-разпределението (приложение 5) при $\alpha = 0{,}01$, степени на свобода $\phi = 30 + 22 - 2 = 50$ и едностранна критична област се намира теоретичната стойност на характеристиката $t_т = 2{,}39$. Щом $t > t_т$, нулевата хипотеза се отхвърля и се приема алтернативната.

Казано по друг начин, заключението гласи: получените данни от експеримента дават основание да се твърди с 1 %-ов риск за грешка, че разликата между средните добиви от новия и от контролния сорт е статистически значима (неслучайна).

6.4. Проверка на хипотеза относно разлика между относителни дялове

Критерият за проверка на хипотеза относно разликата между относителни дялове ($p_1 - p_2$) е същият, както при разлика между средни величини. Формулата на характеристиката *t* се различава само по начина, по който се намира стандартната грешка, т.е. стандартното отклонение. Известно е, че то е $\sqrt{pq}$. Затова формула 6.11 ще се модифицира в:

(6.13) $$t = \frac{|p_1 - p_2|}{\sqrt{\frac{p_1 q_1 n_1 + p_2 q_2 n_2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$, или

по аналогия с формула 6.12:

(6.14) $$\frac{|p_1 - p_2|\sqrt{n_1 + n_2 - 2}}{\sqrt{(p_1 q_1 n_1 + p_2 q_2 n_2)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Във формулите p е относителният дял на единиците, които имат дадено качество, а q - единиците, които нямат това качество. Степените на свобода са $\phi = n_1 + n_2 - 2$.

(Решен пример е даден в точка 6.7).

6.5. Проверка на хипотеза относно разлика между дисперсии

Постановката при такъв вид хипотези също може да бъде различна. Например числовата стойност на дисперсията в една генерална съвкупност е известна от минали наблюдения и тя е $\sigma_0^2 = a$. От направена случайна извадка е изчислена оценка на дисперсията, която е различна от a. В случая се проверява дали в действителност оценката по извадката не се различава от a (нулева хипотеза) или разликата е значима (алтернативна хипотеза). В други случаи се налага да се проверява хипотеза относно разликата между оценките на дисперсиите, изчислени от две извадки ($\hat{\sigma}_1^2$ и $\hat{\sigma}_2^2$).

6.5.1. Проверка на хипотеза относно дисперсията на генералната съвкупност

Нулевата хипотеза в този случай е:

$H_0 : \sigma_0^2 = a$,

което означава, че по предположение дисперсията в генералната съвкупност е такава, каквато е установена в предходно изследване.

Според алтернативната хипотеза:

$$H_1 : \sigma_0^2 \neq a \ ,$$

което означава, че са настъпили изменения и дисперсията вече не е равна на a, т.е. разликата е статистически значима. (В някои случаи алтернативната хипотеза може да се дефинира с едностранна критична област: $H_1 : \sigma_0^2 > a$ или $H_1 : \sigma_0^2 < a$).

Хипотезата се проверява чрез критерий с характеристика χ^2. Емпиричната ѝ стойност се изчислява по формулата:

$$\chi^2 = \frac{(n-1)\hat{\sigma}^2}{a}, \qquad (6.15)$$

където $\hat{\sigma}^2$ е оценката на дисперсията, изчислена по данните от извадката, а степените на свобода са $\phi = n-1$. Тъй като дисперсията на извадката е изместена и затова се изчислява по формулата $\hat{\sigma}^2 = \frac{\sum (x-\bar{x})^2}{n-1}$, формула 6.15 може да приеме вида:

$$\chi^2 = \frac{(n-1)\frac{\sum (x-\bar{x})^2}{n-1}}{a} = \frac{\sum (x-\bar{x})^2}{a}. \qquad (6.16)$$

Теоретичната стойност на характеристиката ($\chi^2_{\text{т}}$) се намира в таблицата за χ^2-разпределението на Пирсън (приложение 6) при зададеното равнище на значимост (α) и степените на свобода $\phi = n-1$.

Проверката на такава хипотеза може да се илюстрира с ***пример***.

От минали наблюдения в един отрасъл на промишлеността е известно, че дисперсията на заплатите е $\sigma_0^2 = a = 5200$. През септември 2008 г. е направена случайна извадка от 81 заети лица и е изчислена оценка на дисперсията $\hat{\sigma}^2 = 4800$. Необходимо е да се провери дали разликата между 4800 и 5200 е случайна (нулева хипотеза) или е статистически значима (неслучайна) и дава основание да се смята, че са

настъпили промени във вариацията на заплатите. Заключението ще се направи с равнище на значимост $\alpha = 0{,}01$.

Емпиричната стойност на характеристиката е:

$$\chi^2 = \frac{(n-1)\hat{\sigma}^2}{a} = \frac{(81-1)4800}{5200} = 73{,}84 .$$

При степени на свобода $\phi = 81-1 = 80$ и равнище на значимост $\alpha = 0{,}01$ в таблицата за χ^2-разпределението (приложение 6) се намира $\chi^2_{\text{т}} = 112{,}3$. Следователно $\chi^2 < \chi^2_{\text{т}}$. Това означава приемане на нулевата хипотеза, т.е. няма основание да се твърди, че са настъпили промени във вариацията на заплатите.

6.5.2. Проверка на хипотеза относно разлика между две оценки на дисперсията

Една от възможните познавателни задачи, изискваща проверка на хипотеза е тази, когато от дадена генерална съвкупност са направени две независими случайни извадки, но излъчени по различен начин. По информацията в извадките са изчислени две независими оценки на дисперсията. Необходимо е в този случай да се провери дали разликата между оценките е случайна и следователно двете оценки възпроизвеждат еднакво дисперсията в генералната съвкупност, или разликата е статистически значима (неслучайна) и трябва да се потърси причината за това. В друг случай може да се проверява разлика между оценки на дисперсията на две генерални съвкупности или подсъвкупности на една съвкупност.

Нулевата хипотеза ще гласи, че $\hat{\sigma}_1^2 = \hat{\sigma}_2^2$.

Критерият (тестът) за проверка обикновено е *F*-характеристика, която представлява отношение на две независими оценки на дисперсията:

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} . \qquad (6.17)$$

Ако двете извадки са с обеми n_1 и n_2, като се има предвид, че дисперсиите на извадките са изместени оценки, те (оценките) се изчисляват по познатите вече формули:

$$\hat{\sigma}_1^2 = \sigma_1^2 \frac{n_1}{n_1 - 1} \qquad \text{или} \qquad \hat{\sigma}_1^2 = \frac{\sum (x_1 - \bar{x}_1)^2}{n_1 - 1};$$

$$\hat{\sigma}_2^2 = \sigma_2^2 \frac{n_2}{n_2 - 1} \qquad \text{или} \qquad \hat{\sigma}_2^2 = \frac{\sum (x_2 - \bar{x}_2)^2}{n_2 - 1}.$$

Теоретичната стойност на характеристиката ($F_т$) се намира в таблицата за *F*-разпределението на Фишер (приложение 7) при зададеното равнище на значимост и степени на свобода $\phi_1 = n_1 - 1$ и $\phi_2 = n_2 - 1$. Трябва при това да се има предвид, че в стандартната таблица обикновено *F* е по-голямо от единица. Затова, ако емпиричното *F* , изчислено по формула 6.17 е по-малко от единица, с теоретичното трябва да се сравнява $\frac{1}{F}$, или в числителя на формулата на емпиричното *F* да се поставя по-голямата оценка на дисперсията.

Да приемем като ***пример***, че в един пункт за екологични наблюдения са направени измервания с два уреда на съдържанието на серен окис във въздуха. С единия уред са направени 24 измервания (n_1), а с другия - 25 измервания (n_2). От измерванията с първия уред е изчислена оценка на дисперсията $\hat{\sigma}_1^2 = 1{,}69$, а от измерванията с втория уред - $\hat{\sigma}_2^2 = 1{,}21$.

Необходимо е да се провери хипотезата относно разликата между двете оценки на дисперсията при равнище на значимост $\alpha = 0{,}05$.

Емпиричната характеристика, изчислена по формула 6.17 е:

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{1{,}69}{1{,}21} = 1{,}4.$$

При приетото равнище на значимост 0,05 и степени на свобода $\phi_1 = 24 - 1 = 23$ и $\phi_2 = 25 - 1 = 24$ в таблицата за *F*-разпределението (приложение 7) се намира $F_т = 1{,}98$. Очевидно е, че $F < 1{,}98$. Следовател-

но, не може да се отхвърли нулевата хипотеза, т.е. не се доказва значимо различие в точността на измерванията с двата уреда. Разликата може да се приеме за случайна. Това заключение се прави при 5 %-ов риск за грешка.

6.6. Проверка на хипотеза относно съответствието между емпирично и теоретично разпределение

Емпиричните разпределения са разнообразни по форма. Голяма част от тях обаче са сходни с определени теоретични разпределения, т.е. теоретичните могат да служат като модели на емпиричните. При статистическия анализ често се използват свойствата на теоретичните разпределения и преди всичко на нормалното разпределение и свързаните с него извадкови разпределения. Но дали дадено емпирично разпределение съответства на теоретичното не може да се установи само чрез визуално сравняване на графичните им образи, или само чрез коефициентите за асиметрия и ексцес, когато еталонът е нормалното разпределение. Необходимо е по-прецизно средство за проверка съгласува ли се емпиричното разпределение с предполагаемото теоретично и в частност с нормалното. Такова средство е проверката на хипотези и по-конкретно съдържащите се в тази теория ***критерии за съгласуваност***.

Теоретично са изведени и доказани различни критерии, включително и непараметрични.[1] Най-популярен и широко прилаган е критерият на ***Пирсън***.

Критерият χ^2 (хи-квадрат) на Пирсън е с характеристика:

$$\chi^2 = \sum_{i=1}^{k} \left[\frac{\left(f_i - \hat{f}_i\right)^2}{\hat{f}_i} \right] , \qquad (6.18)$$

където:

f_i са честотите на емпиричното разпределение (емпиричните честоти);

[1] Относно някои широко прилагани непараметрични методи вж. **Lind, D., W. Marchal, S. Wathen**, Statistical Techniques in Business and Economics, New York, 2005.

$\hat{f}_i$ - честотите на теоретичното разпределение (теоретичните честоти);

k - броят на груповите интервали, по които са разпределени честотите (броят на обособените групи при разпределението).

Ако емпиричните честоти напълно съвпадат с теоретичните, $\chi^2 = 0$. Обикновено няма такова пълно съвпадение и затова практически χ^2 характеристиката, изчислена по формула 6.18 е по-голяма от 0 и то толкова по-голяма, колкото са по-големи различията между емпиричното и теоретичното разпределение. Задачата на проверката е да се установи дали различията са случайни (статистически незначими) и може да се приеме, че емпиричното разпределение се подчинява на закона на теоретичното или тези различия са значими (неслучайни).

Следователно нулевата хипотеза в случая ще гласи, че емпиричните честоти са равни на теоретичните. Доколкото има разлики, те са случайни.

Алтернативната хипотеза ще гласи, обратно, че разликите не са случайни, дължат се на това, че даденото теоретично разпределение не е адекватния модел на емпиричното. Заключението ще се направи, както при всякакви хипотези, чрез сравняване на емпиричната стойност на χ^2 с теоретичната, определена по таблицата на χ^2-разпределението (приложение 6) при предварително определеното равнище на значимост (α) и степените на свобода.

В този случай степените на свобода са $\phi = k - m - 1$, където m е броят на параметрите на теоретичното разпределение. Ако теоретичното разпределение примерно е Поасоновото, то има единствен параметър λ и степените на свобода са $\phi = k - 1 - 1 = k - 2$. Ако е нормалното разпределение, то има два параметъра - математическото очакване ($\bar{x}$) и дисперсията (σ^2), затова степените на свобода са $\phi = k - 2 - 1 = k - 3$.

За да е коректно приложението на χ^2-критерия на Пирсън, трябва да се спазват някои условия, изведени и теоретично, и от многократни опити. Задължително условие е емпиричните данни да са получени от случайна извадка. Обемът на извадката да съдържа най-малко 50 единици. Препоръчва се броят на единиците (честотите) в отделните

групови интервали да не е под 5. Практически това означава, че ако има интервали с по-малки честоти от 5, те трябва да се обединят със съседните. [1]

От техническа гледна точка основният въпрос при изчисляване на характеристиката χ^2 е намирането на теоретичните честоти ($\hat{f}_i$). За това има различни механизми. Ако теоретичното разпределение е нормалното, за намирането на $\hat{f}_i$ може да се приложи практически удобен алгоритъм, който се състои от няколко стъпки.

1. Намират се средната аритметична ($\bar{x}$) и стандартното отклонение (σ) на емпиричното разпределение.
2. Изчисляват се нормираните отклонения $z_i = \dfrac{x_i - \bar{x}}{\sigma}$.
3. По таблицата за функцията $f(z)$ на нормалното разпределение се намират стойностите на $f(z)$ за всяка стойност на z_i.
4. Намерените по този начин стойности на $f(z)$ се умножават по общата сума на честотите, т.е. обема на извадката $n = \sum f$ и по ширината на груповите интервали (h) и получените произведения се разделят на стандартното отклонение. Получените по този начин числа са търсените теоретични честоти ($\hat{f}_i$) на нормалното разпределение.

Приложението на критерия на Пирсън може да се илюстрира със следния ***пример***. Да приемем, че е дадено разпределението на 300 младежи донаборници по признака ръст (табл. 6.2).

[1] В литературата се срещат и други граници относно минималния обем на извадката и на честотите в отделните интервали. Вж. **Закс, Л.**, Цит. съч., с.295.

Таблица 6.2

Разпределение на младежи - донаборници по ръст

Групови интерва-ли в см.	Среди на интерва-лите x_i	Брой на младе-жите f_i	$z_i = \frac{x_i - \bar{x}}{\sigma}$	$f(z)$	$\hat{f}_i = \frac{f(z).300.2}{4{,}78}$	$\frac{(f_i - \hat{f}_i)^2}{\hat{f}_i}$
1	2	3	4	5	6	7
162 - 164	163	19	- 1,67	0,09893	13	2,769
164 - 166	165	21	- 1,25	0,18265	23	0,174
166 - 168	167	35	- 0,83	0,28269	36	0,027
168 - 170	169	49	- 0,41	0,36678	46	0,196
170 - 172	171	55	0,00	0,39894	51	0,314
172 - 174	173	47	0,42	0,36526	46	0,022
174 - 176	175	32	0,84	0,28034	36	0,444
176 - 178	177	22	1,26	0,18037	23	0,043
178 - 180	179	20	1,68	0,09728	13	3,769
		300			297	7,758

Предполага се, че това разпределение се подчинява на закона на нормалното разпределение и доколкото има разлики между емпиричните и теоретичните честоти, те са случайни. Това е нулевата хипотеза. За да се провери тя, необходимо е да се изпълнят описаните процедури.

Средната аритметична (средният ръст на младежите) е $\bar{x} = 170{,}98$ см, а стандартното отклонение - $\sigma = 4{,}78$ см.

Намират се нормираните отклонения на средите на интервалите (x_i) от средната аритметична

$$\left(z_i = \frac{x_i - \bar{x}}{\sigma} \right). \frac{163 - 170{,}98}{4{,}78} = 1{,}67; \ \frac{165 - 170{,}98}{4{,}78} = -1{,}25 \text{ и т.н.}$$

(показани са в колона 4 на табл. 6.2).

В таблицата за функцията $f(z)$ на нормалното разпределение се намира $f(z)$ за всяка стойност на z_i (приложение 3). Показани са в колона 5 на табл. 6.2. Получените стойности на $f(z)$ се умножават по сумата на честотите $(\sum f = 300)$ и по ширината на интервалите $(h = 2)$ и разликата се разделя на стандартното отклонение $(\sigma = 4{,}78)$. Така се получават теоретичните честоти ($\hat{f}_i$), показани в колона 6 на табл. 6.2.

По принцип сумата на теоретичните честоти трябва да е равна на сумата на емпиричните. В примера се получава разлика от 3 единици. Това се дължи на обстоятелството, че честотите на двата крайни интервала (първият и последният) са твърде много различни от нула, а нормалното разпределение има граница от $-\infty$ до $+\infty$, т.е. двата му края се приближават асимптотично към нулата. В такива случаи, когато се получават по-големи разлики между $\sum f_i$ и $\sum \hat{f}_i$ може да се прибавят към скалата на емпиричното разпределение по един групов интервал преди първия и след последния с емпирични честоти 0. Тогава сумата на теоретичните честоти ще се изравни със сумата на емпиричните.

Разликите между емпиричните (f_i) и теоретичните ($\hat{f}_i$) честоти, повдигнати на квадрат $\left(f_i - \hat{f}_i\right)^2$ са разделени на теоретичните честоти ($\hat{f}_i$) и са записани в последната колона на табл. 6.2 - $\left(\frac{\left(f_i - \hat{f}_i\right)^2}{\hat{f}}\right)$. Тяхната обща сума е търсената емпирична характеристика $\chi^2 = 7{,}758$.

Груповите интервали са $k = 9$. Нормалното разпределение има 2 параметъра - математическото очакване и дисперсията. Следователно степените на свобода са $\phi = 9 - 2 - 1 = 6$. Приемаме заключението да се направи при равнище на значимост $\alpha = 0{,}01$. В стандартната таблица за χ^2-разпределението (приложение 6) се намира при $\alpha = 0{,}01$ и $\phi = 6$ теоретичната (критичната) стойност на $\chi^2_т = 16{,}812$.

Тъй като $\chi^2 < \chi^2_т$ (7,758 < 16,812) следва ***заключение:***

Критерият χ^2 на Пирсън потвърждава нулевата хипотеза, т.е. нормалното разпределение е адекватен модел на емпиричното. Разликите между емпиричните и теоретичните честоти могат да се приемат за случайни. Това заключение се прави при равнище на значимост $\alpha = 0{,}01$, т.е. с 1 %-ов риск за грешка.

6.7. Практикум

6.7.1. Въпроси за самопроверка

1. Какво е статистическото понятие за хипотеза?
2. Какво се разбира под грешка от първи род и грешка от втори род?
3. Какво значи равнище на значимост?
4. Какъв е смисълът на критичната област?
5. Какво се разбира под мощност на критерия?
6. Какво значи теоретична (критична) стойност на характеристиката?
7. Каква е характеристиката на критерия при проверка на хипотеза относно разлика между средни на две извадки?
8. Колко са степените на свобода при проверка на хипотеза относно разликата между относителните дялове на нестандартните изделия, произведени през първия и последния работен ден на седмицата?
9. Какво представлява критерият на Пирсън за проверка на съответствието между емпирично и теоретично разпределение?

6.7.2. Задачи за упражнение

Задача 1. В опитно поле е направен експеримент, за да се провери ефекта от нова технология за наторяване на захарно цвекло. От засадени общо 42 опитни парцели, 22 (n_1) са наторявани по новата и 20 (n_2) по старата технология. От първите е получен 7200 кг среден добив от декар ($\bar{x}_1$), а от вторите - 6900 кг от декар ($\bar{x}_2$). Стандартното отклонение при първите (σ_1) е 260 кг, а при вторите (σ_2) - 275 кг.

Необходимо е да се провери значима (неслучайна) ли е разликата между средните добиви при наторяване по различните технологии. Поради практическото значение на експеримента, проверката да се направи с равнище на значимост $\alpha = 0{,}01$.

Отговор: Нулевата хипотеза не се отхвърля, т.е. не се доказва значима разлика между добивите.

Задача 2. В една фирма специалистите по контрол на качеството са решили да проверят дали има разлика между произвежданите нестандартни изделия от едни и същи работници през първия и последния ден на работната седмица. Проверени са чрез случайна извадка 28 изделия (n_1), произведени в понеделник и 34 изделия (n_2), произведени в петък. Установено е, че от произведените в понеделник нестандартни са 2 % (p_1), а от произведените в петък - 2,6 % (p_2).

Необходимо е да се провери хипотеза относно статистическата значимост на разликата ($p_1 - p_2$) в относителните дялове. Проверката да се направи при равнище на значимост $\alpha = 0{,}05$.

Решение:

Хипотезите (нулевата и алтернативната) се дефинират по следния начин:

$$H_0 : p_1 - p_2 = 0$$
$$H_1 : p_2 > p_1$$

Характеристиката на критерия (теста) се изчислява по формулата:

$$t = \frac{|p_1 - p_2|}{\sqrt{\frac{p_1 q_1 n_1 + p_2 q_2 n_2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} =$$

$$\frac{0{,}026 - 0{,}020}{\sqrt{\frac{0{,}020.0{,}98.28 + 0{,}026.0{,}974.34}{28 + 34 - 2}\left(\frac{1}{28} + \frac{1}{24}\right)}} = 0{,}141.$$

По таблицата за *t*-разпределението (приложение 5) при степени на свобода $\phi = 28 + 34 - 2 = 60$ и равнище на значимост $\alpha = 0{,}05$ се намира теоретичната стойност на характеристиката $t_т = 1{,}67$.

Заключение: Тъй като $t < t_т$, няма основание да се отхвърли нулевата хипотеза, т.е. емпиричните данни не дават основание разликата между относителните дялове на нестандартните изделия, произведени в

понеделник и в петък, да се приеме за статистически значима. Това заключение се прави при 5 %-ов риск за грешка.

Задача 3. При статистическия контрол на качеството на продукцията и технологичните процеси в една машиностроителна фирма е направена случайна извадка от 143 бр. детайли от даден вид, произвеждани през определено време от автоматичен струг. Размерите им са измерени с микрометър и са разпределени в интервали с еднаква ширина. Резултатите се съдържат в следващата таблица.

Таблица 6.3

Разпределение на проверените детайли по размер в милиметри

Групови интервали в мм	Брой на детайлите
40,25 - 40,26	3
40,26 - 40,27	8
40,27 - 40,28	19
40,28 - 40,29	27
40,29 - 40,30	36
40,30 - 40,31	25
40,31 - 40,32	16
40,32 - 40,33	7
40,33 - 40,34	2

Предполага се (дефинира се хипотеза), че това емпирично разпределение се подчинява на закона на нормалното разпределение, че различието между емпиричните и теоретичните честоти е случайно. Необходимо е да се провери хипотезата при равнище на значимост $\alpha = 0{,}01$.

Отговор: $\chi^2 = 0{,}847$

$\chi^2_{\text{т}} = 13{,}27$ $\chi^2 < \chi^2_{\text{т}}$,

затова нулевата хипотеза се потвърждава.

7. ДИСПЕРСИОНЕН АНАЛИЗ

> "Статистическият анализ ... може да направи мисленето и умението по-ефективни за постигането на правилни научни резултати."
>
> М. Езекиел и К. Фокс

Съдържанието на главата предлага специфична методология, която по същество също е проверка на хипотези, но по отношение на предполагаемото действие на определени фактори върху даден резултат – печалба, качество на продукта, доход, пазарен дял, лечебен ефект от прилагане на определени медикаменти и много други. Читателят ще се запознае с общите принципи и подходи и по-конкретно с еднофакторния анализ, илюстриран с конкретен последователно развит пример и с интерпретирането на получените резултати.

7.1. Обща постановка

Дисперсионният анализ е статистическа методология, която намира широко приложение в много области и преди всичко в експерименталното дело. Неговото начало, както и наименованието му, е поставено от ***Роналд Фишер*** през 20-те години на миналия век. Понататък се развива сравнително бързо като цялостна теория за анализ на резултати от експерименти и други наблюдения, намиращи се под влиянието на различни, едновременно действащи фактори.

Основната познавателна задача на дисперсионния анализ е да се провери дали може да се приеме, че даден предполагаем фактор (или повече фактори) влияе значимо върху интересуващ ни резултативен признак. По същество това е проверка на хипотеза.

В предходната глава беше установено, че ако по даден факторен признак са обособени две съвкупности, направени са от тях две извадки и са изчислени техните средни величини, то хипотезата относно разликата между средните може да се провери например чрез критерий с *t*–

характеристика. Ако се проверява разлика между дисперсиите на две извадки, използва се *F*–характеристиката. Описаният конкретен подход обаче е неприложим, когато по предполагаемия фактор са обособени три или повече съвкупности, въз основа на случайни извадки са изчислени три или повече средни и трябва да се провери хипотеза относно разликата между тях.

Може да се предполага, че ако има разлика между средните, те се дължат на влиянието на фактора, но може тези разлики да са случайни. Нулевата хипотеза ще гласи: $H_0 \quad \bar{x}_1 = \bar{x}_2 = \bar{x}_3 = ... = \bar{x}_k$, а алтернативната - $H_1 \quad \bar{x}_1 \neq \bar{x}_2 \neq \bar{x}_3 \neq ... \neq \bar{x}_k$. Казано по друг начин, ако има факторно влияние, ще има ***факторна вариация*** между средните. Тя обаче не може да се измери пряко. Дали има или няма такава вариация може да се установи чрез апарата на дисперсионния анализ. Основният принцип се състои в разлагането на общата ***девиация*** (сумата от квадратите на разликите) на компоненти, чрез които се изчисляват и се съпоставят независими оценки на дисперсията, за да се получи дисперсионно отношение. Както при всички други хипотези, чрез сравняване на емпиричната с теоретичната му стойност се прави заключението.

Приложението на дисперсионния анализ е свързано с някои условия (предпоставки).

1. Необходимо е резултативният признак да е представен на интервална скала, т.е. да има числови измерения. Факторните признаци може да са представени на каквато и да е скала, обикновено на номинална или ординална.

2. Информацията да е получена чрез независими случайни извадки.

3. Извадките да произлизат от генерални съвкупности с нормално разпределение и еднакви дисперсии.[1]

[1] При еднакви дисперсии се използва терминът **хомоскедастичност**, а при различни дисперсии - **хетероскедастичност**. Относно проверката за хомоскедастичност (критериите на Бартлет, на Кокран и Хартли) вж. **Закс**, **Л**. Цит. съч., с. 448 и сл. Относно последиците при отсъствието на посочените условия вж. **Гласс**, **Дж.** и **Дж. Стенли**, Статистические методы в педагогике и психологии, М., 1976, с. 333.

Може да се създаде впечатление, че тези условия са твърде строги и ограничаващи приложението на дисперсионния анализ. В действителност те почти винаги се осигуряват при конкретните изследвания.

Когато се проверява предполагаемото влияние на един фактор, дисперсионният анализ е ***еднофакторен***, при два фактора - ***двуфакторен***, при повече фактори - ***многофакторен***. Според разполагаемата информация, съдържаща се в извадките, се получават също различни варианти-изходна таблица с еднакъв или с различен брой случаи (честоти) в клетките. Конкретните модели на анализа също са различни.

В следващото изложение се разглежда еднофакторен дисперсионен анализ. Двуфакторният и многофакторният са извън рамките на тази книга. Трябва обаче да се има предвид, че основните принципи и подходи са общи.[1]

7.2. Еднофакторен дисперсионен анализ

Да приемем, че k съвкупности (с нормално разпределение и еднакви, но неизвестни дисперсии) са отграничени по разновидности на даден признак. От тях са направени k извадки (по една от всяка съвкупност), които съдържат $n_1, n_2, n_3,..., n_k$ единици. Общият брой на единиците във всички извадки е

$$n = n_1 + n_2 + n_3 + ... + n_k.$$

Отделните единици в извадките се различават по значенията на резултативния признак. Ако се намерят разликите между тези значения (x) и общата средна аритметична ($\bar{x}$), разликите се повдигнат на квадрат и се сумират, ще се получи общата сума на квадратите на разликите $\left(\Sigma(x-\bar{x})^2\right)$, която се нарича ***обща девиация***. Отделните значения на признака във всяка група (извадка) варират и около груповите средни ($\bar{x}_i$). Сумата на квадратите на отклоненията от груповите средни $\left(\Sigma(x-\bar{x}_i)^2\right)$ се нарича ***вътрешногрупова девиация***. За вариацията на

[1] За двуфакторния и многофакторния анализ вж. **Стефанов**, **Ив.** и **А**. **Тотев**, Теория на статистиката, С., 1960; **Шеффе**, **Г**. Дисперсионный анализ, М., 1980; **Хьютсон**, **А.** Дисперсионный анализ, М., 1971.

груповите средни ($\bar{x}_i$) около общата средна ($\bar{x}$) може да се намери ***междугруповата девиация*** $\left(\Sigma(\bar{x}_i - \bar{x})^2 n_i\right)$. Тези три девиации са свързани адитивно:

$$\sum(x - \bar{x})^2 = \sum(x - \bar{x}_i)^2 + \sum(\bar{x}_i - \bar{x})^2 n_i \,. \qquad (7.1)$$

Тази адитивна връзка между девиациите дава възможност да се намери всяка от тях при наличие на другите две.

На всяка от тези девиации съответствуват определени степени на свобода: на общата девиация - (n - 1); на вътрешногруповата - (n - k)[1]; на междугруповата - (k - 1). Те също са свързани адитивно:

$$(n - 1) = (n - k) + (k - 1). \qquad (7.2)$$

Като се разделят междугруповата и вътрешногруповата девиация на съответствуващите им степени на свобода, ще се получат две ***независими оценки на дисперсията***. Оценката, определена по междугруповата девиация ($\hat{\sigma}_1^2$), е

$$\hat{\sigma}_1^2 = \frac{\sum(\bar{x}_i - \bar{x})^2 n_i}{k - 1}, \qquad (7.3)$$

а по вътрешногруповата девиация - ($\hat{\sigma}_2^2$) -

$$\hat{\sigma}_2^2 = \frac{\sum(x - \bar{x}_i)^2}{n - k}. \qquad (7.4)$$

Като отношение на по-голямата оценка към по-малката[2] се получава емпиричната характеристика F, (дисперсионното отношение):

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}. \qquad (7.5)$$

[1] Получени са като сбор от степените на свобода за отделните групи (извадки):
(n_1 - 1) + (n_2 - 1) + (n_3 - 1) +...+ (n_k - 1) = (n - k)

[2] Таблиците за F-разпределението са съставени за $F_T > 1$. Затова е необходимо емпиричното F да се изчислява като отношение на по-голямата оценка към по-малката.

Може да се предполага, че разликата между двете оценки на дисперсията се предизвиква от фактора. Но тя може да бъде и съвсем случайна.

За да се провери дали различието между оценките, изразено в F, се причинява от действието на групировъчния признак (фактор) или не, трябва F да се сравни с теоретичното (табличното) $F_т$, което фиксира границата на допустимите случайни колебания, зад която се намира критичната област. По таблицата за F-разпределението (приложение 7) се намира $F_т$ при предварително прието равнище на значимост (α) и дадените степени на свобода- (k - 1) и (n - k).

Ако $F > F_т$, нулевата хипотеза се отхвърля и се приема алтернативната. Иначе казано, приема се, че разликата между груповите средни (средните на извадките) е статистически значима, което означава, че има основание да се твърди, че съществува влияние на факторния признак, по който са обособени групите. Ако $F \le F_т$, няма основание да се отхвърли нулевата хипотеза.

Изложения ход на проверката на хипотезата при еднофакторния дисперсионен анализ може да се обобщи в схема (табл. 7.1).

Таблица 7.1

Схема на еднофакторния дисперсионен анализ

Източници на вариация	Сума на квадратите на отклоненията на средните (девиация)	Степени на свобода	Оценки на дисперсията	Емпирично F	Теоретично F	
					при $\alpha = 0{,}05$	при $\alpha = 0{,}01$
Вътре в групите	$\sum (x - \bar{x}_i)^2$	n - k	$\hat{\sigma}_2^2 = \frac{\sum (x - \bar{x}_i)^2}{n - k}$	$F = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2}$	$F_т$	$F_т$
Между групите	$\sum (\bar{x}_i - \bar{x})^2 n_i$	k - 1	$\hat{\sigma}_1^2 = \frac{\sum (x_i - \bar{x})^2 n_i}{k - 1}$			
Обща	$\sum (x - \bar{x})^2$	n - 1				

Ще илюстрираме еднофакторния дисперсионен анализ с ***пример***.

В опитно поле е изпробвано наторяването на царевицата с комбиниран тор, имащ различен състав, предполагайки, че това има значение за средните добиви от декар. Съставени са 4 варианта торове, с всеки от които са наторявани по 5 опитни участъци (парцели) от по 1 декар, засети с царевица. Вариантите на торовете са означени с а, б, в, г, а опитните участъци са номерирани от 1 до 5.

Получените резултати са представени в табл. 7.2

Таблица 7.2

Средни добиви от царевица по опитни участъци

Вариан-ти тор	Добиви в кг. от декар по опитни участъци					Суми по редове	Средни на извадките $\overline{x}_i$
	1	2	3	4	5		
а	460	458	463	450	454	2285	457,00
б	470	470	465	452	453	2310	462,00
в	458	456	451	460	450	2275	455,00
г	462	460	458	472	468	2320	464,00
Суми по колони	1850	1844	1837	1834	1825	9190	(459,50)
Средни по участъци	462,50	461,00	459,25	458,50	456,25	-	-

Броят на участъците, наторявани с отделен вариант на комбиниран тор, може да се разглежда като отделна извадка от предполагаема генерална съвкупност. Дадени са следователно 4 случайни извадки, всяка от които с обем 5 опитни участъци. Предполага се, че експериментът е направен съгласно всички правила в теорията на опитното дело.

Между получените добиви има разлики (вариация), които се дължат на различни фактори, но в случая се предполага, че има факторна вариация, дължаща се на различните хранителни вещества в торовете. Но

може и да няма факторна вариация и разликите между добивите да са случайни (статистически незначими). Това трябва да се провери.

Нулевата хипотеза ще гласи, че разликите са случайни, т.е. $H_0 \quad \bar{x}_1 = \bar{x}_2 = \bar{x}_3 = \bar{x}_4$. Алтернативната ще гласи, че разликите са статистически значими (неслучайни), т.е. $H_1 \quad \bar{x}_1 \neq \bar{x}_2 \neq \bar{x}_3 \neq \bar{x}_4$.

Средните на извадките (по варианти торове) се намират по формулата на непретеглената средна аритметична ($\bar{x} = \frac{\sum x}{n}$). За вариант "а" тя е $\bar{x}_1 = \frac{2285}{5} = 457{,}0$ кг.; за вариант "б"- $\bar{x}_2 = \frac{2310}{5} = 462{,}0$ кг. и т.н. Получените средни са посочени в последната колона на табл. 7.2.

Общата средна се изчислява по същия начин - $\bar{x} = \frac{9190}{20} = 459{,}5$ кг., или от средните на извадките - $\bar{x} = \frac{457 + 462 + 455 + 464}{4} = 459{,}5$ кг. (Изчислява се в случая като непретеглена, тъй като средните на извадките се отнасят за еднакъв брой участъци).

Ако от добивите, получени от всичките опитни парцели (x) се извади общата средна ($\bar{x} = 459{,}5$), разликите се повдигнат на квадрат и се сумират, ще се получи общата девиация. Изчислена по данните в таблицата, тя е $\sum (x - \bar{x})^2 = 999$.

Изчисляването на междугруповата девиация е показано в табл. 7.3.

Таблица 7.3

$\bar{x}_i$	n_i	$\bar{x}_i - \bar{x}$ $(\bar{x} = 459{,}5)$	$(\bar{x}_i - \bar{x})^2$	$(\bar{x}_i - \bar{x})^2 n_i$
457	5	-2,5	6,25	31,25
462	5	2,5	6,25	31,25
455	5	-4,5	20,25	101,25
464	5	4,5	20,25	101,25
	20			265,00

Получава се междугрупова девиация $\left(\bar{x}_i - \bar{x}\right)^2 n_i = 265$.

Вътрешногруповата девиация може да се изчисли като разлика между общата и междугруповата:

$$\sum\left(x - \bar{x}_i\right)^2 = \sum\left(x - \bar{x}\right)^2 - \sum\left(\bar{x}_i - \bar{x}\right)^2 n_i = 999 - 265 = 734.$$

Степените на свобода за вътрешногруповата девиация са $\left(n - k\right) = 20 - 4 = 16$ и за междугруповата - $\left(k - 1\right) = 4 - 1 = 3$.

От междугруповата и вътрешногруповата девиация и дадените степени на свобода се изчисляват двете независими оценки на дисперсията:

$$\hat{\sigma}_1^2 = \frac{\sum\left(\bar{x}_i - \bar{x}\right)^2 n_i}{k - 1} = \frac{265}{3} = 88{,}33;$$

$$\hat{\sigma}_2^2 = \frac{\sum\left(x - \bar{x}_i\right)^2}{n - k} = \frac{734}{16} = 45{,}88.$$

Дисперсионното отношение е:

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{88{,}33}{45{,}88} = 1{,}93.$$

Тъй като заключението от експеримента би имало твърде важно практическо значение, нека приемем равнище на значимост $\alpha = 0{,}01$. В таблицата за *F*-разпределението (приложение 7) при $\alpha = 0{,}01$, $\phi_1 = 3$ и $\phi_2 = 16$ теоретичната (критичната) стойност на дисперсионното отношение е $F_{\text{Т}} = 5{,}29$.

Тъй като $F < F_{\text{Т}}$, нулевата хипотеза се приема, т.е. емпиричните данни от експеримента не потвърждават предполагаемото влияние на различния състав на торовете върху добивите. Доколкото има разлика в получените добиви, тя е в границите на допустимите случайни колебания. Това заключение е валидно при 1%-ов риск за грешка. В примера същото заключение ще се направи и при равнище на значимост $\alpha = 0{,}05$ (5%-ов риск за грешка), защото теоретичната стойност ($F_{\text{Т}} = 3{,}24$) също е по-голяма от емпиричната.

Необходимо е да се има предвид, че при тази, както и при всички подобни проверки, не трябва да се твърди категорично, че дадения фактор изобщо не влияе върху интересуващия ни резултативен признак. И не само защото заключението е вероятностно, а и защото е възможно в някои случаи да не са спазени изискванията относно провеждането на опити и за филтрирането на евентуалните други фактори, които могат да деформират в известна степен експерименталните данни.

Необходимо е специалистът да е уверен, че са спазени всички изисквания, свързани с формирането на извадките, както и другите условия за коректно приложение на дисперсионния анализ.

7.3. Практикум

7.3.1. Въпроси за самопроверка

1. Каква е познавателната задача на дисперсионния анализ?
2. Какво означава хомоскедастичност?
3. Какво представлява девиацията?
4. Как се намират независимите оценки на дисперсията?
5. Как се определят степените на свобода за вътрешногруповата и междугруповата девиация?

7.3.2. Задачи за упражнение

Задача 1. Предполага се, че мащабите на търговските фирми влияят върху скоростта на оборота на стоковите запаси. Направено е наблюдение на 30 фирми, разделени на три групи-малки, средни и големи според средния месечен оборот, формирани като случайни извадки. За всяка фирма е изчислена продължителността на оборота в дни. Резултатите са представени в следващата таблица.

Таблица 7.4

Продължителност на оборота на групите фирми

Групи фирми	Брой на фирмите в извадката (n_c)	Продължителност на оборота на отделните фирми в дни (x)	Средна продължителност на оборота за извадката ($\bar{x}_i$)
малки	12	15, 14, 11, 18, 12, 16, 12, 14, 15, 13, 12, 18	14,2
средни	10	14, 11, 12, 14, 12, 14, 13, 12, 11, 12	12,5
големи	8	16, 10, 11, 11, 12, 11, 10, 14	11,9
	30		

Необходимо е да се провери чрез дисперсионен анализ хипотезата относно влиянието на мащабите на фирмите върху скоростта на оборота при равнище на значимост $\alpha = 0{,}05$.

Отговор: $F = 3{,}88$

$F_т = 3{,}35$

Задача 2. Ергономи твърдят, че при характера и технологията на производството в дадена фирма, различното оцветяване на стените на работните помещения влияе върху настроението на работниците и по този начин върху тяхната производителност на труда. Направен е експеримент при 4 цвята на стените в 4 работни помещения, в които са работили общо 50 работници, изпълняващи еднакви производствени операции и които преди експеримента са имали еднаква средна производителност. За периода на наблюдението общата девиация на часовата производителност на труда е била $\sum (x - \bar{x})^2 = 271{,}4$, а междугруповата девиация - $\sum (\bar{x}_i - \bar{x})^2 n_i = 10{,}14$.

Необходимо е:

а) да се дефинират нулевата и алтернативната хипотези;

б) да се изчислят двете независими оценки на дисперсията и дисперсионното отношение (F);

в) да се намери теоретичната стойност на F при равнище на значимост $\alpha = 0{,}01$;

г) да се направи заключение относно предполагаемото влияние на цвета на стените върху часовата производителност на труда на работниците.

Отговори: по точки б) и в)

$\hat{\sigma}_1^2 = 3{,}38$; $\hat{\sigma}_2^2 = 5{,}68$; $F = 1{,}68$; $F_т = 4{,}3$.

8. РЕГРЕСИОНЕН И КОРЕЛАЦИОНЕН АНАЛИЗ

..”Научното съждение - това е съждение, опиращо се на знанието на взаимозависимостите, освободено от индивидуалните качества на изследователя.”

К. Пирсън

Тази глава съдържа изключително важна и широко прилагана методология за анализ на зависимости в масовите явления, наречени корелационни. Усвоявайки тази методология, читателят ще придобие знания да дефинира правилно изследователските си задачи при анализа на зависимостите в своята професионална област, да се предпазва от некоректен подход, да избира и правилно да прилага подходящи методи за моделиране и измерване на зависимостите при различните им форми и обхват. Той ще може да специфицира методологичните процедури, когато зависимостите се разглеждат в тяхната динамика или когато анализираната информация е представена на неинтервални скали. Възможностите за прилагане на придобитите знания са извънредно широки, защото в каквато и област да работи, явленията се намират в многостранни и многопосочни връзки и зависимости. Тяхното изследване и конкретно моделиране и измерване задоволява не само всеобщия стремеж към знания, а и практическа необходимост във връзка с управлението на процесите, за вземането на стратегически и оперативни решения.

8.1. Същност и задачи на регресионния и корелационния анализ

Едно от главните направления на статистическия анализ е анализът на зависимостите между факторни и резултативни признаци и между изучаваните явления. Но зависимостите са различни по същество и по форма на проявление. Самото понятие за зависимост се дефинира различно във философията, математиката, физиката и др. Специфично съдържание има и в статистиката.

Преди всичко то означава зависимости, които се проявяват в масовите явления в реално време и пространство. По своята форма те не

са функционални и не се отнасят за всеки отделен случай (единица) на съвкупността.

Само като съвкупност отделните случаи на проявление на масовото явление са подложени на общото влияние на системно действуващи причини, фактори. Зависимостта се съдържа в определена степен във всеки отделен случай, но в случайна форма и тя може да се прояви като такава само в съвкупността, в която се неутрализира действието на случайностите.

Казано по друг начин, на определено изменение на дадено явление-фактор не съответствува точно определено изменение при всички единици на явлението-резултат. Това е съществена особеност на зависимостите, които са обект на статистически анализ, при който те получават числови характеристики.

Може да се дефинира още по-конкретно: тази зависимост, при която на дадено значение на факторния признак са възможни няколко или много различни значения на резултативния признак се нарича ***корелационна зависимост*** (от лат. *correlatio* - съотношение, взаимозависимост)[1]. Тя следователно се различава от функционалната зависимост, при която на всяко значение на независимата променлива съответствуват напълно определени значения на зависимата променлива. Известно е например, че съществува зависимост между потреблението на определени хранителни продукти в домакинствата и доходите им. Но това не означава, че при еднакъв доход и потреблението във всички домакинства е еднакво. Зависимостта се проявява при определена вариация в потреблението и в доходите.

Трябва да се има предвид, че корелационната зависимост не винаги е по характер причинно-следствена. Дали две явления се намират помежду си в отношение на причина и следствие, по какъв начин възникват причините и как те обуславят едни или други следствия и др. са въпроси, на които търсят и дават обяснение науките, изучаващи по същество съответната област от действителността. Може да се установи

[1] Терминът корелация е заимствуван от естествознанието. Той се среща в трудовете на френския изследовател **Хорх Кювие (1769-1832)** като принцип на "корелация на частите на организма". В същия смисъл се използва от **Ч. Дарвин**. В статистическите изследвания е въведен за пръв път от **Ф. Галтон (1822-1911)** при измерване на зависимостта между признаците на родителите и на децата.

корелационна зависимост между две или повече явления, но нейната дълбока причина да се корени в други явления и тя да остане скрита, неуловима при статистическия анализ.

Методите за анализ на корелационните зависимости имат големи познавателни възможности. Но те, както изобщо всички статистически методи, са безразлични към качествената природа на явленията. Затова коректното приложение на тези методи предполага добре да се познават изследваните явления. В противен случай може да се стигне до абсурдни заключения. Във връзка с това се използва терминът ***лъжекорелация***. Това е привидна зависимост, каквато в действителност не съществува. Лъжлива представа за зависимост се получава например при паралелно изменение на две явления, които са независими помежду си.

Когато се изучават корелационни зависимости, обикновено се поставят две основни познавателни задачи: 1) Да се моделира формата на зависимостта и чрез съставения модел се характеризира влиянието на явлението (явленията)-фактор върху явлението-резултат. Тази задача се решава чрез ***регресионния анализ***[1] 2) Да се измери теснотата на зависимостта (взаимозависимостта) между интересуващите ни явления - тази задача се решава посредством ***корелационния анализ***. Това са две страни на анализа на зависимостите, които имат общи теоретико-методологични основи.

Когато се изследва зависимост между едно явление-резултат и едно явление-фактор, говори се за ***единична регресия и корелация*** или за ***еднофакторен регресионен и корелационен анализ***, а когато се обхващат две или повече явления-фактори - за ***множествена регресия и корелация*** или за ***многофакторен регресионен и корелационен анализ***.

Според графичния образ, който приема, корелационната зависимост е ***линейна*** (праволинейна) и ***нелинейна*** (криволинейна). Това разграничение е важно, тъй като при двата случая има особености в моделирането и измерването на зависимостите.

Класическата теория на регресионния и корелационния анализ е разработена при две съществени предпоставки.

[1] Терминът регресия (от лат. *regressio* - движение назад) е преминал в статистиката от биологията. **Галтон** го е употребил в по-тесен смисъл при статистическото изучаване на наследствеността, но в статистическата литература се е наложил с по-широко съдържание.

Първо. Факторните и резултативните признаци да имат числови характеристики, т.е. да са представени на интервални скали. С това се осигурява възможност да се прилагат прецизни параметрични методи за моделиране и измерване на теснотата на зависимостите.

Второ. Зависимостите да се измерват въз основа на разпределения в статика, т.е. за определен момент или период при независимост между значенията на факторните и съответно на резултативните признаци.

Практически обаче възниква необходимост от изследване на зависимости при неинтервални скали, както и в динамика, т.е. при динамични редове. В такива случаи класическите методи се модифицират в известна степен или се прилагат специфични методи.

В следващото изложение ще бъде направено кратко въведение в класическата теория на регресионния и корелационния анализ, без да се изчерпва цялата методология в тази област. Отделно се разглеждат особеностите на анализа на зависимостите при неинтервални скали и при динамични статистически редове.

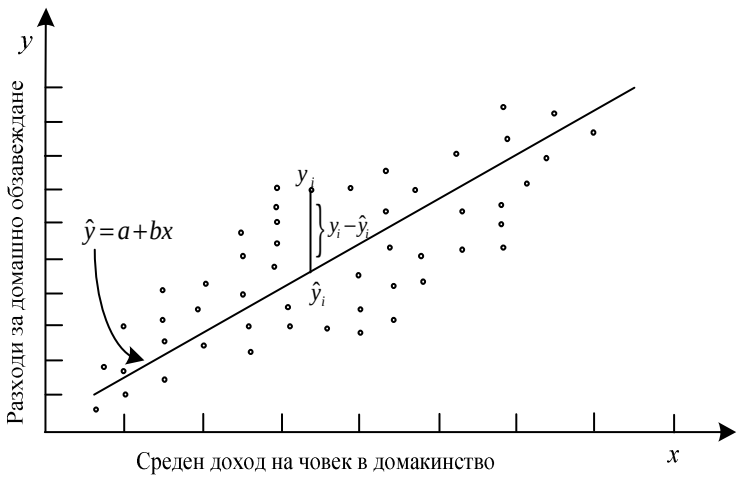
8.2. Еднофакторен регресионен и корелационен анализ

При емпиричните изследвания често се изследва зависимостта между два признака, от които единият е резултативен, а другият - факторен. Аналогична е постановката и когато двата признака са взаимозависими. Не се държи сметка за други признаци, които могат да се намират в зависимост с изследваните. В тези случаи анализът е ***еднофакторен***, или се казва още единична регресия и корелация.

8.2.1. Линейна (праволинейна) регресия и корелация

Ще приемем, че зависимостта на един резултативен признак от един факторен признак по форма е праволинейна. Дадени са например данни по наблюдавани домакинства за средния доход на член от домакинството и разходите за домашно обзавеждане през 2007 година. Най-обща визуална представа за зависимостта на разходите от доходите може да се получи от графичното представяне на данните. Съставя се точкова диаграма, като по скала на абцисната ос се отчита доходът (x), а

по ординатната ос-разходите за домашно обзавеждане (*y*). Всяко домакинство заема по двата признака положението на точка. Така се получава множество точки, разположени в определено поле, наречено ***корелационно поле***. В случая то има формата на елипса и затова може да се нарече ***корелационна елипса*** (фиг. 8.1.).



Фиг. 8.1

Макар, че са разсеяни в корелационното поле, все пак точките не са разпилени безразборно, а са разположени около въображаема възходяща линия. Ако зависимостта би била функционална, всички точки биха лежали само на тази линия. Но тъй като е корелационна, точките са разположени около линията. В случая тя е права, следователно зависимостта е линейна (праволинейна). Но тази линия не може да бъде само въображаема. Необходимо е да се намери нейното точно положение според емпиричните данни.

На диаграмата има много точки и през всяка точка може да се прекара линия. Необходимо е да се намери една обща линия, която да бъде графичен модел на зависимостта. Тя трябва да минава между всички точки така, че сумата от квадратите на разликите между емпиричните стойности на y и техните оценки ($\hat{y}$), които се намират на линията, да е минимум, т.е. $\sum (y - \hat{y})^2 = minimum$. Това означава, че тази линия,

наречена ***регресионна линия,*** трябва да се намери по ***метода на най-малките квадрати***. Но това е геометрична интерпретация. Регресионната линия е графичният образ на зависимостта. За да се намери тя, трябва да се намери аналитичния модел, т.е. функцията на регресионната линия. Уравнението, което описва зависимостта се нарича ***регресионно уравнение***. Това по същество е ***линеен регресионен модел***.

За да се получи регресионния модел при линейна зависимост се тръгва от линейната функция (уравнение на права) - $y = a + bx$. В геометричната интерпретация a е точката, в която правата пресича ординатната ос, а b е ъглов коефициент, изразяващ ъгъла, който правата сключва с абсцисната ос. За да се намерят тези параметри, може да се състави система от две нормални уравнения.

$$(8.1) \qquad \begin{vmatrix} \sum y = Na + b\sum x \\ \sum xy = a\sum x + b\sum x^2 \end{vmatrix}$$

Ако двете уравнения се разделят на N и се направят възможните съкращения, ще се получат две удобни формули за намиране на a и b:

$$(8.2) \qquad b = \frac{\sum xy - N\,\bar{x}\,\bar{y}}{\sum x^2 - N\,\bar{x}^2}; \quad a = \bar{y} - b\bar{x}.$$

След като са намерени a и b от изходното уравнение $y = a + bx$, съставя се регресионният модел:

$$(8.3) \qquad \hat{y} = a + bx \ .$$

В този модел a се нарича свободен член и не се интерпретира съдържателно, а b се нарича ***регресионен коефициент***. Той показва с колко единици се изменя резултативния признак, при изменение на факторния признак с единица (според приетата мярка). В посочения условен пример регресионният коефициент би показал с колко лв. средно се увеличават разходите на домакинствата за домашно обзавеждане, при увеличаване на доходите средно на член от домакинствата примерно с 1000 лв. за година.

Тъй като регресионният коефициент b измерва изменението на y при изменение на x с една единица, той може да се означи с $b_{y/x}$. Възможно е да се състави регресионно уравнение като се разменят

местата на *x* и *y*, и то ще съдържа регресионен коефициент за *x* по отношение за *y*. Той ще показва колко единици изменение на *x* съответстват на изменение на *y* с единица (според приетата мярка) и може да се означи с $b_{x/y}$. (По-нататък се разглежда връзката между $b_{y/x}$ и $b_{x/y}$ и между тях и коефициентите на корелацията).

Чрез регресионното уравнение $\hat{y} = a + bx$, могат да се намерят оценките $\hat{y}$ за всяка стойност на *x* при известни вече *a* и *b*:

$$\hat{y}_1 = a + bx_1, \quad \hat{y}_2 = a + bx_2, \quad \hat{y}_3 = a + bx_3 \quad \text{и т.н.}$$

При това трябва да се получи $\sum y = \sum \hat{y}$.

Тези оценки показват какви биха били значенията на резултативния признак, при положение, че зависимостта се проявява еднакво при всички единици. Разликите между фактическите значения (y) и изчислените ($\hat{y}$) чрез регресионното уравнение ($y - \hat{y}$) са конкретните грешки на оценките при всяка единица. Общата ***стандартна грешка на оценката*** (S_y) може да се изчисли като средна квадратична величина от тези разлики ($S_y = \sqrt{\frac{\sum (y - \hat{y})^2}{N}}$).

Регресионният модел и съдържащият се в него регресионен коефициент осигуряват много полезна информация за формата на зависимостта и за изменението на резултативния признак под влияние на факторния. При определени условия регресионният модел може да се използва за прогнозиране на възможното изменение на резултативния признак при зададен бъдещ размер на факторния признак.

Регресионният модел и регресионният коефициент не отговарят на въпроса колко силна е корелационната зависимост. Отговор на този въпрос дава ***корелационният анализ***.

Между значенията на резултативния признак има вариация. Тя се дължи на различни фактори, може да се нарече ***обща вариация*** и да се измери с дисперсията σ_y^2. Ако интересуващият ни факторен признак наистина влияе в някаква степен върху резултативния, той предизвиква част от общата вариация. Това ще се изрази във вариацията на оценките на резултативния признак ($\hat{y}$), получен чрез регресионен модел. Тази вариация се нарича ***факторна вариация*** или обяснена вариация. Тя може

да се измери с дисперсията $\sigma_{\hat{y}}^2$. Вариацията, която се дължи на други фактори, се нарича ***остатъчна вариация*** и може да се измери с S_y^2. По такъв начин общата вариация (общата дисперсия) се разлага на две части - факторна и остатъчна:

(8.4) $$\sigma_y^2 = \sigma_{\hat{y}}^2 + S_y^2 \ .$$

Колкото по-силна е зависимостта на резултативния признак от факторния, толкова по-голям ще е делът на факторната вариация в общата. Затова от уравнение (8.4) $\sigma_{\hat{y}}^2$ трябва да се пренесе вляво на равенството и двете му части да се разделят на σ_y^2 :

(8.5) $$\frac{\sigma_{\hat{y}}^2}{\sigma_y^2} = \frac{\sigma_y^2}{\sigma_y^2} - \frac{S_y^2}{\sigma_y^2} .$$

Ако $\frac{\sigma_{\hat{y}}^2}{\sigma_y^2}$ се означи с r^2, ще се получи:

(8.6) $$r^2 = \frac{\sigma_{\hat{y}}^2}{\sigma_y^2} = \frac{\sigma_y^2 - S_y^2}{\sigma_y^2} = 1 - \frac{S_y^2}{\sigma_y^2} \ .$$

Полученият коефициент r^2 се нарича ***коефициент на детерминацията*** и показва каква част (или колко процента) от вариацията на резултативния признак се обуславя от вариацията на факторния признак.

Разликата $1 - r^2$ (или $100 - r^2$) се нарича ***коефициент на индетерминацията*** и показва каква част (или колко процента) от вариацията на резултативния признак се дължи на други фактори.

Положителният квадратен корен на r^2 се нарича ***коефициент на корелацията*** (известен още като корелационен коефициент на ***Карл Пирсън***):

(8.7) $$r = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}} ,$$

където $S_y^2 = \frac{\sum(y-\hat{y})^2}{N}$ и $\sigma_y^2 = \frac{\sum(y-\bar{y})^2}{N}$. (Ако анализът се прави по информация, получена от извадки, оценките на S_y^2 и σ_y^2 трябва да се изчислят по формулите $S_y^2 = \frac{\sum(y-\hat{y})^2}{n-2}$ и $\sigma_y^2 = \frac{\sum(y-\bar{y})^2}{n-1}$).

В дадения случай, когато зависимостта е праволинейна, r се нарича ***коефициент на линейна корелация***.

Корелационният коефициент като измерител на теснотата на зависимостта може да приема стойности от 0 до 1, т.е. $0 \le r \le 1$. Колкото зависимостта е по-силна, толкова r е по-близък до 1. За по-определеното му тълкуване относно теснотата на зависимостта, често се използва условна скала: когато е до 0,3, корелационната зависимост се оценява като слаба, над 0,3 до 0,5 - умерена, над 0,5 до 0,7 - значителна, над 0,7 до 0,9 - силна и над 0,9 - много силна.

Трябва да се има предвид, че корелационният коефициент, изчислен по формула 8.7, е винаги положителна величина. В действителност зависимостта може да бъде положителна (позитивна), когато с увеличаване на значенията на факторния признак (x) се увеличават значенията на резултативния признак (y), и отрицателна (негативна) - когато при увеличаване на значенията на факторния признак значенията на резултативния намаляват. Дали зависимостта е положителна, или отрицателна, т.е дали пред r трябва да се запише положителен или отрицателен знак, се съди по алгебричния знак на регресионния коефициент b.

Между коефициента на линейната корелация и регресионните коефициенти има определена връзка. Тя може да се изрази аналитично и да се изведат някои правила, които имат значение при емпиричните изследвания.

1. Може да се докаже, че:

(8.8) $b_{y/x} = r\frac{\sigma_y}{\sigma_x}$; (8.9) $b_{x/y} = r\frac{\sigma_x}{\sigma_y}$,

където σ_x е средноквадратично (стандартното) отклонение за x, а σ_y - средноквадратично (стандартното) отклонение за y.

2. От формули 8.8 и 8.9 следва, че:

(8.10) $r = b_{y/x}\frac{\sigma_x}{\sigma_y}$; (8.11) $r = b_{x/y}\frac{\sigma_y}{\sigma_x}$.

3. Щом $b_{y/x} = r\frac{\sigma_y}{\sigma_x}$ и $b_{x/y} = r\frac{\sigma_x}{\sigma_y}$, тогава $b_{y/x}.b_{x/y} = r^2$.

Следователно:

(8.12) $r = \sqrt{b_{y/x}.b_{x/y}}$.

4. Беше установено, че $a = \bar{y} - b\bar{x}$. Ако в регресионното уравнение $\hat{y} = a + bx$ се замести a с този израз, като същевременно b се замести с $r\frac{\sigma_y}{\sigma_x}$, ще се получи друг израз на регресионното уравнение:

(8.13) $\hat{y} = \left(\bar{y} - r\frac{\sigma_y}{\sigma_x}\right) + r\frac{\sigma_y}{\sigma_x}\bar{x}$.

След съответни съкращения се получава:

(8.14) $\hat{y} = \bar{y} + r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$,

или още

(8.15) $\hat{y} = \bar{y} + b(x - \bar{x})$.

Тази модификация на регресионното уравнение предлага някои практически удобства при анализа.

Ще илюстрираме изложения начин за съставяне на регресионното уравнение и за изчисляване на корелационния коефициент и на коефициента на детерминацията с ***пример***. Да приемем, че са наблюдавани 10 работници, произвеждащи еднакви изделия. Установени са трудовият им стаж и средната производителност през наблюдавания период. Изходните данни и изчисленията са дадени в табл. 8.1. (Данните са условни и се отнасят за малък брой случаи. С примера се илюстрират само изчислителните процедури и тълкуването на резултатите.)

Таблица 8.1.

Зависимост между трудовия стаж на работниците и тяхната производителност на труда във фирма "Н" през м. април 2008 г.

Пор. No на работ-ника	Трудов стаж в навър-шени години (x)	Средна дневна произво-дител-ност на труда, бр. (y)	xy	x^2	$\hat{y}$	$(y-\hat{y})$	$(y-\hat{y})^2$	$(y-\bar{y})$	$(y-\bar{y})^2$
1	8	180	1440	64	190	-10	100	20	400
2	2	135	270	4	130	5	25	25	625
3	10	240	2400	100	210	30	900	80	6400
4	6	180	1080	36	170	10	100	20	400
5	1	110	110	1	120	-10	100	-50	2500
6	4	142	568	16	150	-8	64	-18	324
7	2	165	330	4	130	35	1225	5	25
8	7	128	896	49	180	-52	2704	-32	1024
9	1	115	115	1	120	-5	25	-45	2025
10	9	205	1845	81	200	5	25	45	2025
	50	1600	9054	356	1600	0	5268	0	15748

По данните от таблицата се получават:

$$\bar{x} = \frac{\sum x}{N} = \frac{50}{10} = 5 \text{ год.}; \qquad \bar{y} = \frac{\sum y}{N} = \frac{1600}{10} = 160 \text{ бр.};$$

$$b = \frac{\sum xy - N\,\bar{x}\,\bar{y}}{\sum x^2 - N\,\bar{x}^2} = \frac{9054 - 10.5.160}{356 - 10.5^2} = 9{,}94 \text{ (кръгло 10 бр.);}$$

$$a = \bar{y} - b\bar{x} = 160 - 9{,}94.5 = 110{,}3 \text{ бр. (кръгло 110 бр.).}$$

Регресионното уравнение (линейният регресионен модел) е:

$$\hat{y} = 110 + 10x\,.$$

Регресионният коефициент b = 10 бр. показва, че с увеличаване на трудовия стаж средно с 1 година дневната производителност на труда се увеличава средно с 10 бр.

Като се заместят последователно съответните стойности на x в регресионното уравнение, се получават оценките на $\hat{y}$:

$\hat{y}_1 = 110 + 10.8 = 190$ бр.

$\hat{y}_2 = 110 + 10.2 = 130$ бр. и т.н.

Ако изчисленията са извършени правилно и точно, $\sum \hat{y} = \sum y$. В случая това условие е изпълнено.

За да се изчисли корелационният коефициент, трябва да се изчислят предварително S_y^2 и σ_y^2:

$$\sigma_y^2 = \frac{\sum (y - \bar{y})^2}{N} = \frac{15748}{10} = 1574{,}80;$$

$$S_y^2 = \frac{\sum (y - \hat{y})^2}{N} = \frac{5268}{10} = 526{,}80;$$

$$r = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}} = \sqrt{1 - \frac{526{,}80}{1574{,}80}} = \sqrt{0{,}67} = 0{,}82.$$

Тъй като регресионният коефициент b е положителна величина, корелационният коефициент също е с положителен знак. Този корелационен коефициент показва силна зависимост на производителността на труда от трудовия стаж на работниците.

Коефициентът на детерминацията е $r^2 = 0{,}67$ или 67 %. Това означава, че 67 на сто от вариацията (различията) в производителността на труда на наблюдаваните работници може да се обясни с вариацията (различията) в продължителността на трудовия им стаж в предприятието. Останалите 33 на сто (коефициент на индетерминацията) се дължат на други фактори.

Коефициентът на единичната линейна корелация може да се изчисли и по друг метод, без да се съставя предварително регресионното уравнение. Това е ***методът на нормираните (стандартизираните) отклонения***, наричан още метод на ***Аугуст Браве (1811-1863)***[1].

Известно е, че когато разликите между значенията на признака и средната аритметична се разделят на средното квадратично (стандартно)

[1] На името на френския учен **А. Браве**, на когото принадлежи идеята за такъв коефициент, доразвита от **Б. Шепард**, **Ф. Галтон** и **К. Пирсън**.

отклонение, получават се нормирани (стандартизирани) отклонения. В случая те са

$$z_x = \frac{x - \bar{x}}{\sigma_x} \qquad \text{и} \qquad z_y = \frac{y - \bar{y}}{\sigma_y}.$$

Коефициентът на линейната корелация е средна аритметична величина от произведенията на нормираните отклонения:

$$(8.16) \qquad r = \frac{\sum z_x z_y}{N}.$$

Чрез елементарна преработка могат да се получат модификации на тази формула, които при емпиричния анализ предлагат удобства. Ако например z_x и z_y се заместят с $\frac{x - \bar{x}}{\sigma_x}$ и $\frac{y - \bar{y}}{\sigma_y}$, тогава

$$(8.17) \qquad r = \frac{\sum\left[\left(\frac{x - \bar{x}}{\sigma_x}\right).\left(\frac{y - \bar{y}}{\sigma_y}\right)\right]}{N} \quad \text{или, още:}$$

$$(8.18) \qquad r = \frac{\frac{\sum (x - \bar{x})(y - \bar{y})}{N}}{\sigma_x . \sigma_y}.$$

Числителят на формула 8.18 е средна аритметична от произведенията на разликите и се нарича ***ковариация*** или ***момент на произведенията*** (по-точно - първи смесен момент). Затова и изчисляването на r по тази формула понякога се нарича метод на момента на произведенията.

Формулата на *r* може да се запише и така:

$$(8.19) \qquad r = \frac{\frac{\sum xy}{N} - \bar{x}\,\bar{y}}{\sigma_x . \sigma_y}.$$

След съответна преработка може да се получат други, много удобни за практическо използване формули:

(8.20) $$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}};$$

(8.21) $$r = \frac{\sum xy - N\,\bar{x}\,\bar{y}}{\sqrt{\left(\sum x^2 - N\,\bar{x}^2\right)\left(\sum y^2 - N\,\bar{y}^2\right)}};$$

(8.22) $$r = \frac{N\sum xy - \sum x \sum y}{\sqrt{\left[N\sum x^2 - \left(\sum x\right)^2\right]\left[N\sum y^2 - \left(\sum y\right)^2\right]}}.$$

По данните от таблица 8.2. корелационният коефициент по метода на нормираните отклонения (на Браве), изчислен чрез формула 8.20 е:

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}} = \frac{1054}{\sqrt{106.15748}} = 0{,}82\,.$$

Вижда се, че се получава същият резултат, както по предходния метод. Трябва обаче да се има предвид, че методът на нормираните отклонения е приложим само при линейна (праволинейна) корелация.

По данните от табл. 8.2 могат да се изчислят двете стандартни отклонения:

$$\sigma_x = \frac{106}{10} = 3{,}256; \qquad \sigma_y = \frac{15748}{10} = 39{,}68\,.$$

Като се имат предвид формули 8.8 и 8.15, може да се изчисли регресионният коефициент и да се състави регресионното уравнение.

$$b_{y/x} = r\frac{\sigma_y}{\sigma_x} = 0{,}82\ \frac{39{,}68}{3{,}256} = 10 \text{ (кръгло);}$$

$$\hat{y} = \bar{y} + b(x-\bar{x});\ \ \hat{y} = 160 + 10(x-5).$$

Оттук следват:

$$\hat{y} = 160 + 10(8-5) = 160 + 10.3 = 190;$$

$$\hat{y} = 160 + 10(2-5) = 160 + 10.(-3) = 130 \text{ и т.н.}$$

Таблица 8.2

Изчисляване на коефициента на линейната корелация по метода на нормираните отклонения

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
8	180	3	20	60	9	400
2	135	-3	-25	75	9	625
10	240	5	80	400	25	6400
6	180	1	20	20	1	400
1	110	4	-50	200	16	2500
4	142	-1	-18	18	1	324
2	165	-3	15	-15	9	25
7	128	2	-32	-64	4	1024
1	115	-4	-45	180	16	2025
9	205	4	45	180	16	2025
50	1600	0	0	1054	106	15748

Изложените начини за съставяне на регресионното уравнение и за изчисляване на регресионния коефициент са приложими в този вид при негрупирани данни, т.е когато са дадени отделните единици със съответните им значения на двата признака. Когато единиците са много и предварително са групирани по два признака, получава се двумерно разпределение със съответни честоти. То се представя в шахматна по форма таблица, наричана ***корелационна таблица***.(вж. табл. 8.3). Графично може да се представи чрез стереограма, подобна на хистограмата, но с три измерения. Може да се представи с повърхност на разпределението, подобна на полигона на разпределението, но също в стереометрична форма с три измерения.

При такива двумерни разпределения описаните методи по принцип остават в сила, но формулите и изчислителните процедури се модифицират, като се взема под внимание различният брой на единиците (честотите *f*) за комбинираните значения на признаците.

Нормалните уравнения за изчисляване параметрите на регресионното уравнение приемат вида:

$$(8.23) \qquad \begin{cases} \sum yf = a\sum f + b\sum xf \\ \sum yxf = a\sum xf + b\sum x^2 f \end{cases}$$

При такива двумерни разпределения теснотата на зависимостта може да се измери чрез ***корелационното отношение*** (емпиричното корелационно отношение), което по смисъл не се различава от корелационния коефициент. Основава се на съотношението между дисперсията на средните значения на резултативния признак по отделните значения на факторния признак ($\sigma^2_{\bar{y}_x}$) и общата дисперсия на резултативния признак (σ^2_y). Означава се с гръцката буква η (ета):

$$(8.24) \qquad \eta = \sqrt{\frac{\sigma^2_{\bar{y}_x}}{\sigma^2_y}}, \quad \text{където}$$

$$\sigma^2_{\bar{y}_x} = \frac{\sum (\bar{y}_i - \bar{y})^2 f}{\sum f}.$$

Дисперсията на средните на *y* по групи, обособени по значенията на *x*, е част от общата дисперсия на *y* и се обуславя от действието на факторния признак *x*. Колкото зависимостта е по-силна, толкова нейният дял в общата дисперсия на *y* е по-голям.

Ще илюстрираме изчисляването на корелационното отношение с пример, като видоизменяме примера от табл. 8.1.

Таблица 8.3

Разпределение на работниците на фирма "Н" по трудов стаж и по средна дневна производителност на труда през април 2008 година

Групови интервали по производителност на труда, бр.(y)	Групови интервали по трудов стаж, години (x)									Брой на работниците (f)
	0-2	3-5	6-8	9-11	12-14	15-17	18-20	21-23	24-26	
20-24	1	2	1							4
25-29		1	5	3						9
30-34		2	3	7	3	1				16
35-39			3	4	8					18
40-44			1	2	7	8	2			20
45-49				2	2	5	4	2		15
50-54					1	1	5	3		10
55-59						1	2	2		5
60-64							1	1	1	3
Брой на работниците	1	5	13	18	21	19	14	8	1	100
$\bar{y}_x$	22,00	27,00	31,23	35,06	39,62	43,32	50,57	53,25	62,00	40,45

Средните аритметични $(\bar{y}_x)$ са изчислени от средите на интервалите на y и честотите по колони на таблицата (за всяка среда на интервала за x). Посочени са в последния ред на таблицата. Общата средна аритметична на y е $\bar{y}$ = 40,45 бр. От средите на интервалите на y и общата средна ($\bar{y}$) е изчислена общата дисперсия - $\sigma_y^2 = 91{,}35$.

Изчисляването на $\sigma_{\bar{y}_x}^2$ е показано в табл. 8.4

Таблица 8.4

$\bar{y}_x$	f	$\bar{y}_x - \bar{y}$ $(\bar{y} = 40{,}45)$	$(\bar{y}_x - \bar{y})^2$	$(\bar{y}_x - \bar{y})^2 f$
22,00	1	-18,45	340,40	340,40
27,00	5	-13,45	180,90	904,51
31,23	13	-9,22	85,00	1105,11
35,06	18	-5,39	29,05	522,94
39,62	21	-0,83	0,69	14,47
43,32	19	2,87	8,24	156,50
50,57	14	10,12	102,41	1433,80
53,25	8	12,80	163,84	1310,72
62,00	1	21,55	464,40	464,40
	100			6252,85

$$\sigma^2_{\bar{y}_x} = \frac{6252{,}85}{100} = 62{,}53; \qquad \eta = \sqrt{\frac{62{,}53}{91{,}35}} = \sqrt{0{,}6845} = 0{,}83.$$

Полученото корелационно отношение показва силна зависимост на производителността на труда от трудовия стаж на работниците. Коефициентът на детерминацията е 0,68 или 68 % и показва, че 68 на сто от вариацията на средната дневна производителност на труда се дължи на вариацията (различията) в трудовия стаж на работниците.

Формулите на коефициента на линейната корелация по метода на нормираните отклонения (на Браве) също се допълва с честотите. Твърде удобна и често употребявана формула е:

$$(8.25) \qquad r = \frac{\frac{\sum xyf}{\sum f} - \bar{x}\,\bar{y}}{\sigma_x \sigma_y},$$

където: $\sigma_x = \sqrt{\frac{\sum (x - \bar{x})^2 f}{\sum f}}$ и $\sigma_y = \sqrt{\frac{\sum (y - \bar{y})^2 f}{\sum f}}$.

Когато двумерното разпределение е представено в корелационна таблица (като табл. 8.3), формула 8.25 може да се модифицира, като се има предвид едно свойство на корелационния коефициент. То гласи: ако от всички значения на двата признака се извади някакво постоянно число

и получените разлики се разделят на постоянно число, коефициентът на корелацията не се изменя. Това създава възможност да се приложи значително съкратена изчислителна процедура, известна като ***метод на четирите полета***, или още като ***таблична корелация***.[1]

8.2.2. Нелинейна (криволинейна) регресия и корелация

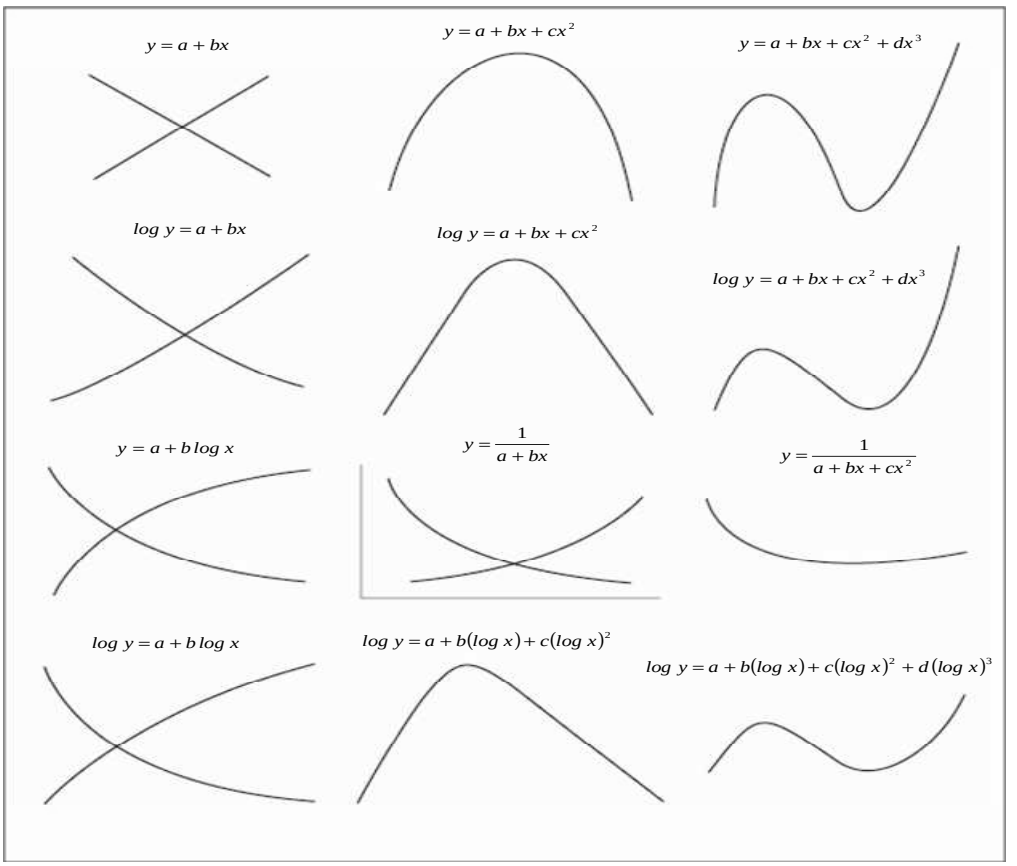
В много случаи графичният образ на зависимостта между интересуващите ни признаци не е права, а някаква крива линия. Тогава тя не може да се моделира с линейно регресионно уравнение. Ако в такива случаи се изчислят линейни корелационни коефициенти, те не биха измервали вярно теснотата на зависимостта и могат да доведат до неверни заключения.

Изложените общи логически и методологични положения относно моделирането и измерването на линейната корелационна зависимост са валидни и при нелинейна (криволинейна) зависимост. За да се състави обаче такова регресионно уравнение, което да е адекватен модел на емпиричната зависимост, трябва да се избере съответната подходяща функция. Тук изследователят трябва да бъде особено внимателен. Той трябва да изучи предварително характера и формата на зависимостта и да избере правилно подходящата функция (предполага се, че наличието на зависимост не подлежи на съмнение или е предварително доказано). Добра ориентация в това отношение може да даде ***диаграмата на емпиричните данни***. Тя може да покаже как приблизително изглежда кривата, описваща зависимостта. На фиг. 8.2 са показани някои ***типове криви и математическите им функции***.[2] Когато не може категорично да се определи само една крива като най-подходяща и трябва да се избира между две или повече криви, след като всяка от тях е изпитана, може като критерий да се използва стандартната грешка на оценката. Като най-подходяща може да се приеме тази функция, при която се получава най-малка стандартна грешка на оценката. Има и по-прецизни методи за

[1] Вж. **Гатев, К.** Въведение в статистиката. С., 1995, с. 225 и сл.

[2] Заимствано от **Езекиэл, М**., и **К. Фокс**, Методы анализа корреляции и регресий, М., 1966.

проверка на адекватността на моделите, чието разглеждане не е възможно в рамките на тази книга.[1]



Фиг. 8.2

Когато е избрана съответната функция, намирането на ***параметрите на регресионното уравнение*** става по метода на най-малките квадрати, както при линейна регресия. Ако например зависимостта се описва с парабола от втора степен, изхожда се от нейното уравнение $y = a + bx + cx^2$, като се съставя системата от нормални уравнения:

[1] Вж. **Съйкова, И.**, Статистически изследвания на зависимости и други връзки в социално-икономическата област, С., 1991; **Съйкова, И., А. Стойкова – Къналиева, С. Съйкова**, Статистическо изследване на зависимости, С., 2002.

$$(8.26)\qquad \begin{vmatrix} \sum y = Na + b\sum x + c\sum x^2 \\ \sum xy = a\sum x + b\sum x^2 + c\sum x^3 \\ \sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4 \end{vmatrix}.$$

Изчисляването на параметрите може да се облекчи, ако вместо x се използват отклоненията на x от тяхната средна аритметична - $(x-\bar{x})$. Тогава могат да се направят съкращения, тъй като $\sum(x-\bar{x})=0$ и $\sum(x-\bar{x})^3=0$ и системата ще приеме вида:

$$(8.27)\qquad \begin{vmatrix} \sum y = Na + c\sum(x-\bar{x})^2 \\ \sum y(x-\bar{x}) = b\sum(x-\bar{x})^2 \\ \sum y(x-\bar{x})^2 = a\sum(x-\bar{x})^2 + b\sum x^3 + c\sum(x-\bar{x})^4 \end{vmatrix}.$$

Във второ уравнение се съдържа само неизвестната b и тя ще се изрази с формулата:

$$(8.28)\qquad b = \frac{\sum y(x-\bar{x})}{\sum(x-\bar{x})^2}.$$

Чрез решаване на системата от останалите две уравнения се намират a и c.

При редица емпирични изследвания, в т.ч. в икономическата и социалната област, се оказват подходящи функциите

$y = a + bx + cx^2 + dx^3$; $\quad log\, y = a + bx$; $\quad y = a + b\, log\, x$;

$log\, y = a + b\, log\, x$; $\quad y = ab^x$; $\quad y = ax^b$;

$y = \frac{1}{a+bx}$; $\quad y = \frac{1}{a+bx+cx^2}$; $\quad$ и др.

Съществуват възможности нелинейните функции да се трансформират в линейни. Например функцията $y = ab^x$ може чрез логаритмуване да се трансформира в $log\, y = log\, a + x\, log\, b$. Това улеснява изчисляването на параметрите. Освен това чрез трансформацията на нелинейните функции в линейни може по емпиричен път да се получи ориентация относно формата на зависимостта.

За измерване на теснотата на нелинейната корелационна зависимост служи корелационен коефициент, аналогичен на коефициента на линейната корелация, който в случая често се нарича ***индекс на корелацията***, чиято формула е:

$$r = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}} \,. \qquad (8.29)$$

Както се вижда, тази формула не се различава по външен вид от познатата вече формула на ***К. Пирсън*** за изчисляване на коефициента на линейната корелация. По същество разликата се състои в това, че S_y^2 се изчислява при значения на $\hat{y}$, намерени чрез съответен вид нелинейно регресионно уравнение. Измерването на теснотата на нелинейната корелация може да стане и чрез корелационното отношение (теоретичното корелационно отношение):

$$\eta = \sqrt{\frac{\sigma_{\hat{y}}^2}{\sigma_y^2}}, \quad \text{където} \quad \sigma_{\hat{y}}^2 = \frac{\sum (\hat{y} - \bar{y})^2}{N} \,. \qquad (8.30)$$

Трябва да се има предвид, че докато при линейната корелация коефициентът не се изменя, ако се разменят местата на *x* и *y*, при нелинейната корелация индексът на корелацията ще се измени при такава размяна. Всъщност при нелинейната зависимост могат да се изчислят ***два индекса на корелацията*** - за *y* по отношение на *x* и за *x* по отношение на *y*. Затова винаги трябва предварително да се определя положението на двата признака в регресионното уравнение. Без това условие индексът на корелацията няма достатъчно ясен смисъл.

8.3. Многофакторен регресионен и корелационен анализ

При изучаването на много явления не можем да се задоволим с характеризиране само на зависимостта между два признака (променливи). Интересът често е насочен към общото влияние на два или повече признака върху даден резултативен признак или пък към всеки от избраните факторни признаци при условно елиминиране на останалите.

Преди да се пристъпи към моделиране и измерване на зависимостите, при многофакторния анализ трябва предварително да се решат два въпроса.

Първият въпрос се отнася за избора и броя на факторите (факторните признаци). Кои факторни признаци да бъдат включени в анализа, решава специалиста, пристъпващ към анализ. Познавайки добре същността на изследваните явления, естествено е той да подбере важните, съществените фактори, от които в най-голяма степен зависи съществуването и развитието на явлението-резултат (резултативният признак). По отношение на броя на факторните признаци теорията няма категорично предписание и еднозначно решение. Многогодишният опит обаче подсказва, че не е целесъобразно да се включват много факторни признаци едновременно в един модел. И не защото не е възможно да се състави и реши моделът, а защото се затруднява съдържателната интерпретация на получените резултати. Ако има интерес и към други признаци, те могат да се включат в друг модел.

Ако анализът се прави въз основа на извадка, броят на факторните признаци трябва да се определя съобразно обема на извадката. Ако при даден обем на извадката броят на факторните признаци в модела е твърде голям, стандартната грешка може да бъде толкова голяма, че да се обезсмисли заключението. Затова обикновено се прилага правилото обемът на извадката да е поне 8 пъти по-голям от броя на включените в модела факторни признаци.

Вторият въпрос е свързан с понятието ***мултиколинеарност***. Съгласно теорията на регресионния и корелационния анализ факторните признаци трябва да са независими помежду си. Когато има зависимост между два факторни признака, казва се, че има ***колинеарност***, а когато има такава зависимост между повече признаци - ***мултиколинеарност***. Включването в анализа (в един модел) на зависими помежду си факторни признаци може силно да деформира резултатите от анализа. От това следва, че не би трябвало да се включват едновременно взаимно зависими факторни признаци. Когато наличието на такава зависимост не е очевидно, може предварително да се провери, като се изчислят коефициенти на единичната корелация по двойки факторни признаци, преди те да се включат в модела. Тези коефициенти ще покажат дали има

зависимост (колинеарност и мултиколинеарност) и от какъв порядък е тя. Трябва обаче да се има предвид, че в много случаи могат да се получат всички или почти всички коефициенти различни от нула. И това би направило невъзможен анализа, ако се прилага строго и безусловно правилото да не се включват факторни признаци, между които има каквато и да е степен на зависимост. Затова се допуска разумен компромис. Установено е, че зависимостта между факторните признаци оказва съществено влияние (деформация) върху резултатите от анализа, когато корелационните коефициенти, измерващи тази зависимост са от порядъка 0,8 и повече. Когато се получат такива коефициенти, някои от факторните признаци следва да отпаднат.

След като са решени въпросите относно факторните признаци, анализът може да протече в двата му аспекта-регресионен и корелационен.

При многофакторния регресионен и корелационен анализ е необходимо да се внесе малко изменение в означенията. За удобство по-нататък ще приемем, че значенията на резултативния признак (зависимата променлива) са x_1, а на факторните признаци (независимите променливи) - $x_2, x_3, x_4, ..., x_k$.

Ако x_1 се намират в линейна зависимост с $x_2, x_3, x_4, ..., x_k$, за намиране на уравнението на множествената регресия се изхожда от следната функция (линеен полином):

$$x_1 = a + b_2 x_2 + b_3 x_3 + ... + b_k x_k . \qquad (8.31)$$

Коефициентите b изразяват зависимостта на x_1 от съответните факторни признаци, при положение, че влиянието на останалите условно е елиминирано. За да се покаже за кои фактори се отнасят те и кои са елиминирани, се записват: $b_{12.34...k}$, $b_{13.24...k}$, $b_{1k.23...(k-1)}$. В такъв случай регресионният модел ще приеме вида:

$$\hat{x}_1 = a + b_{12.34....k} x_2 + b_{13.24....k} x_3 + ... + b_{1k.23...(k-1)} x_k . \qquad (8.32)$$

По него могат да се намерят стойностите $\hat{x}_1$ за дадените емпирични стойности $x_2, x_3, ..., x_k$.

Ако са избрани например три факторни признака, ще се състави система от четири нормални уравнения:

(8.33)

$$\begin{cases} \sum x_1 = N\,a + b_{12.34}\sum x_2 + b_{13.24}\sum x_3 + b_{14.23}\sum x_4 \\ \sum x_1x_2 = a\sum x_2 + b_{12.34}\sum x_2^2 + b_{13.24}\sum x_3x_2 + b_{14.23}\sum x_4x_2 \\ \sum x_1x_3 = a\sum x_3 + b_{12.34}\sum x_3x_2 + b_{13.24}\sum x_3^2 + b_{14.23}\sum x_4x_3 \\ \sum x_1x_4 = a\sum x_4 + b_{12.34}\sum x_4x_2 + b_{13.24}\sum x_3x_4 + b_{14.23}\sum x_4^2 \ . \end{cases}$$

След като се намерят a, $b_{12.34}$ и $b_{14.23}$,съставя се регресионното уравнение

(8.34) $$\hat{x}_1 = a + b_{12.34}x_2 + b_{13.24}x_3 + b_{14.23}x_4 .$$

Когато броят на факторните признаци е голям, възникват трудности по решаването на системата от голям брой уравнения. В такъв случай е по-удобно да се приложат средствата на матричната алгебра. Съвременната електронна техника обаче има големи възможности за решаване на далеч по-сложни задачи.

Регресионните коефициенти, съдържащи се в регресионното уравнение, се наричат ***частни регресионни коефициенти*** (наричат се още частични, нетни, парциални). Всеки от тях измерва (в съответни единици според приетата мярка) влиянието на дадения фактор, като се елиминират останалите фактори, включени в модела (означени след точката в индекса към регресионния коефициент). Поради това, ако се изчислят частни регресионни коефициенти при определен брой признаци и след това се прибавят нови, ще се получат като правило други числови стойности на регресионните коефициенти за първоначалните факторни признаци. Преди включването им новите фактори скрито са влияли в някаква степен чрез първоначално включените. Ето защо, когато се интерпретират частните регресионни коефициенти, трябва винаги да се има предвид колко и какви фактори са включени в модела, а следователно и влиянието на колко и какви фактори се елиминира.

За измерване на теснотата на корелационната зависимост при многофакторния анализ могат да се изчислят частни (частични, нетни, парциални) и множествени корелационни коефициенти.

Частните корелационни коефициенти измерват теснотата на зависимостта между резултативния признак и съответните факторни признаци при условно елиминирано влияние на останалите. Те носят

същия алгебричен знак, какъвто е знакът на съответните частни регресионни коефициенти. Според това, колко факторни признака са включени в анализа, могат да се получат частни корелационни коефициенти при различен брой елиминирани фактори. Затова частните корелационни коефициенти са от различен ***порядък***. Ако се измерва зависимостта между два признака, без да се елиминира влиянието на други, коефициентите са от ***нулев порядък***. Това всъщност са коефициентите на единичната корелация. Когато се елиминира влиянието само на един фактор, коефициентите са от ***първи порядък***, при елиминиране на два фактора - от ***втори порядък*** и т.н. Между коефициентите от различен порядък има връзка, която създава възможност коефициентите от даден порядък да се получат от коефициентите от по-нисък порядък (обикновено от предходния порядък). Частните корелационни коефициенти от първи порядък се изчисляват от коефициентите от нулев порядък по формулите:

$$(8.35) \qquad r_{12.3} = \frac{r_{12} - r_{13}.r_{23}}{\sqrt{\left(1 - r_{13}^2\right)\left(1 - r_{23}^2\right)}};$$

$$(8.36) \qquad r_{13.2} = \frac{r_{13} - r_{12}.r_{23}}{\sqrt{\left(1 - r_{12}^2\right)\left(1 - r_{23}^2\right)}}.$$

Частните корелационни коефициенти от втори порядък могат да се изчислят от съответните коефициенти от първи порядък по формулите:

$$(8.37) \qquad r_{12.34} = \frac{r_{12.3} - r_{14.3}.r_{24.3}}{\sqrt{\left(1 - r_{14.3}^2\right)\left(1 - r_{24.3}^2\right)}};$$

$$(8.38) \qquad r_{13.24} = \frac{r_{13.2} - r_{14.2}.r_{34.2}}{\sqrt{\left(1 - r_{14.2}^2\right)\left(1 - r_{34.2}^2\right)}};$$

$$(8.39) \qquad r_{14.23} = \frac{r_{14.2} - r_{13.2}.r_{34.2}}{\sqrt{\left(1 - r_{13.2}^2\right)\left(1 - r_{34.2}^2\right)}}.$$

Като се има предвид връзката между отделните корелационни коефициенти от различен порядък, може да се състави обща формула на частния коефициент на корелацията между x_1 и x_2 при общо k признака.

$$(8.40) \qquad r_{12.34...k} = \frac{r_{12.34...(k-1)} - r_{1k.34...(k-1)}.r_{2k.34...(k-1)}}{\sqrt{\left(1-r^2_{1k.34...(k-1)}\right)\left(1-r^2_{2k.34...(k-1)}\right)}}.$$

От тази формула могат да се съставят формулите на всеки частен коефициент от съответен порядък.

За илюстрация на изчисляването на частните корелационни коефициенти ще приведем ***пример***. В опитно поле е направено наблюдение за установяване на зависимостта на средните добиви от царевица (x_1) от наторяването с минерални торове (x_2) и влагата в почвата (x_3) или напояването.

Изчисленият коефициент на единичната корелация между добива и наторяването (коефициента от нулев порядък) е $r_{12} = 0{,}904$, а между добива и влагата в почвата (напояването) - $r_{13} = 0{,}938$. Във връзка с изчисляването на частичните корелационни коефициенти е изчислен и $r_{23} = 0{,}852$.

Частният корелационен коефициент, измерващ зависимостта на добива от наторяването, при елиминиране влагата, т.е. коефициент от първи порядък е:

$$r_{12.3} = \frac{r_{12} - r_{13}.r_{23}}{\sqrt{\left(1-r^2_{13}\right)\left(1-r^2_{23}\right)}} = \frac{0{,}904 - 0{,}938.0{,}852}{\sqrt{\left(1-0{,}938^2\right)\left(1-0{,}852^2\right)}} = 0{,}577.$$

Частният коефициент, измерващ зависимостта на добива от влагата при елиминирано влияние на наторяването е:

$$r_{13.2} = \frac{r_{13} - r_{12}.r_{23}}{\sqrt{\left(1-r^2_{12}\right)\left(1-r^2_{23}\right)}} = \frac{0{,}938 - 0{,}904.0{,}852}{\sqrt{\left(1-0{,}904^2\right)\left(1-0{,}852^2\right)}} = 0{,}750.$$

Очевидно е, че много големият коефициент на единичната корелация между добив и наторяване е повлиян в неявна форма от влагата, която не е елиминирана. Същото важи и за корелацията между добива и влагата, когато не е елиминирано влиянието на наторяването. Частните коефициенти са значително по-малки, щом са елиминирани съответно факторите влага и наторяване. Частните коефициенти от втори порядък ще имат други стойности, ако се прибавят като факторен признак примерно температурите през вегетационния период в развитието на растенията.

Коефициентът на множествената корелация може да се изведе принципно по същия начин, както коефициентът на единичната корелация. Ако е съставено регресионното уравнение и са намерени значенията на x_1, може да се намери стандартната грешка на оценката или нейният квадрат - остатъчната дисперсия, която се означава с $S^2_{1,23...k}$:

$$S^2_{1,23...k} = \frac{\sum (x_1 - \hat{x}_1)^2}{N}. \qquad (8.41)$$

По-нататък трябва да се намери и общата дисперсия на x_1:

$$\sigma^2_1 = \frac{\sum (x_1 - \bar{x}_1)^2}{N}. \qquad (8.42)$$

Коефициентът на множествената корелация ($R_{1,23...k}$) ще се получи по формулата:

$$R_{1,23...k} = \sqrt{1 - \frac{S^2_{1,23...k}}{\sigma^2_1}}, \qquad (8.43)$$

или по нейната модификация:

$$R_{1,23...k} = \sqrt{1 - \frac{\sum (x_1 - \hat{x}_1)^2}{\sum (x_1 - \bar{x}_1)^2}}. \qquad (8.44)$$

Коефициентът на множествената корелация показва колко е силна зависимостта между резултативния признак и всички факторни признаци, включени в регресионния модел, взети в тяхното едновременно съвкупно действие.

Квадратът на коефициента на множествената корелация ($R^2_{1,23...k}$) се нарича ***коефициент на множествената детерминация***. Той показва каква част от общата вариация на признака x_1 се обуславя от вариацията на включените в анализа факторни признаци.

Квадратът на стандартната грешка на оценката може да се изчисли и от предварително изчислените единични и частни корелационни коефициенти

$$S^2_{1,23...k} = \sigma^2_1 (1 - r^2_{12})(1 - r^2_{13.2})(1 - r^2_{14.23})...(1 - r^2_{1k.23...(k-1)}). \qquad (8.45)$$

Ако $S^2_{1,23...k}$ се замести с този израз във формула 8.43, ще се получи:

$$R_{1,23...k} = \sqrt{1-\left(1-r^2_{12}\right)\left(1-r^2_{13.2}\right)\left(1-r^2_{14.23}\right)...\left(1-r^2_{1k.23...(k-1)}\right)}. \quad (8.46)$$

По данните от ***примера*** може да се изчисли коефициентът на множествената корелация, който измерва зависимостта на добива едновременно от наторяването и влагата в почвата:

$$R_{1,23...k} = \sqrt{1-\left(1-0{,}904^2\right)\left(1-0{,}750^2\right)} = 0{,}96.$$

Полученият коефициент показва много голяма зависимост на добива от двата фактора. Коефициентът на множествената детерминация е $R^2_{1,23...k} = 0{,}92$ или 92 % и показва, че 92 на сто от различията (вариацията) в средния добив се обуславят от различията (вариацията) в наторяването и влагата в почвата (напояването). Останалите 8 % (коефициент на множествената индетерминация) се дължат на други фактори.

При нелинейна множествена регресия и корелация възникват някои методологични проблеми, свързани главно с необходимите условия за коректно приложение на метода на най-малките квадрати. Част от нелинейните модели могат чрез логаритмични и други процедури да се трансформират в линейни. При други такива трансформации не са възможни и изискват по-специфичен подход. Тези проблеми излизат извън рамките на настоящата книга.[1]

8.4. Стохастични грешки и статистическа значимост на корелационните коефициенти

Известно е от гл. 5, че в много случаи е необходимо и целесъобразно параметрите на генералните съвкупности да се оценяват въз основа на репрезентативни извадки. Това важи и за регресионните и корелационните коефициенти като параметри на двумерните и многомерните разпределения. Когато обаче те са получени от извадки, винаги са обременени със стохастични грешки. Поради това не могат да се приемат "на доверие" като надеждни оценки на неизвестните коефициенти на генералните съвкупности, без да се изчисли размерът на

[1] По тези проблеми вж. **Съйкова, И.**, Цит. съч.; **Величкова, Н.**, и **И. Кацарска**, Приложение на регресионния и корелационния анализ при моделиране на икономически процеси, С., 1975; **Езекиэл, М**. и **К. Фокс**, Цит. съч., М., 1966.

грешките им или да се провери статистическата им значимост. Тук важат общите принципи и правила на статистическото оценяване и на статистическата проверка на хипотези. Ще разгледаме само някои от възможностите за изчисляване на стандартните грешки и на доверителните интервали, както и за проверка на ***значимостта на корелационните коефициенти***, без подробна обосновка и доказателства.

Съгласно класическия подход при статистическото оценяване, коефициентът на корелацията, изчислен по данни от случайна извадка, е точкова оценка на коефициента на корелацията в генералната съвкупност. За да се намери доверителният интервал и да се провери статистическата значимост на корелационния коефициент, необходимо е да се знае какво е разпределението на корелационните коефициенти. Ако разпределението е нормално, ***стандартната грешка на коефициента на единичната корелация*** може да се изчисли по формулата

$$\mu_r = \frac{1-r^2}{\sqrt{n-1}}. \qquad (8.47)$$

Като се умножи μ_r по доверителния коефициент, съответстващ на зададена вероятност, ще се получи максимално допустимата грешка (Δ_r). Доверителният интервал ще се получи, като се прибави и се извади максималната грешка от корелационния коефициент на извадката ($r \pm \Delta_r$).

Възможностите за приложение на формула 8.47 обаче са твърде ограничени поради ограничените възможности разпределението на корелационните коефициенти да е нормално. То е нормално или близко до нормалното при големи извадки (n > 100) и когато коефициентът на корелацията в генералната съвкупност е близък до 0. Разбира се, при емпиричните изследвания не знаем коефициента на корелацията в генералната съвкупност. Следователно не можем сигурно да се опрем на предположението за нормално разпределение, за да използваме горната формула. Изход от това затруднение е намерен в ***Z-трансформацията на Р. Фишер***. Тя се изразява в трансформиране на корелационния коефициент в една абстрактна характеристика Z_r чрез формулата:

$$Z_r = \frac{1}{2} ln \frac{1+r}{1-r}, \qquad (8.48)$$

където *ln* е натурален логаритъм с основа e = 2,71828.

Практически не е необходимо да се правят изчисления по тази формула, тъй като са съставени таблици, по които се намира Z_r за всяко значение на r от 0 до 1 (вж. приложение 8).

Разпределението на Z_r е близко до нормалното независимо от обема на извадката (n).[1] По такъв начин се удовлетворява изискването за нормалност на разпределението, за да се построи доверителният интервал. Стандартната грешка на Z_r е:

$$\mu_{Z_r} = \frac{1}{\sqrt{n-3}}. \qquad (8.49)$$

Известно е, че за нормално разпределение доверителният коефициент при вероятност 0,95 е 1,96, а при вероятност 0,99 - 2,58. Следователно 95 %-вия доверителен интервал е:

$$\text{от } Z_r - 1{,}96\frac{1}{\sqrt{n-3}} \text{ до } Z_r + 1{,}96\frac{1}{\sqrt{n-3}}, \qquad (8.50)$$

а 99 %-вия доверителен интервал -

$$\text{от } Z_r - 2{,}58\frac{1}{\sqrt{n-3}} \text{ до } Z_r + 2{,}58\frac{1}{\sqrt{n-3}}. \qquad (8.51)$$

За да се намери доверителният интервал на самия корелационен коефициент, е необходимо да се направи ***обратна трансформация*** на граничните стойности на Z_r в гранични стойности на r чрез формулата:

$$r = \frac{\left(e^{2Z} - 1\right)}{\left(e^{2Z} + 1\right)}, \qquad (8.52)$$

по която също има съставена таблица (вж. приложение 9).

По данните от ***примера*** в т. 8.2.1 беше изчислен коефициентът на корелацията между производителността на труда и трудовия стаж на 100 работници - r = 0,83. Да приемем, че тези 100 работници са случайна извадка (n), направена от генерална съвкупност от 1800 работници (N). Необходимо е да се намери доверителният интервал с вероятност 0,95, в

[1] Строго погледнато тази трансформация е приложима при n > 10. Предложена е (от Хотелинг) модификация на формулата за n < 50. Вж. **Закс, Л**., Цит. съч., к.393

който се съдържа корелационният коефициент на генералната съвкупност.

Намираме в приложение 8, че на $r = 0{,}83$ съответствува $Z_r = 1{,}19$ (закръглено до втория знак). При доверителен коефициент 1,96 (вероятност 0,95) максималната грешка е:

$$\Delta_{Z_r} = 1{,}96\frac{1}{\sqrt{100-3}} = 0{,}11.$$

Доверителният интервал е от 1,08 до 1,30, тъй като 1,19 - 0,11 = 1,08 и 1,19 + 0,11 = 1,30.

В приложение 9 за обратната трансформация намираме, че на $Z_r = 1{,}08$ отговаря $r = 0{,}79$ и на $Z_r = 1{,}30$ отговаря $r = 0{,}86$. Следователно 95 %-ият доверителен интервал на коефициента на корелацията е с граници от 0,79 до 0,86. Може следователно да се приеме (с вероятност 0,95), че коефициентът на корелацията в генералната съвкупност не е по-малък от 0,79 и не е по-голям от 0,86 ($0{,}79 \le r_0 \le 0{,}86$).

Друг аспект на статистическото заключение относно корелационния коефициент, изчислен от извадка, е проверката на неговата статистическа значимост. Числовата му стойност може да бъде повлияна от случайни фактори и може да се окаже сравнително висока дори и тогава, когато реално няма зависимост в генералната съвкупност.

Проверката на значимостта на корелационния коефициент се състои всъщност в проверка на нулевата хипотеза. Тя може да се направи по различни методи. И тук може да се приложи Z-трансформацията, като се изчисли характеристиката z по формулата

$$z = Z_r\sqrt{n-3}. \qquad (8.53)$$

След това се намира теоретичната стойност на z по таблицата за нормалното разпределение при предварително прието равнище на значимост. Ако емпиричното z е по-голямо от табличното, нулевата хипотеза се отхвърля и се приема алтернативната, т.е. приема се, че корелационният коефициент е статистически значим. Обратно, ако z е по-малко от $z_т$, нулевата хипотеза се потвърждава.

Доказано е, че когато коефициентът на корелацията в генералната съвкупност е равен на 0, т.е. когато нулевата хипотеза е вярна, характеристиката

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad (8.54)$$

има *t*-разпределение със степени на свобода (*n* - 2). Затова може да се използва таблицата за *t*-разпределението на Стюдент. Ако емпиричното *t* е по-голямо от табличното, нулевата хипотеза се отхвърля (при приетото равнище на значимост и (*n* - 2) степени на свобода).

При множествената корелация стандартната грешка може да се изчисли по формулата

$$\mu_{Z_r} = \frac{1}{\sqrt{n-k-1}}, \qquad (8.55)$$

където k е броят на признаците (променливите), а n – обемът на извадката.

Трансформацията на $R_{1,23\ldots k}$ в $Z_{R_{1,23\ldots k}}$ става по същия начин, както при единична корелация и процедурите за намиране на доверителния интервал са същите.

Проверката на значимостта на коефициента на множествената корелация се основава обикновено на *F*-разпределението на Фишер. Емпиричната характеристика *F* се намира по формулата

$$F = \frac{R^2_{1,23\ldots k}(n-k)}{(1-R^2_{1,23\ldots k})(k-1)}. \qquad (8.56)$$

При дадено равнище на значимост (α) теоретичната характеристика се намира по таблицата за *F*-разпределението при ($n - k$) и (k - 1) степени на свобода. Коефициентът на множествената корелация се смята за статистически значим (значимо различен от 0), ако емпиричната стойност на *F* е по-голяма от теоретичната.

Посочихме само някои от възможностите за определяне на доверителните интервали и за проверка на значимостта на коефициентите на корелацията. В тази област може да се подложи на проверка разликата между коефициенти, получени от две извадки, да се изчислят стандартни грешки и доверителни интервали на регресионните коефициенти и др.[1]

[1] Вж. **Съйкова, И.**, Цит. съч.

8.5. Регресионен и корелационен анализ при динамични редове

При емпиричните статистически изследвания, особено в областта на икономиката, се налага да се моделират и измерват зависимости между явления (процеси), разглеждани в развитие, т.е. представени в динамични редове. Интересно е например да се изследва чрез регресионния и корелационния анализ зависимостта на икономическия растеж от инвестициите в наука и образование през определен период, или на измененията през този период на разходите на домакинствата за транспортни нужди от изменението на доходите им и др. По принцип при тези и при други подобни случаи могат да се съставят регресионни модели, да се изчисляват регресионни и корелационни коефициенти. Много често обаче възниква положение, което влиза в противоречие с класическата теория на регресионния и корелационния анализ. И ако то не се преодолее, съществува риск да се направят погрешни изводи и оценки въз основа на резултатите от анализа.

Класическата теория на регресионния и корелационния анализ е изградена върху някои предпоставки, които не са налице при динамичните редове. Едно от основните изисквания на тази теория е отделните значения на признаците (променливите) да са независими помежду си. При динамичните редове те обикновено не са независими. В посочения пример не може да се твърди, че инвестициите в наука и образование през дадена година нямат никаква връзка с достигнатите в предходните години. Това важи и за изменението по години на брутния вътрешен продукт, чрез който се характеризира икономическия растеж. Във всеки от динамичните редове се проявява някаква тенденция (тренд).

Зависимостта, която съществува между членовете на динамичните редове (всеки член на реда се намира в зависимост от предходните), се нарича ***автокорелация***. Ако при анализа не се държи сметка за наличието на автокорелация, могат да се получат регресионни и корелационни коефициенти с високи стойности, които не се дължат на интересуващата ни зависимост между явленията, а на автокорелацията. Несъстоятелни биха били и изводите относно максималната грешка, доверителния

интервал и статистическата значимост на регресионните и корелационните коефициенти. Нерядко има автокорелация и между остатъчните елементи (ε) около линията на регресията, която прави недостоверни регресионните коефициенти, техните стохастични грешки и др. Ето защо е необходимо подходът при моделирането и измерването на зависимостите между динамичните статистически редове да се видоизмени така, че да се отстрани смущаващото влияние на автокорелацията. Възможно е и първо да се измери автокорелацията, което означава и да се провери дали има такава. Ако се установи, че няма или че е назначителна, тя може да се пренебрегне и по-нататък да се прилага класическият подход, описан в предходните параграфи. Това е по-скоро теоретична възможност, защото в анализираните динамични редове обикновено се съдържа трайна тенденция на изменение, в тях се съдържа автокорелация.

Разработени са и се прилагат в емпиричните изследвания различни методи за елиминиране на автокорелацията. Тук накратко ще бъде изложена същността на някои от тях, без подробни извеждания, обосновки и доказателства.[1]

8.5.1. Корелация между последователните разлики (диференчен метод)

Един от възможните начини за елиминиране на автокорелацията е да се измерва зависимостта между ***последователните разлики*** (прирасти). Този метод се нарича ***диференчен***.

Наличието на автокорелация в динамичните редове се изразява обикновено в наличието на някаква трайна тенденция (тренд), която не е обусловена само от зависимостта между явленията. Следователно да се отстрани автокорелацията, ще рече да се отстрани влиянието на тренда. Ако явлението, представено чрез динамичен ред, се развива с праволинейна трайна тенденция (праволинеен тренд), тази тенденция може да се опише графично с гладка права линия и аналитично с

[1] По-подробно по тези въпроси вж. **Величкова, Н.**, Статистически методи за изучаване и прогнозиране развитието на социално-икономическите явления, С., 1981, с.272 и сл.; **Андерсон**, **Т.**, Статистический анализ временных рядов, М., 1976, с.20 и сл.

уравнение на права. В такъв случай развитието би било с постоянен прираст, т.е. първите последователни разлики биха били еднакви, а вторите последователни разлики (прирастите на прирастите) ще бъдат равни на 0. Ако има зависимост между двете явления (двата реда), тя би се проявила в първите последователни разлики, в които е елиминиран трендът, т.е. елиминирана е автокорелацията. Затова регресионното уравнение и корелационният коефициент трябва да се намерят не непосредствено от значенията на x и y, а от ***първите последователни разлики*** $\Delta x = (x_t - x_{t-1})$ и $\Delta y = y_t - y_{t-1}$ по някоя от познатите вече формули.

Еднофакторният линеен регресионен модел би имал вида:

(8.57) $$\Delta y = a + b\Delta x .$$

Корелационният коефициент съответно може да се изчисли по известните вече формули, но представени с Δx и Δy вместо x и y:

(8.58) $$r_{\Delta y \Delta x} = \sqrt{1 - \frac{S^2_{\Delta y}}{\sigma^2_{\Delta y}}}$$ и

(8.59) $$r_{\Delta y \Delta x} = \frac{\sum (\Delta x - \overline{\Delta} x)(\Delta y - \overline{\Delta} y)}{\sqrt{\sum (\Delta x - \overline{\Delta} x)^2 \sum (\Delta y - \overline{\Delta} y)^2}} .$$

Ако трайната тенденция (трендът) се моделира с парабола от втора степен, регресионният модел и корелационният коефициент трябва да се изчислят от вторите последователни разлики. Изобщо, ако трайната тенденция се описва от полином от k-ти порядък, последните са ***k-те разлики*** (следващите са равни на 0) и следва да се измерва зависимостта между тези разлики. Има обаче достатъчно основания да се препоръчва по възможност да се оперира с последователни разлики от по-нисък порядък. При последователни разлики от по-висок порядък се затруднява съдържателното интерпретиране на резултатите от анализа.

Диференчният метод е практически удобен и в съдържателно отношение е ясен и лесно разбираем. Има обаче някои ограничителни условия за приложението му. Едно от изискванията е да няма цикличност в динамичните редове, т.е. трайната тенденция да се описва от гладка, плавна линия. Освен това е необходимо остатъчните елементи

(отклоненията) около регресионната линия да бъдат независими помежду си. Не е логично също да се измерва зависимост между последователни разлики от различен порядък, ако например трайната тенденция на развитието в единия ред е праволинейна по форма, а в другия има формата на парабола или друга крива.

8.5.2. Корелация между отклоненията (остатъчните елементи) около тренда

Логически този метод е сходен с диференчния метод. Разликата е в това, че предварително се намират нови (изравнени) стойности на динамичните редове, каквито те биха имали, ако явленията се развиват плавно, без колебания, по заложената в тях трайна тенденция. Тези стойности се намират обикновено по метода на най-малките квадрати, като значенията на x и y се разглеждат като функция от времето $\hat{x}_t = a + bt$ и $\hat{y}_t = a + bt$ (вж. гл. 10). Предполага се, че разликите между емпиричните стойности x и y и изчислените $\hat{x}$ и $\hat{y}$ не са автокорелирани. По-нататък се изчислява корелационен коефициент по познатите формули от тези разлики - $(x - \hat{x})$ и $(y - \hat{y})$. Той ще измерва зависимостта между двете явления при елиминирана автокорелация.

Този метод също не е безупречен във всички случаи и не може да се приеме като универсален. И тук е възможно например тенденциите в двата реда да са различни и следователно намирането на $\hat{x}$ и $\hat{y}$ да става по различни модели. Тогава също ще бъде некоректно да се измерва зависимостта между $(x - \hat{x})$ и $(y - \hat{y})$ щом адекватните модели за $\hat{x}$ и $\hat{y}$ са различни по форма.

8.5.3. Включване на времето като факторен признак (факторна променлива)

При прилагане на предходните методи времето не фигурира в явен вид. То участвува косвено доколкото значенията на признаците (променливите) се разглеждат като функция от времето, когато се изхожда от последователните разлики или от изравнените стойности на динамичните редове. То обаче може да се включи пряко в регресионното

уравнение като факторен признак (променлива) и по този начин се изключва автокорелацията. Моделът може да изглежда примерно така:

$$\hat{y} = a + bx + ct \,. \qquad (8.60)$$

Автокорелацията ще се поеме от параметър *c* пред *t*, който ще обхваща действието и на всички останали фактори, които не се съдържат във факторната променлива. Параметрите на регресионното уравнение и корелационните коефициенти могат да се интерпретират в съдържателно отношение по-определено и разбираемо. Препоръчва се наред с включването на времето да се оперира не непосредствено със значенията на *x* и *y*, а с логаритмите на техните верижни индекси (верижни темпове), т.е. $log\frac{x_t}{x_{t-1}}$ и $log\frac{y_t}{y_{t-1}}$.

В този метод, оперирайки с верижните темпове, по същество се включва под друга форма и принципът, заложен в диференчния метод.

Ако се използват логаритмите на верижните темпове, моделът по формула 8.60 ще приеме вида:

$$log\frac{y_t}{y_{t-1}} = a + b\,log\frac{x_t}{x_{t-1}} + ct \,. \qquad (8.61)$$

При регресионния и корелационния анализ на динамични редове възникват и редица други проблеми. Един от тях е този за закъсняващото действие (лаг) на факторите. Той се състои в това, че някои фактори действуват върху интересуващото ни явление не веднага, а след известно време. Разходите за наука и образование като съществен фактор за икономически растеж не дават ефект още през същата година. Необходимо е при анализа на зависимостите да се държи сметка за закъснението в действието на факторите. Това закъснение трябва по съответен начин да се отрази и в регресионните модели. Такива модели се наричат ***модели с разпределени лагове***.[1]

[1] Вж. **Величкова, Н.**, Цит. съч., с.315 и сл.

8.6. Измерване на зависимости при неинтервални скали

В началото на тази глава беше посочено, че в много случаи признаците, между които има зависимост, не могат да се представят на интервална скала и че са възможни комбинации на скали. От това се обуславят някои специфични особености на методите за измерване на зависимости при неинтервални скали, известни в литературата като ***непараметрични методи***. До някои от тях учените са достигнали по емпиричен път, а други са изведени въз основа на общите принципи на регресионния и корелационния анализ. Тези методи съдържат известни условности. Без да претендират за голяма точност и прецизна математическа обоснованост, те все пак осигуряват надеждна информация за зависимостите, когато параметричните методи са неприложими. Тук ще бъдат описани някои методи, които могат да се прилагат при съответни комбинации на скали, без подробно математическо извеждане и доказване.

8.6.1. Коефициенти на корелацията на ранговете

Да приемем, че единиците на дадена съвкупност са подредени (ранжирани) по значенията на два признака, между които има зависимост. Това означава, че е приложена рангова скала за двата признака. Тогава може да се измери зависимостта между двата реда от рангове. Коефициентът, който се изчислява в случая, се нарича ***коефициент на корелацията на ранговете***. Строго погледнато, така може да се постъпи и при интервални скали, като конкретните значения на признаците се заместят с техните рангове (поредни места, определени възходящо или низходящо). Такова преминаване обаче от по-силна скала към по-слаба не е оправдано, защото се губи възможността да се приложат по-съвършени методи. Има области, в които коефициентите на корелацията на ранговете са единствено възможните измерители на теснотата на зависимостта. Те са много необходими например, когато отделни обекти са ранжирани по експертни оценки.

Според конкретния случай се прилагат различни варианти на коефициентите на корелацията на ранговете.

а) Коефициент на корелацията на ранговете на Спирман

Ако единиците са ранжирани по два признака, зависимостта ще се проявява във взаимното разположение на двата реда рангове. Ако има пълна положителна зависимост, ранговете ще съвпадат: на ранг 1 ще съответствува ранг 1, на ранг 2 ще съответствува ранг 2 и т.н Ако зависимостта е пълна, но отрицателна, ранговете на двата признака ще се подреждат противоположно: на ранг 1 ще съответствува ранг N, на ранг 2 ще съответствува ранг $N-1$ и т.н. При липса на зависимост ще се получава безпорядъчно разположение на двата реда рангове. За характеризиране на фактическата зависимост е предложен от ***К. Спирман*** коефициент на корелацията на ранговете (ρ - ро), чиято формула е:

$$\rho = 1 - \frac{6\sum d^2}{N\left(N^2-1\right)}, \qquad (8.62)$$

където с d се означават разликите между ранговете по двата признака, а с N – броят на двойките рангове.

Ще илюстрираме изчисляването на коефициента на корелацията на ранговете на Спирман с ***пример***. Да допуснем, че посредством експертни оценки са ранжирани 10 села по благоустроеност и демографско състояние. Данните се съдържат в табл. 8.5.

Таблица **8.5**

Рангове на селата по благоустроеност и демографско състояние

Рангове по благо-устроеност	Рангове по демографско състояние	Разлики между ранговете	Квадрати на разликите между ранговете
		d	d^2
1	**2**	**3**	**4**
1	2	-1	1
2	1	1	1
3	4	-1	1
4	3	1	1
5	6	-1	1
6	5	1	1
7	8	-1	1
8	10	-2	4
9	7	2	4
10	9	1	1
			16

$$\rho = 1 - \frac{6.216}{10\left(10^2 - 1\right)} = 1 - 0{,}97 = 0{,}9$$

Изчисленият коефициент показва много голяма корелационна зависимост между благоустроеност на селата и демографското им състояние (по рангове според експертните оценки).

б) Коефициент на корелацията на ранговете на Кендал

Този коефициент, наричан още ***тау (τ) - коефициент на М. Кендал***, се различава по начин на извеждане, а обикновено и по числова стойност, от коефициента на Спирман. При него се въвежда понятието ***бал*** за изразяване на порядъка, в който са разположени взаимно, от една страна, всяка двойка рангове по единия признак, и съответно, двойките рангове по другия признак. Ако двойка рангове от единия ред е разположена в същия порядък, както и кореспондиращата й

двойка рангове от другия ред, тя получава бал +1, а ако е разположена в противоположен порядък - бал -1. Общата сума на всички балове се означава с *S*. При общо *N* единици максималният брой балове е $\frac{N(N-1)}{2}$. Отношението на фактическата сума на баловете (*S*) към максимално възможната $\frac{N(N-1)}{2}$ ще изразява степента на зависимост между ранговете. Това е ***коефициентът на корелацията на ранговете на Кендал***:

$$\tau = \frac{S}{\frac{N(N-1)}{2}} = \frac{2S}{N(N-1)}. \qquad (8.63)$$

Има възможност да се опрости и облекчи изчисляването на баловете. Ако единият ред рангове са подредени възходящо, той ще бъде ред от натурални числа от 1 до *N*. В такъв случай може да се оперира само с втория ред рангове.

Установява се колко ранга след първия са по-големи от него и колко са по-малки. Броят на по-големите се нарича брой на ***съответствията***, а броят на по-малките рангове - брой на ***инверсиите***. Общата сума на баловете (*S*) е равна на разликата между броя на съответствията (*P*) и броя на инверсиите (*Q*). По такъв начин формулата на коефициента на корелацията на ранговете на Кендал придобива вида:

$$\tau = \frac{P-Q}{\frac{N(N-1)}{2}}. \qquad (8.64)$$

По данните от ***примера*** коефициентът ще се изчисли по следния начин (табл. 8.6).

След ранг 2 от втората колона 8 са по-големи и 1 е по-малък. След ранг 1 по-големи са 8 и 0 по-малки. След ранг 4 по-големите са 6 и 1 е по-малък и т.н.

Таблица 8.6

Рангове на селата по благоустроеност и демографско състояние

Рангове по благо-устроеност	Рангове по демографско състояние	Брой на съответствията	Брой на инверсиите
1	2	8	1
2	1	8	0
3	4	6	1
4	3	6	0
5	6	4	1
6	5	4	0
7	8	2	1
8	10	0	2
9	7	1	0
10	9	0	0
		$P = 39$	$Q = 6$

$$\tau = \frac{P-Q}{\frac{N(N-1)}{2}} \frac{39-6}{\frac{10.9}{2}} = \frac{33}{45} = 0{,}73\,.$$

Формула 8.64 може да се преобразува така, че да съдържа само броя на съответствията или само броя на инверсиите:

$$\tau = \frac{4P}{N(N-1)} - 1; \qquad (8.65)$$

$$\tau = 1 - \frac{4Q}{N(N-1)}. \qquad (8.66)$$

По данните от примера

$$\tau = \frac{4.39}{10.9} - 1 = 1{,}73 - 1 = 0{,}73,$$

$$\tau = 1 - \frac{4.6}{10.9} = 1 - 0{,}27 = 0{,}73.$$

Като правило между числовата стойност на коефициентите на Спирман и Кендал има разлика.

При изчисляването на коефициентите на корелацията на ранговете може да има случаи, когато се получават т.нар. свързани рангове. Това са еднакви рангове за 2 или повече единици (обекти). Най-елементарният подход в такъв случай е да се приемат средни рангове. Ако например две села имат рангове 6 и 6, ще се запише 6,5 и 6,5 (средна от 6 и 7). При свързани рангове е възможно съответно да се видоизменят формулите на ранговите коефициенти.[1]

в) Коефициент на конкордацията (съгласуваността)

Има случаи, при които за група обекти (единици) може да има не два, а повече редове от рангове - например когато ранжирането на n единици (обекти) се извършва от няколко (m) експерти, независимо един от друг. Тогава за измерване на съгласуваността на експертните оценки чрез ранжиране се използва ***коефициентът на конкордацията*** (на съгласуваността) на ***М. Кендал***.

При такава ситуация един експерт ще даде за n обекти общо $\frac{n(n+1)}{2}$ ранга (сума от n члена на натуралния ред на числата), а общата сума на всички рангове, дадени от m експерти е $\frac{m.n(n+1)}{2}$. Може да се намери средната сума на ранговете за един обект, дадени от всичките m експерти:

$$\frac{m.n(n+1)}{2} : n = \frac{m(n+1)}{2}.$$

Може да се намери отклонението (D) на сумата на ранговете за един обект от средната сума.

Коефициентът на конкордацията (съгласуваността), означен с W, се намира по формулата

$$W = \frac{12\sum D^2}{m^2\left(n^3 - n\right)}. \qquad (8.67)$$

[1] Вж. **Кендэл, М**. Ранговые корреляции. М., 1975.

Ако ранговете съвпадат, т.е. при пълна съгласуваност, $W = 1$. Ако ранговете не съвпадат, т.е., ако няма пълна съгласуваност, $W < 1$, като най-малката му възможна стойност е 0.

Ще илюстрираме коефициента на конкордацията с ***пример***. Да приемем, че 5 дегустатори дегустират 4 марки вина, независимо един от друг. Всеки ги подрежда (ранжира) от първо до четвърто място. Дегустаторите са означени с А, Б, В, Г, Д, а вината-с I, II, III, IV. Резултатите от ранжирането и изчисленията са показани в табл. 8.7. Коефициентът на конкордацията трябва да покаже каква е степента на съгласуваност (сходство) в експертните оценки.

***Таблица* 8.7**

Ранжиране на вината и изчисляване на коефициента на конкордацията

Дегуста-тори	Вина				Сума на ранговете
	I	II	III	IV	
А	3	1	2	4	10
Б	1	2	4	3	10
В	2	1	3	4	10
Г	4	3	1	2	10
Д	2	4	3	1	10
Сума	12	11	13	14	50
D	-0,5	-1,5	0,5	1,5	0
D^2	0,25	2,25	0,25	2,25	5

Средната сума на ранговете за един обект (фактор) е

$$\frac{m(n+1)}{2} = \frac{5(4+1)}{2} = 12{,}5.$$

Коефициентът на конкордацията е

$$W = \frac{12\sum D^2}{m^2\left(n^3 - n\right)} = \frac{12.5}{5^2\left(4^3 - 4\right)} = 0{,}1.$$

Полученият коефициент показва ниска степен на съгласуваност (сходство) в оценките на дегустаторите, т.е. те имат твърде различни критерии и вкусове относно качествата на дегустираните вина.

8.6.2. Коефициенти на контингенцията

Когато се измерва зависимост между два признака и единият от тях или двата са скалирани по номинална или ординална скала, разгледаните вече методи са неприложими. Тогава се прилагат коефициенти, които могат да се обединят под общото наименование ***коефициенти на контингенцията.***[1] Всеки от тях има приложение според комбинацията от скали.

а) Коефициент фи (φ) на Пирсън

Този коефициент се използва, когато двата признака имат по две алтернативни значения, т.е. скалирани са по дихотомна скала. Данните в този случай се представят в таблица, в която единиците се разпределят в 4 клетки според комбинациите на значенията на признаците. (Затова понякога се говори за четириклетъчна корелация).

Ако двете алтернативни значения се записват условно с ДА и НЕ, а броят на единиците в клетките - с a, b, c, d, таблицата ще има вида:

Таблица 8.8

Първи признак / Втори признак	ДА	НЕ	Общо
ДА	a	b	$a + b$
НЕ	c	d	$c + d$
Общо	$a + c$	$b + d$	N

[1] В литературата се срещат различни наименования на коефициентите от тази група.

Коефициентът φ на Карл Пирсън се изчислява по формулата:

$$\varphi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}. \qquad (8.68)$$

Този коефициент може да приема стойности от 0 до ±1. Алгебричният му знак зависи от това, как се разполагат двете алтернативни значения на признаците.

Ще илюстрираме разглеждания коефициент с ***пример***. Да предположим, че са анкетирани 500 посетители на един туристически комплекс, от които 200 български граждани и 300 чуждестранни. Помолени са да отговарят на въпроса доволни ли са (ДА) или са недоволни (НЕ) от обслужването. Резултатите са представени в табл. 8.9. Необходимо е да се изчисли φ коефициента на контингенцията на Пирсън, който ще покаже доколко отговорите в двете алтернативи (ДА и НЕ) зависят от това дали са български или чуждестранни граждани.

Таблица 8.9

Разпределение на анкетираните по резултати от анкетата

Отговори / Групи анкетирани	Доволни (ДА)	Недоволни (НЕ)	Общо
Чужденци	220	80	300
Българи	140	60	200
Общо	360	140	500

$$\varphi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} = \frac{220.60 - 140.80}{\sqrt{300.200.360.140}} = 0{,}04.$$

Изчисленият коефициент показва, че зависимостта е слаба, т.е изискванията на посетителите относно обслужването почти не зависят от това дали са чуждестранни или български граждани.

б) Коефициент на асоциацията на Юл

При дихотомна скала за двата признака, както в примера, е възможно използването и на предложения от ***Дж. Юл коефициент на асоциацията:***

$$Q = \frac{ad - bc}{ad + bc}. \qquad (8.69)$$

По данните от примера той е :

$$Q = \frac{220.60 - 140.80}{220.60 + 140.80} = 0{,}08.$$

Коефициентът, изчислен по тази твърде елементарна по конструкция формула обикновено надценява зависимостта, по-голям е от коефициента на Пирсън.

в) Коефициент на колигацията

Това е коефициент, също предложен от ***Дж. Юл***, който е значително по-близък до коефициента φ на К. Пирсън.

$$Y = \frac{1 - \sqrt{\frac{bc}{ad}}}{1 + \sqrt{\frac{bc}{ad}}}. \qquad (8.70)$$

По данните от примера той е

$$Y = \frac{1 - \sqrt{\frac{140.80}{220.60}}}{1 + \sqrt{\frac{140.80}{220.60}}} = 0{,}04.$$

Както се вижда, по формула 8.70 на Юл се получава същия по числова стойност коефициент, както по формула 8.68 на Пирсън.

г) Коефициенти на взаимната свързаност на Пирсън, Чупров и Крамер

Когато признаците са скалирани по номинална или ординална скала, но имат повече от две значения (разновидности), т.е. таблицата е с произволен брой клетки (най-малко 4), в които се разполагат единиците, се използват предложените от ***К. Пирсън, А. А. Чупров*** и ***Х. Крамер*** коефициенти, основаващи се на χ^2-анализа. Те най-често се наричат коефициенти на взаимната свързаност и принадлежат към общата група коефициенти на контингенцията.

Коефициентът на Пирсън (С) може да се запише с формулата:

$$C = \sqrt{\frac{\varphi^2}{1+\varphi^2}}\ . \qquad (8.71)$$

Коефициентът на Чупров (К) се изчислява по формулата:

$$K^2 = \frac{\varphi^2}{\sqrt{(k_1-1)(k_2-1)}}\ . \qquad (8.72)$$

В двете формули $\varphi^2 = \frac{\chi^2}{N}$, където

$$\chi^2 = \sum_1^k \left[\frac{\left(f_{ij} - \hat{f}_{ij}\right)^2}{\hat{f}_{ij}} \right], \qquad (8.73)$$

където:

f_{ij} са фактическите честоти (броят на единиците) в клетките;

$\hat{f}_{ij}$ - теоретичните честоти в клетките;

k_1 - броят на групите (клетките) по единия признак;

k_2 - броят на групите (клетките) по другия признак.

Практически φ^2 се намира по значително опростен алгоритъм, като се изпълняват следните стъпки:

1. Честотите във всяка клетка на таблицата се повдигат на квадрат;
2. Квадратите се разделят на сбора на честотите по колони;
3. Получените числа се сумират по редове;

4. Получените суми се разделят на сбора на честотите по редове;
5. Получените числа се сумират, от сумата се изважда 1 и това е φ^2.

Трябва да се има предвид, че коефициентът *C* на Пирсън никога не достига единица. Освен това максимално възможната му стойност зависи от броя на групите по двата признака. При еднакъв брой групи (клетки) по единия и другия признак ($k_1 = k_2$) неговата максимално възможна стойност е $\sqrt{\frac{k-1}{k}}$. Ако $k_1 = 2$ и $k_2 = 2$, т.е. таблица с размери 2x2, коефициентът *C* не може да превишава 0,707; при $k_1 = 5$ и $k_2 = 5$ максималната му стойност е 0,894 и т.н. Това прави несравними коефициентите, изчислени при групировки с различен брой групи.

Коефициентът *K* на Чупров е по-прецизна мярка на зависимостта. Той може да приема стойности в границите от 0 до 1, какъвто и да е броят на групите (клетките) по двата признака, ако те за двата признака са еднакъв брой ($k_1 = k_2$). Когато обаче броят на групите по двата признака не е еднакъв ($k_1 \neq k_2$), коефициентът не може да достигне 1 при пълна зависимост. Практически не е трудно да се изравни броят на групите по двата признака.

Коефициентът на Крамер (V) може да се разглежда като модификация на формулата на Чупров, целяща коефициентът да се постави в теоретични граници от 0 до 1 и в случаите, когато $k_1 \neq k_2$:

$$V^2 = \frac{\varphi^2}{min[(k_1 - 1),(k_2 - 1)]}. \qquad (8.74)$$

Записът в знаменателя означава, че при изчисляването на коефициента се записва по-малкото от (k_1 - 1) и (k_2 - 1). Когато $k_1 = k_2$, $V = K$.

Ще илюстрираме изчисляването на трите коефициента по описаната процедура с ***пример***. Да приемем, че в една фирма е разработен проект за нова организация на труда. Направена е анкета, за да се проучи отношението на персонала към проекта. По данните от анкетата може да се изчисли коефициент, измерващ зависимостта на изразеното отношение към проекта от степента на образованието на анкетираните.

Таблица 8.10

Разпределение на анкетираните и изчисляване на φ^2

Образо-вание	Отношение към проекта						О б щ о		
	положително		отрицателно		безразлично				
Основно	30	900	10	100	20	400	60	11,255	0,188
		3,676		1,428		6,154			
Средно	120	14400	50	2500	40	1600	210	119,105	0,567
		58,776		35,714		24,615			
Висше	95	9025	10	100	5	25	110	38,324	0,348
		36,837		1,428		0,076			
Общо	245	x	70	x	65	x	380	x	1,103

В лявата част на клетките са записани честотите (броят на анкетираните) в групите според двата признака (30, 10, 20 и т.н.). В горната дясна част на клетките са записани квадратите на честотите ($30^2 = 900$; $10^2 = 100$ и т.н). В долната дясна част на клетките са записани числата, получени като се разделят квадратите на честотите на сумите на честотите по колони (900:245 = 3,673; 100:70 = 1,428 и т.н.). Сумите на получените числа се съдържат в предпоследната колона. В последната колона са записани числата, получени като се разделят сумите в предпоследната колона на сумите на честотите по редове (11,255:60 = 0,188 и т.н.).

$\varphi^2 = 1{,}103 - 1 = 0{,}103$.

Коефициентът на Пирсън е:

$$C = \sqrt{\frac{\varphi^2}{1+\varphi^2}} = \sqrt{\frac{0{,}103}{1+0{,}103}} = 0{,}31.$$

Коефициентът на Чупров:

$$K^2 = \frac{\varphi^2}{\sqrt{(k_1-1)(k_2-1)}} = \frac{0{,}103}{\sqrt{(3-1).(3-1)}} = 0{,}0515;$$

$$K = \sqrt{0{,}515} = 0{,}23\,.$$

Коефициентът на Крамер е:

$$V^2 = \frac{\varphi^2}{(k_1 - 1)} = \frac{0{,}103}{3-1} = 0{,}0515\,;$$

$$V = \sqrt{0{,}0515} = 0{,}23\,,$$

т.е. колкото е и коефициента на Чупров, тъй като $k_1 = k_2$.

Получените коефициенти показват слаба зависимост на отношението на анкетираните към проекта от тяхното образование.

8.6.3. Бисериални коефициенти

Когато се измерва зависимост между два признака, от които единият е по дихотомната скала (с две значения), а другият - по интервалната скала, възможни са два подхода. Първо, възможно е значенията на признака, скалиран по интервалната скала, да се сведат условно до две или повече и да се приложи някой от коефициентите на контингенцията. Такъв подход обаче е нецелесъобразен, защото това означава да се изостави по-силната скала и да се премине към по-слаба, а по този начин се губи част от информацията. Вторият подход се състои в изчисляването на ***бисериални (двусерийни) коефициенти***. Прилагат се обикновено два такива коефициента: точково-бисериален и бисериален, обосновани от ***К. Пирсън***.

Бисериалният коефициент е приложим тогава, когато единия признак е скалиран по дихотомната, а другият по интервалната скала, но е известно, че разпределението е нормално или близко до нормалното. Това изискване за нормално разпределение не винаги може да бъде удовлетворено, затова този коефициент има ограничено приложение.

Когато едната скала е рангова, а другата - интервална, използва се рангово-бисериален коефициент.

а) Точково-бисериален коефициент

Точково-бисериалният коефицент се прилага обикновено, когато видът на разпределението не е известен и като правило се смята, че то не е нормално.

Една от формулите на точково-бисериалния коефициент (r_{pb}), която е удобна за практическа работа, има следния вид:

$$r_{pb} = \frac{\bar{x}_1 - \bar{x}_0}{\sigma_x} \cdot \sqrt{\frac{n_1 n_0}{N(N-1)}}, \qquad (8.75)$$

където:

$\bar{x}_0$ е средна аритметична на значенията на признака по интервалната скала за единиците (n_0) със значение 0 по дихотомната скала;

$\bar{x}_1$ - средната аритметична на значенията на същия признак за единиците (n_1) със значение 1 по дихотомната скала;

n_0 - броят на единиците със значение 0 по дихотомната скала;

n_1 - броят на единиците със значение 1 по дихотомната скала;

N - общият брой на всички единици, т.е. $n_0 + n_1$.

Точково-бисериалният коефициент може да приема стойности от 0 до ±1.

Като ***пример*** да допуснем, че в дадена фирма едно изделие (или производствена операция) се произвежда (извършва) от мъже и жени, които имат съответна индивидуална производителност на труда. С цел да се проучи зависимостта на производителността от пола на работниците, направено е наблюдение на общо 60 работници (N), от които 40 мъже (n_1) и 20 жени (n_0). Средната дневна производителност на труда през периода на наблюдението за мъжете е 152 бр. ($\bar{x}_1$), а за жените - 146 бр. ($\bar{x}_0$). Средното квадратично отклонение на производителността на труда е $\sigma_n = 9{,}8$.

Точково-бисериалният коефициент по формула 8.75 е :

$$r_{pb} = \frac{\bar{x}_1 - \bar{x}_0}{\sigma_x} \cdot \sqrt{\frac{n_1 n_0}{N(N-1)}} = \frac{152 - 146}{9{,}8} \sqrt{\frac{40.20}{60(60-1)}} = 0{,}29.$$

Изчисленият коефициент показва слаба зависимост на производителността на труда на работниците от техния пол.

б) Рангово-бисериален коефициент

При емпирични изследвания може да възникне необходимост да се измери зависимостта между дихотомен признак и признак по рангова скала. За такива случаи е подходящ ***рангово-бисериалният коефициент***. Той се основава на подхода на ***М. Кендал*** при изчисляване на коефициента на корелацията на ранговете и една от практически удобните формули е:

$$r_{Rb} = \frac{2(\bar{y}_1 - \bar{y}_0)}{N}, \qquad (8.76)$$

където $\bar{y}_0$ и $\bar{y}_1$ са средните рангове на двете групи единици, обособени по дихотомния признак, а $N = n_0 + n_1$, т.е. общият брой на всички единици.

Рангово-бисериалният коефициент също може да приема стойности от 0 до ±1.

Да приемем ***например***, че на състезание (конкурс) по стенография са се явили 5 мъже (n_1) и 5 жени (n_0) и класирането им е представено с поредни места (рангове) от 1 до 10 според бързината на стенографирането. Изчислени са средните рангове на мъжете - $\bar{y}_1 = 6{,}2$ и на жените - $\bar{y}_0 = 4{,}8$.

Рангово-бисериалният коефициент, изчислен по формула 8.76 е:

$$r_{Rb} = \frac{2(\bar{y}_1 - \bar{y}_0)}{N} = \frac{2(6{,}2 - 4{,}8)}{10} = 0{,}28.$$

Като се има предвид, че максимално възможната стойност на коефициента е 1, може да се заключи, че класирането по бързина на стенографирането почти не зависи от половата принадлежност на участващите в състезанието.

8.7. Практикум

8.7.1. Въпроси за самопроверка

1. Как се дефинира корелационната зависимост?
2. Какво значи “лъжекорелация”?

3. Какви познавателни задачи решава регресионният и какви корелационният анализ?
4. Какво се разбира под “регресионен модел”?
5. Какво е познавателното значение на регресионните коефициенти?
6. Какво изразява (измерва) корелационният коефициент и коефициентът на детерминацията?
7. Как се избира адекватният регресионен модел при “конкуриращи се” функции?
8. Какъв е познавателният смисъл на частните (частичните, нетните) регресионни и корелационни коефициенти?
9. Как се обяснява понятието мултиколинеарност и какво е значението му при корелационния анализ?
10. От какво се налага Z–трансформацията на Фишер?
11. Какво се разбира под автокорелация и какъв проблем поражда тя при регресионния и корелационния анализ?
12. Какви са възможните методи за елиминиране на автокорелацията при анализа?
13. Кога се прилагат коефициентите на корелацията на ранговете?
14. Кога се прилагат коефициентите на контингенцията?
15. Какво се разбира под закъсняващо действие на факторите?

8.7.2. Задачи за упражнение

Задача 1. Фирма за радио-телевизионна техника има 10 филиала. Ръководството на фирмата се интересува от ефекта на разходите за реклама по отношение на оборота (постъпленията от продажбите). Данните се съдържат в следващата таблица. Те са примерни, обхващат малък брой единици и чрез тях само се илюстрира методологията на анализа.

Таблица 8.11

Филиали на фирма "Н" по разходи за реклама и реализиран оборот през 2007 г.

Филиали - N	Средномесечни разходи за реклама - хил. лв.	Средномесечен оборот - хил. лв.
1	40	800
2	25	600
3	38	1200
4	52	3400
5	65	4000
6	43	3100
7	70	3800
8	120	5200
9	80	4600
10	167	5300

Иска се:

1. Да се състави диаграма за зависимостта на оборота от разходите за реклама.
2. Да се състави регресионен модел при условие, че зависимостта е линейна.
3. Да се интерпретира регресионният коефициент.
4. Да се изчисли коефициент на корелацията и коефициент на детерминацията по метода на Пирсън и на Браве.
5. Да се изчислят регресионните коефициенти $b_{y/x}$ и $b_{x/y}$ чрез коефициента на корелацията и стандартните отклонения.
6. Да се изчисли корелационният коефициент чрез регресионните коефициенти.
7. Да се състави доверителен интервал на корелационния коефициент чрез Z-трансформацията на Фишер при доверителна вероятност 0,95. (Наблюдаваните филиали на фирмата се приемат условно като извадка от генерална съвкупност). Да се обосноват предположения относно причината за твърде широкия доверителен интервал.

Отговори:

2) $\hat{y} = 875{,}30 + 33{,}21x$

4) $r = 0{,}82;\qquad r^2 = 0{,}673.$

5) $\sigma_x = 41{,}27;\quad \sigma_y = 1671{,}53;$

$$b_{y/x} = 0{,}82\frac{1671{,}53}{41{,}27} = 33{,}21;\ b_{x/y} = 0{,}82\frac{41{,}27}{1671{,}53} = 0{,}0202.$$

6) $r = \sqrt{33{,}21.0{,}0202} = 0{,}82$

7) $0{,}397 \le r_0 \le 0{,}956$

Задача 2. Направен е анализ на производителността на труда в една фирма. Установено е, че за наблюдавания период коефициентът на корелацията, измерващ зависимостта на производителността (x_1) от образователното равнище на работниците (x_2) е $r_{12} = 0{,}969$, корелационният коефициент, измерващ зависимостта на производителността от възрастта на работниците (x_3) е $r_{13} = 0{,}426$, а коефициентът на корелацията между възрастта и образованието е $r_{23} = 0{,}362$.

Иска се:

1. Да се изчислят частни (частични, нетни) коефициенти на корелацията - $r_{12.3}$ и $r_{13.2}$ и да се направят съответните изводи.
2. Да се изчислят коефициентите на множествената корелация и на множествената детерминация и да се направят произтичащите от тях изводи.

Отговори:

1) $r_{12.3} = 0{,}965;\quad r_{13.2} = 0{,}321.$

2) $R_{1,23} = 0{,}972;\quad R^2_{1,23} = 0{,}945$ или 94,5 %.

Задача 3. Обявен е анонимен конкурс за архитектурен проект за луксозен хотел. Представени са 4 проекта, означени с I, II, III, IV. Възложено е на 5 експерти (А, Б, В, Г, Д) да оценят независимо един от друг отделните проекти и да ги ранжират от първо до четвърто място. Резултатите са представени в следващата таблица.

Таблица 8.12

Ранжиране на проектите и изчисляване на коефициента на конкордацията

Експерти	Проекти			
	I	II	III	IV
А	2	3	1	4
Б	3	1	2	4
В	4	2	1	3
Г	2	4	1	3
Д	4	1	3	2

Иска се да се изчисли коефициента на конкордацията.

Отговор: $W = 0{,}328$

Задача 4. В една фирма е направено за определен период наблюдение относно трудовата дисциплина. Наблюдавани са 220 заети лица, разпределени по две дихотомни (бинарни) скали: 1) семейни и несемейни; 2) идващи навреме на работа и закъсняващи. Резултатите са показани в следващата таблица.

Таблица 8.13

Брой на наблюдаваните лица

Идващи навреме / Неидващи навреме	Семейни или несемейни		Общо
	ДА	НЕ	
ДА	130	60	190
НЕ	10	20	30
	140	80	220

Иска се да се изчислят:

1. Коефициентът на контингенцията (φ) на Пирсън.
2. Коефициентът на асоциацията (Q) на Юл.
3. Коефициентът на колигацията (Y) на Юл.

Отговори:

1) $\varphi = 0{,}29$ 2) $Q = 0{,}625$ 3) $Y = 0{,}35$

9. СТРУКТУРЕН СТАТИСТИЧЕСКИ АНАЛИЗ

> "Законодателите и политиците често са понасяли несполуки затова, че статистическите им познания са били недостатъчни."
>
> Ф. Найтингейл

Съдържанието на тази глава запознава с основни теоретико-методологични положения на една сравнително нова област на теорията на статистиката и на нейните практико-приложни аспекти. Отнася се за анализ на различни структури – икономически, социални, демографски, професионални, технологични и др. Запознатият с това съдържание ще може да преценява как и с какви методи да измери и оцени измененията в структурите с течение на времето или различията между еднотипни структури по страни, области и други региони, както и концентрацията на определени дейности, диференциацията на доходите и т.н. Акцентът е поставен върху обосновката на един сравнително нов измерителен апарат и върху съдържателното интерпретиране на информацията, която той осигурява.

9.1. Обща постановка

При изследване на явленията в различни области от действителността много често е необходимо да се проникне във вътрешния им строеж, т.е в тяхната структура. И това е продиктувано не само от чисто познавателен интерес, но и от конкретни практически задачи. В областта на икономиката например въпросите, отнасящи се до възпроизводствената, отрасловата, продуктовата и други структури на производството са свързани пряко или косвено с икономическия растеж, с икономическата и социалната ефективност. Това важи и за структурата на разходите за производството, на стоковите запаси, външнотърговския стокообмен, личното потребление и т.н. Привличат интереса на много специалисти също демографските структури, структурата на доходите и

разходите на населението, професионално-квалификационния състав на работната сила и др. Всичко това изисква да се анализират структурите, да се оценяват настъпващите в тях изменения и поражданите от тези изменения ефекти. За тази цел е необходимо използването на подходящи статистически методи, каквито предлага едно сравнително ново направление в теорията на статистиката - ***структурният статистически анализ***.

Една от важните насоки на този анализ е измерването на ***структурните изменения***. Необходими са съответни обобщаващи измерители, за да се оцени интензитета на измененията, тяхното ускоряване или затихване.

Съществено значение има измерването на ***различията между две или повече структури*** в тяхното статично състояние (в даден момент или период). Необходимо е например да се измерят и оценят различията в производствените или други структури между две или повече страни, области, общини и др., или различията в структурата на потреблението на различни групи от населението и т.н.

Не по-малко значение има измерването на ***неравномерността на структурите***. То е свързано например с анализа на концентрацията на производството, с диференциацията на доходите, с териториалната локализация и специализация в областта на селското стопанство, промишлеността и др.

При емпиричните изследвания икономическите, социалните и други структури приемат формата на ***статистически структури***, които са техни специфицирани статистически отражения (образи). В този смисъл статистическата структура се разглежда като свойство на статистическата съвкупност, изразяващо нейния вътрешен строеж, представен числово чрез относителните дялове на отделните части (структурни елементи).

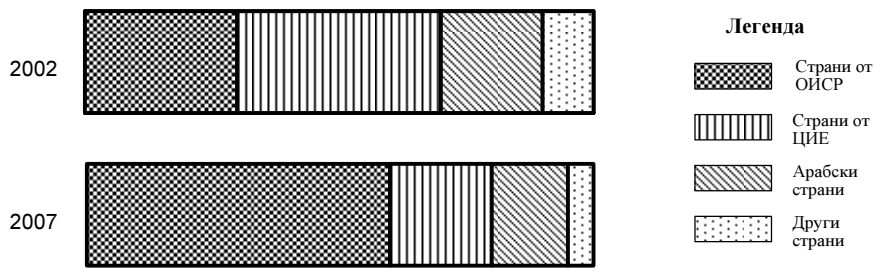
9.2. Графично представяне на структури

Графичното изобразяване на структурите и структурните различия е сравнително елементарно, но много полезно средство на анализа. То осигурява визуалните образи на това, което предстои по-нататък да се

измерва и интерпретира. За графично представяне на структурите обикновено се използват плоскостни диаграми, при които графичният образ се изобразява чрез правоъгълници, кръгове и др., разделени на части в такова съотношение, в каквото съотношение се намират структурните части на съвкупностите.

Общата ширина или височина на ***правоъгълника*** се приема за единица или 100 % и се разделя на толкова части, от колкото части се състои цялото, чиято структура се изобразява. С два или повече правоъгълници, съставени по този начин, отнасящи се за различни периоди, за различни териториални или други единици, може да се онагледи и изменението на структурите, различието между структурите на две съвкупности и пр.

На фигура 9.1. е представена диаграма по данните, съдържащи се в табл. 9.1.



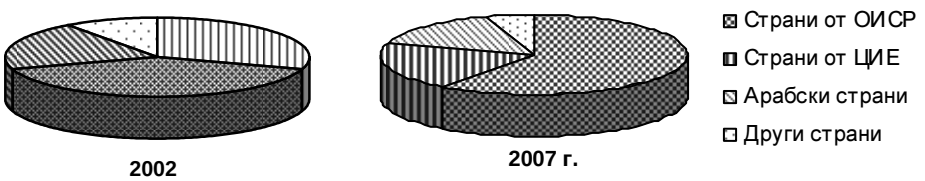
Фиг. 9.1. Географска структура на износа на фирма "Н"

Много подходяща за изобразяване на структури е ***секторната кръгова диаграма***. За да се раздели кръгът на сектори, пропорционални на представяните структурни части (елементи), се има предвид, че окръжността има 360 градуса, а цялата съвкупност, чиято структура се представя се приема за 100. Затова като се разделят 360 градуса на 100 %, се намира, че на 1% отговарят 3,6 градуса. Следователно трябва да се умножат на 3,6 % съответните относителни дялове в %, съставляващи дадената структура, за да се намери дъгата на всеки сектор.

Да вземем за ***пример*** също структурата на износа на фирма "Н" по групи страни от табл. 9.1. Изчисленията в отделните сектори са:

Страни	2002 г.	2007 г.
Страни от ОИСР	3,6º.30 % = 108º	3,6º.60 % = 216º
Страни от ЦИЕ	3,6º.40 % = 144º	3,6º.20 % = 72º
Арабски страни	3,6º.20 % = 72º	3,6º.15 % = 54º
Други страни	3,6º.10 % = 36º	3,6º.5 % = 18º
	360º	360º

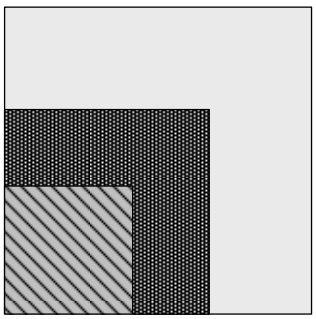
Съставената по този начин диаграма ще се състои от два кръга (фиг. 9.2).



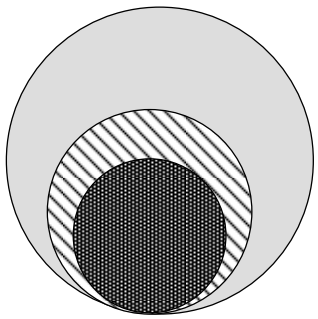
Фиг. 9.2

Има случаи, когато не е необходимо да се представи цялата структура с всичките й части, а само основната й (най-важната) част, от която също се обособява основна част и т.н. Например от стопанисваната селскостопанска земя най-важната част е обработваемата земя, а от нея - нивите.

В такива случаи подходящи са диаграмите, представляващи вписани един в друг квадрати, или кръгове (фиг. 9.3 и фиг. 9.4). Ако лицето на по-големия квадрат или кръг се приеме за единица или 100, лицето на по-малкия, вписан в него, ще заема такава част, какъвто е делът на представената по-малка част в по-голямата част от съвкупността.

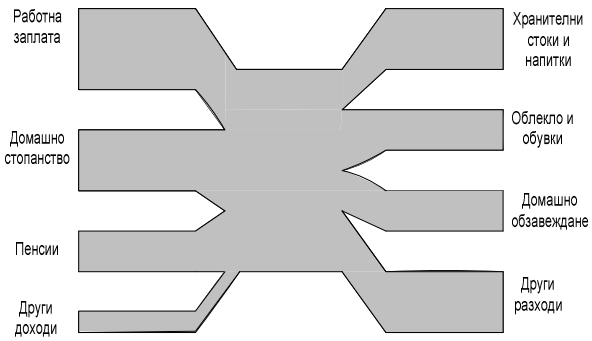


Фиг. 9.3



фиг. 9.4

Особен вид са графичните изображения на балансови структури, т.е. на структурата на приходната и на разходната част на различни баланси или бюджети. Например приходната част (по източници) и разходната част (по направления, по видове разходи) на домакинските бюджети, активната и пасивната част на баланса на фирмата и др. Такава диаграма с примерни данни е показана на фиг. 9.5.



Фиг. 9.5 Структура на доходите и разходите на домакинствата през ... година

Ширината на всеки "ръкав" в лявата и дясната част съответства на относителните дялове на отделните доходи по източници и на видовете (групите) разходи.

Специфична графична форма се прилага за представяне и анализ на неравномерността на структурите (диференциация на доходите,

концентрация на промишлеността и др.). Това е диаграмата на ***Лоренц*** (вж. фиг. 9.6).

9.3. Измерване на структурни изменения

Независимо от характера на структурите, при емпиричния анализ често се налага да се измери изменението им през определен период. То намира количествен израз в изменението на относителните дялове на отделните части на цялата съвкупност. Затова елементарен и много популярен начин за характеризиране на тези изменения е изчисляването на разликите (прирастите) или отношенията (индексите) на относителните дялове.

Ако означим с v_0 и с v_t относителните дялове на отделните части на съвкупността през два сравнявани периода (момента), разликите (прирастите) ще бъдат:

$$\Delta_v = v_t - v_0 . \qquad (9.1)$$

Те показват с ***колко пункта*** се е увеличил или намалил относителният дял на съответната част на съвкупността. И тъй като относителните дялове обикновено се представят в процент, разликата $v_t - v_0$ се чете като прираст (положителен или отрицателен) в процентни пунктове.

Прирастите на относителните дялове (Δ_v) характеризират ***абсолютните структурни изменения***.

Индексите на относителните дялове (I_v) ще бъдат:

$$I_v = \frac{v_t}{v_0} \qquad (9.2)$$

и показват ***колко пъти*** съответните относителни дялове при дадения период (момент) са по-големи или по-малки от тези през базовия период (момент). Те характеризират ***относителните структурни изменения***.

Посочените елементарни измерители на структурните изменения намират широко приложение, лесно разбираеми са и не пораждат проблеми от методологичен и изчислителен характер. Те обаче не дават ***обобщена характеристика*** на силата (степента) на настъпилите

структурни изменения в цялата съвкупност. Поради това са необходими и съответни обобщаващи измерители.

9.3.1. Индекси на различие и на относителна структура

Един възможен обобщаващ измерител на структурните изменения е ***индексът на различие*** (I_S), който се получава като сума от абсолютните разлики (прирасти) на относителните дялове през двата периода:

$$I_S = \sum |v_t - v_0| . \qquad (9.3)$$

Максималната сума на разликите е 2 (или 200, ако относителните дялове са в процент). Следователно I_S може да приема стойности в границите от 0 до 2 (респ. до 200). Ако сумата на разликите се раздели на 2, т.е. на максимално възможната сума, ще се получи:

$$I_S^* = \tfrac{1}{2} \sum |v_t - v_0| . \qquad (9.4)$$

По такъв начин I_S^* се нормира в теоретични граници от 0 до 1, т.е. удовлетворява се условието $0 \le I_S^* \le 1$.

Полусборът на абсолютните стойности на разликите между относителните дялове (I_S^*) се използва твърде често под различни наименования, като индекс на подобие, показател за трансформация на структури, индекс на разместване и др. Във всички случаи обаче той измерва общо ***абсолютните структурни изменения***.

За измерване на ***относителните структурни изменения*** се използва друг измерител, наречен ***индекс на относителна структура*** (I_{Sr}):

$$I_{Sr} = \sum \left| \frac{v_t - v_0}{v_0} \right| = \sum \left| \frac{v_t}{v_0} - 1 \right| . \qquad (9.5)$$

Горната формула показва, че индексът на относителната структура има друго съдържание и друг познавателен смисъл в сравнение с индекса на различие, тъй като се основава на индексите на относителните дялове.

За двата коефициента може да се каже, че са лесно разбираеми, но те не са достатъчно прецизни измерители на структурни изменения.

9.3.2. Линейни и квадратични коефициенти на структурни изменения

В търсене на теоретично по-добре обосновани и по-прецизни измерители са конструирани от ***Л. Казинец*** линейни (средноаритметични) и квадратични коефициенти на структурните изменения, по аналогия на средното аритметично и средното квадратично (стандартно) отклонение.[1] Аналогията се изразява както във формулите, по които се изчисляват, така и в изходната логическа постановка, според която структурните изменения се разглеждат като изменения във вариацията на относителните дялове. Тези коефициенти са конструирани в два варианта - за измерване на абсолютните и на относителните структурни изменения.

Линейният коефициент на абсолютните структурни изменения ($\delta_{\Delta v}$) може да се запише така:

$$\delta_{\Delta v} = \frac{\sum |v_t - v_0|}{k}, \qquad (9.6)$$

където k е броят на частите (структурните елементи) на съвкупността. Очевидно е, че той се получава като се раздели сумата на разликите между относителните дялове, третирана по формула 9.3 като индекс на различие, на броя на тези разлики. Следователно той показва с колко пункта ***средно*** се отличават относителните дялове през дадения период (момент) в сравнение с базовия. Когато няма структурни изменения, $\delta_{\Delta v} = 0$. Колкото по-големи са структурните изменения, толкова коефициентът $\delta_{\Delta v}$ ще бъде по-голям от 0.

Линейният коефициент на относителните структурни изменения (δ_{Iv}) може да се представи така:

$$\delta_{Iv} = \sum \left| \frac{v_t}{v_0} - 1 \right| v_0 \ . \qquad (9.7)$$

[1] **Казинец, Л.**, Измерение структурных сдвигов в экономике, М., 1969.

По замисъл той трябва да покаже средната "интензивност" с която се изменят относителните дялове на частите на съвкупността.

Квадратичният коефициент на абсолютните структурни изменения ($\sigma_{\Delta v}$) има вида:

$$\sigma_{\Delta v} = \sqrt{\frac{\sum (v_t - v_0)^2}{k}} . \qquad (9.8)$$

Следователно той е непретеглена квадратична средна величина на разликите между относителните дялове в двата периода и има същия познавателен смисъл, както и линейния коефициент на абсолютните структурни изменения. Различието е само във формата на осредняването, от което произтича и разликата в числовите им стойности. Известно е, че квадратичната средна е винаги по-голяма от аритметичната средна. Известно е също, че средното квадратично отклонение има преимущества в сравнение със средното аритметично отклонение и следователно трябва да се предпочита квадратичният коефициент.

Квадратичният коефициент на относителните структурни изменения (σ_{Iv}) може да се изчисли по формулата:

$$\sigma_{Iv} = \sqrt{\sum \left(\frac{v_t}{v_0} - 1 \right)^2 v_0} . \qquad (9.9)$$

Очевидно е, че това по същество е квадратична средна величина на относителните прирасти и показва колко средно се отклоняват индексите на относителните дялове от средната им величина, приета за единица. Доказано е, че квадратичният коефициент на относителните структурни изменения превишава по числова стойност коефициента на абсолютните структурни изменения.

Конкретните познавателни задачи при емпиричните изследвания могат да насочат вниманието към коефициентите на абсолютните или на относителните структурни изменения. Най-често обаче ни интересува степента на различие между структурите в двата сравнявани периода (момента). От само себе си се разбира, че различието (изменението от единия период до другия) между структурите е резултат на неравномерността в развитието на отделните части на съвкупността, чиято

структура се анализира. Това означава, че от разгледаните коефициенти най-добре съответствува на познавателната задача при структурния анализ квадратичният коефициент на абсолютните структурни изменения. Този коефициент обаче (формула 9.8) не е нормиран в определени теоретични граници. Той може да се нормира като се раздели на максимално възможната му стойност, която е $\sqrt{\frac{2}{k}}$, тъй като $\sum(v_t - v_0)^2 \leq 2$. В такъв случай ще се получи ***нормираният квадратичен коефициент на абсолютните структурни изменения*** ($\sigma^*_{\Delta v}$).

$$\sigma^*_{\Delta v} = \sqrt{\frac{\sum(v_t - v_0)^2}{2}}, \qquad (9.10)$$

при което се удовлетворява изискването $0 \leq \sigma^*_{\Delta v} \leq 1$.

9.3.3. Интегрален коефициент на структурни изменения

Беше посочено, че квадратичният коефициент на абсолютните структурни изменения се основава на разликите между относителните дялове на съответните части на съвкупността през сравняваните периоди (моменти), без да се вземат под внимание самите размери на относителните дялове, от които са изчислени тези разлики. Например разликата между 90 % и 80 % е 10 пункта и между 20 % и 10 % е също 10 пункта. При еднакви по размер разлики квадратичните коефициенти ще имат една и съща числова стойност, независимо от конфигурацията на структурата. Казано по друг начин, при еднакви абсолютни изменения, както в горния пример, относителните изменения са различни.

Логично е да се изисква обобщаващата характеристика на абсолютните структурни изменения да отразява различията при подобни на примера ситуации.

Целесъобразно е следователно коефициентът на структурните изменения да се конструира така, че: ***първо***, да е нормиран в теоретични граници от 0 до 1; ***второ***, да отразява не само разликите между относителните дялове, но и размера на тези дялове за двата периода; ***трето***, да е достатъчно чувствителен (селективен) по отношение на

промените в структурите. Тези изисквания могат да се удовлетворят, ако средната квадратична разлика между относителните дялове, т.е. квадратичният коефициент на абсолютните структурни изменения ($\sigma_{\Delta v}$), показан във формула 9.8, се раздели на израза $\sqrt{\frac{\Sigma v_0^2}{k} + \frac{\Sigma v_t^2}{k}}$.

В такъв случай и след съответни съкращения ще се получи измерител, който може да се нарече ***интегрален коефициент на структурни изменения***[1] (K_S), защото по същество интегрира абсолютните и относителните структурни изменения:

$$K_S = \sqrt{\frac{\sum (v_t - v_0)^2}{\sum v_0^2 + \sum v_t^2}} . \qquad (9.12)$$

С елементарна преработка формула 9.12 може да се модифицира в:

$$K_S = \sqrt{1 - \frac{2\sum v_0 v_t}{\sum v_0^2 + \sum v_t^2}} . \qquad (9.13)$$

Формула 9.13 разкрива най-добре смисъла на интегралния коефициент, конструиран по описания начин. Не са необходими особени доказателства за това, че когато структурите през сравняваните два периода (момента) са еднакви, т.е. когато за всички части на съвкупността $v_0 = v_t$, тогава $\sum v_0 v_t = \sum v_0^2 = \sum v_t^2$ и следователно $K_S = 0$.

Коефициентът K_S ще бъде равен на 1, когато двете структури са напълно противоположни. Това е теоретична възможност при екстремален случай: когато съвкупността се състои само от две части (двуелементна структура), през първия (базовия) период относителният дял на едната част е 0 и на втората част е 1, а през другия период първата част е 1 и втората - 0. Примерът е абстрактен, но той показва теоретичната възможност K_S да достигне 1, т.е. максималната горна граница. Следователно $0 \le K_S \le 1$. Колкото по-големи са структурните изменения при реални ситуации, толкова повече K_S ще се стреми към 1.

[1] По-подробно относно извеждането и логиката на този коефициент вж. **Гатев, К.**, Методи за анализ на структури и структурни ефекти, С., 2007.

Смисълът на посочения начин на нормиране на квадратичния коефициент и конструирането на интегралния коефициент се състои главно в това, разликите между относителните дялове да се поставят във връзка с размера на самите относителни дялове през двата периода. По такъв начин той измерва абсолютните структурни изменения, но във връзка с относителните, като "цената" на определено увеличение или намаление на относителните дялове се определя според величината на тези дялове.

Ще илюстрираме с конкретен ***пример*** изчисляването на разгледаните: индекс на различие, нормиран квадратичен коефициент на абсолютните структурни изменения и интегрален коефициент на структурните изменения (табл. 9.1).

Да приемем, че е дадена географската структура, т.е. по групи страни, на износа на една фирма през 1995 г. и 2007 г. (табл. 9.1).

Таблица 9.1

Географска структура на износа на фирма "Н"

Групи страни	Относителни дялове		Разлики между относителни дялове	Квадрати на разликите	Квадрати на относителните дялове	
	2002 г.	2007 г.			2002 г.	2007 г.
	v_0	v_t	$(v_t - v_0)$	$(v_t - v_0)^2$	v_0^2	v_t^2
ОИСР	0,30	0,60	0,30	0,0900	0,0900	0,3600
ЦИЕ	0,40	0,20	-0,20	0,0400	0,1600	0,0400
Арабски	0,20	0,15	-0,05	0,0025	0,0400	0,0225
Други	0,10	0,05	-0,05	0,0025	0,0100	0,0025
Сума	1,00	1,00	0,00	0,1350	0,3000	0,4250
	-	-	$\lvert 0,60 \rvert$	-	-	-

$$I_S^* = \tfrac{1}{2}\sum |v_t - v_0| = \frac{0,60}{2} = 0,30;$$

$$\sigma^*_{\Delta v} = \sqrt{\frac{\sum (v_t - v_0)^2}{2}} = \sqrt{\frac{0{,}135}{2}} = 0{,}26;$$

$$K_S = \sqrt{\frac{\sum (v_t - v_0)^2}{\sum v_0^2 + \sum v_t^2}} = \sqrt{\frac{0{,}135}{0{,}30 + 0{,}425}} = 0{,}43.$$

Очевидна е разликата между числовите стойности на трите измерителя. Тя се дължи на посочените особености в тяхната конструкция и изходни положения. Оценката на силата на настъпилите структурни изменения, измерена с който и да е от тях, се опира на интервала от 0 до 1, в който те могат да варират.

При използването на разгледаните измерители на структурни изменения трябва да се има предвид, че числовата им стойност зависи от броя на структурните елементи (броя на частите на съвкупността). Поради това сравняването на коефициенти, изчислени за структури с различен брой елементи, е некоректно. При необходимост от такива сравнения трябва предварително структурите да се приведат в еднакъв брой елементи.

9.4. Измерване на различия между пространствени и други статични структури

При емпиричните изследвания в различни области на действителността често се налага да се измерят не само структурните изменения във времето, но и различията между две или повече структури в тяхното статично състояние. При международни сравнения например е интересно да се установи в каква степен структурата на икономиката на дадена страна се различава от аналогичната структура на други страни. Такива сравнения са възможни и необходими и в рамките на отделна страна по административни или други териториални единици, по отрасли, фирми, обособени групи от населението и др. В редица случаи е целесъобразно сравняването на структури, формирани по различни признаци, например сравняване на секторната структура на дълготрайните активи със секторната структура на заетите лица. Възможно е също сравняването на дадена фактическа структура с нормативна, планова или друга структура

на потреблението на домакинствата със структурата, формирана по определени рационални норми. Изобщо в много области чрез измерването на различията между две или повече структури в тяхното статично състояние може да се получи полезна информация за интересуващите ни явления.

По принцип разгледаните в предходната точка методи за измерване на структурни изменения са приложими и при сравняване на пространствени и други статични структури. Вместо структурни изменения в случая се измерват ***структурни различия***. Съответно вместо за коефициенти на структурни изменения се въвежда терминът ***коефициенти на структурни различия***.

Широко приложение, особено при регионалните изследвания, намира разгледаният индекс на различие, който според случая се нарича ***коефициент на локализация***, ***коефициент на специализация***, ***коефициент на географска асоциация*** и др.

За сравняване на пространствени и други статични структури може да се приспособи интегралният коефициент на структурни изменения, който в случая ще се нарича ***интегрален коефициент на структурни различия*** (K_D) и формулата му ще изглежда така:

$$K_D = \sqrt{\frac{\sum (v_1 - v_2)^2}{\sum v_1^2 + \sum v_2^2}}, \qquad (9.14)$$

или още:

$$K_D = \sqrt{1 - \frac{2\sum v_1 v_2}{\sum v_1^2 + \sum v_2^2}}. \qquad (9.15)$$

И този коефициент е нормиран в граници от 0 до 1, т.е. $0 \le K_D \le 1$.

И тук трябва да се припомни, че числовата стойност на коефициентите на структурни различия зависи от броя на структурните елементи. Затова и в пространствен разрез са сравними само коефициенти за еднакви по брой на елементите структури.

9.5. Измерване на неравномерността на структурите

При редица изследвания в икономическата и социалната област интерес представлява неравномерността на структурите. Това може да бъде например отклонението на реалните структури от една условна (еталонна) равномерна структура. Под ***равномерна структура*** в случая се разбира структурата, отделните части (елементи) на която имат еднакви относителни дялове. С нея се сравнява реалната структура, за да се измери и оцени нейната ***неравномерност***. Неравномерността в този смисъл може да се нарече ***абсолютна неравномерност***.

При анализа по-често ни интересуват ***свързани едномерни структури***. Едната структура изразява относителните дялове на единиците на съвкупността, обособени в групи по даден признак. Другата се отнася за сумарните значения на признака по същите групи. Ако например домакинствата са разпределени по групи според дохода средно на домакинство и за всяка група е посочен сумарния доход (сумарно значение на групировъчния признак) и съответно са изчислени относителните дялове на домакинствата и на доходите по групи, ще се получат две свързани помежду си структури. При сравняването им се установява неравномерността на втората по отношение на първата. В случая става дума за ***сравнителна неравномерност***. В примера това по същество е въпрос за измерване на подоходната диференциация на домакинствата. Същата би била постановката, ако една съвкупност от фирми, е разпределена по брой на заетите лица и за всяка обособена група е посочен общият брой на заетите лица. В случая измерването на сравнителната неравномерност е по същество измерване на концентрацията.

За измерване на неравномерността на структурите и в частност на сравнителната неравномерност има разработени редица методи. Тук ще бъдат разгледани само някои от тях.[1]

[1] По-подробно изложение на тези методи виж в: **Гатев, К**., Методи за анализ на структури и структурни ефекти, С., 2007.

9.5.1. Интегрален коефициент на неравномерността на структурите

Беше вече посочено, че абсолютната неравномерност се разглежда като отклонение на реалната структура от равномерната, приета за еталон. Щом е така, при нейното измерване може да се изходи от същите позиции, от които се формира подходът при измерване изобщо на различията между две структури и на структурните изменения. Обстоятелството, че се сравнява фактическа структура с изкуствено, хипотетично дефинирана равномерна структура, по същество не изменя нещата.

Да приемем за основа формула 9.12. на интегралния коефициент. Ще означим относителните дялове на фактическата структура с v_i, а на равномерната с $\frac{1}{k}$ (или $\frac{100}{k}$, ако относителните дялове са в процент; с k е означен броят на структурните елементи). Формулата на ***коефициента на абсолютната неравномерност*** (K_R) може да се запише така:

$$K_R = \sqrt{\frac{\sum\left(v_i - \frac{1}{k}\right)^2}{\sum v_i^2 + k\left(\frac{1}{k}\right)^2}} \quad . \qquad . \tag{9.16}$$

След съответна преработка ще приеме вида:

$$K_R = \sqrt{1 - \frac{2}{1 + k\sum v_i^2}} \; . \tag{9.17}$$

Ако относителните дялове са в процент,

$$K_R = \sqrt{1 - \frac{20000}{10000 + k\sum v_i^2}} \; . \tag{9.18}$$

Формули 9.17. и 9.18. предлагат значителни практически удобства, тъй като при изчисляването на K_R се оперира само с относителни дялове на реалната структура, чиято неравномерност се изследва.

Както при всички други случаи, интегралният коефициент е нормиран в теоретични граници от 0 до 1 (респ. до 100), т.е. $0 \le K_R \le 1$. Колкото реалната структура е по-неравномерна, толкова коефициентът

K_R е по-голям от 0 и се стреми към 1, но практически не достига тази горна теоретична граница.

При измерване на ***сравнителната неравномерност*** по принцип е приложим същият подход, но се сравняват две свързани едномерни структури. Нека се върнем към примера за диференциацията на доходите. Измерването на степента на сравнителната неравномерност (диференциацията) се свежда до измерване на различието между двете свързани структури.

В този и други подобни случаи могат да се приложат два варианта. При ***първия вариант*** чрез групировка се формират групи с различен брой единици, а следователно и с различни относителни дялове (например групи домакинства по доход). След разпределение на единиците на съвкупността (домакинствата) по тези групи се определя общата сума на значенията на признака за всяка група (сумарният доход на всяка подоходна група домакинства).

При ***втория вариант*** се формират еднакви по относителен дял (следователно и по абсолютни единици) групи от единици на съвкупността и за всяка група се установява относителният дял на сумарните значения на групировъчния признак. Например формират се групи домакинства, всяка от които обхваща по 10 % от всички домакинства (децилни групи), следвайки възходяща градация по размер на дохода. За всяка група се установява каква част тя получава от общата сума на всички доходи. При сравнително равномерно разпределение на доходите всеки 10 % от домакинствата биха получавали по 10 % от общата сума на доходите. Това е хипотетична постановка за равномерност, за да стане възможно измерването на неравномерността като реална и обективно необходима ситуация.

Тук е излишно да се коментират по същество двата варианта. Изследователят на съответните явления може да предпочете единия или другия съобразно с конкретните си интереси и наличната информация. И в двата случая задачата от методологична гледна точка се свежда до измерване на различие между две свързани едномерни структури, което характеризира сравнителната неравномерност (диференциацията). И в двата случая следователно е приложим разгледаният интегрален

коефициент на различията между структури, модифициран съответно в ***интегрален коефициент на сравнителната неравномерност на структурите*** (K_{PR}).

Ако относителните дялове за едната структура (например за броя на домакинствата по групи) се означат с v_i, а за другата структура (например за сумарните доходи по групи) - с v_j, интегралният коефициент на сравнителната неравномерност при описания по-горе първи вариант може да се представи с формулата:

$$K_{PR} = \sqrt{\frac{\sum (v_i - v_j)^2}{\sum v_i^2 + \sum v_j^2}}, \qquad (9.19)$$

или с нейната модификация

$$K_{PR} = \sqrt{1 - \frac{2\sum v_i v_j}{\sum v_i^2 + \sum v_j^2}}. \qquad (9.20)$$

При втория вариант (еднакви относителни дялове за първата структура) v_i е постоянно, равно на $\frac{1}{k}$ (или $\frac{100}{k}$). Тогава е приложима разгледаната вече формула 9.17 на интегралния коефициент на неравномерността, която тук приема вида:

$$K_{PR} = \sqrt{1 - \frac{2}{1 + k\sum v_j^2}}. \qquad (9.21)$$

Когато v_j е в процент (съгласно формула 9.18),

$$K_{PR} = \sqrt{1 - \frac{20000}{10000 + k\sum v_j^2}}. \qquad (9.22)$$

Ще илюстрираме изчисляването на интегралния коефициент на сравнителната неравномерност (концентрация) с ***пример***, съставен съобразно постановката на първия вариант. В табл. 9.2 се съдържат примерни данни за относителните дялове на броя на фирмите и на броя на заетите лица по групи фирми според средния годишен брой на заетите в

в една фирма. За измерването на сравнителната неравномерност (концентрацията) в този случай са приложими формули 9.19 и 9.20.

Таблица 9.2

**Разпределение на промишлените фирми в отрасъл ....
към ..... 2007 г. по брой на заетите лица**

Групови интервали по брой на заетите лица	Относителни дялове в %		Разлики между относи-телните дялове	Квадрати на разликите	Квадрати на относителните дялове	
	фирми	заети лица			фирми	заети лица
	v_i	v_j	$(v_i - v_j)$	$(v_i - v_j)^2$	v_i^2	v_j^2
до 10	0,3	0,1	0,2	0,04	0,09	0,01
11-30	2,5	0,1	2,4	5,76	6,25	0,01
31-50	4,8	0,3	4,5	20,25	23,04	0,09
51-80	7,4	0,9	6,5	42,25	54,76	0,81
81-140	10,8	3,5	7,3	53,29	116,24	12,25
141-200	13,6	2,8	10,8	116,64	184,96	7,84
201-500	31,8	19,8	12,0	144,00	1011,24	392,04
501-1000	15,7	21,3	-5,6	31,36	246,49	453,69
1001-2000	9,2	24,3	-15,1	228,01	84,64	590,49
Над 2000	3,9	26,9	-23,0	529,00	15,21	723,61
	100,0	100,0	0,0	1170,60	1743,32	2180,84

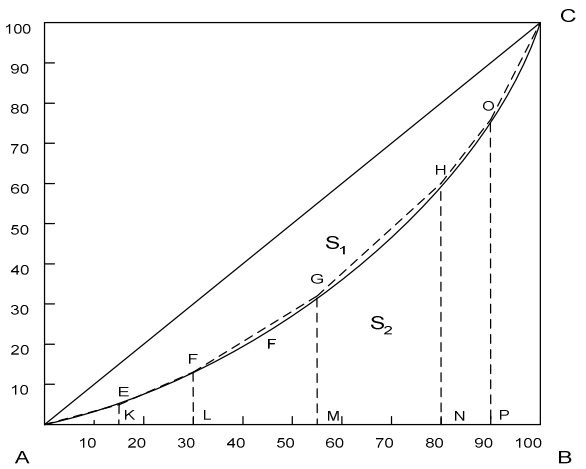
$$K_{PR} = \sqrt{\frac{1170,60}{1743,32 + 2180,84}} = 0,55 .$$

Като се имат предвид теоретичните граници на $K_{PR} (0 \le K_{PR} \le 1)$, концентрацията (сравнителната неравномерност) трябва да се оцени като значителна. Ако такива коефициенти се изчисляват през определен период, може да се характеризира и проследи изменението в степента на концентрацията, респ. на деконцентрацията.

9.5.2. Коефициент на неравномерността на структурите на Лоренц-Джини

Един широко разпространен и достатъчно добре описан в литературата метод за измерване на неравномерността на структурите (концентрацията, диференциацията) е коефициентът на ***Корадо Джини*** , основаващ се на диаграмата на ***Макс Лоренц (1876-1959)***. Затова може да бъде наречен ***метод на Лоренц - Джини***.[1] Тук ще бъдат изложени неговата същност и една от възможните модификации на подхода при извеждане на формулата.

Кривата на Лоренц е графичен модел на неравномерността на структурите. Той представлява линейна диаграма, разположена в поле с квадратна форма (вж. фиг. 9.6). Линията, която описва неравномерността, има съответна математическа интерпретация, която е излишно да разглеждаме. Илюстрирано с примера от предходната точка, построяването на ***диаграмата на Лоренц*** става по следния начин.



Фиг. 9.6

[1] **Lorenz, M**., Methods of Measuring the Concentration of Weahlt. Journal of the American Statistical Association, Vol.70, 1905, **Gini, C**., Indici di conzentratione et di Dependenza. Bologna, 1910.

На абсцисната ос е съставена скала, на която се отчитат кумулативните (с последователно натрупване) относителни дялове на брой на фирмите (C_i), а на ординатната ос - скала за кумулативните относителни дялове на броя на заетите (C_j). Ако има пълна сравнителна равномерност, т.е. няма концентрация (диференциация), относителните дялове на фирмите по групи ще съвпадат с относителните дялове на заетите, т.е. ще съвпадат напълно и кумулативните относителни дялове ($C_i = C_j$). Линията, намерена по двете скали на кумулативните относителни дялове, ще съвпада с диагонала на квадрата, свързващ долния ляв ъгъл с горния десен ъгъл (точките A и C на фиг. 9.6), тъй като на всеки кумулативен относителен дял C_i ще съответства същият кумулативен относителен дял C_j. Когато има неравномерност, между кумулативните относителни дялове C_i и C_j ще има разлика. Тогава ще се получи крива, която повече или по-малко ще се отклонява от диагонала, т.е. от линията на равномерността. Площта, затворена между кривата на неравномерността и линията на равномерността (диагонала) ще се изменя при изменение на степента на неравномерността. Тази площ (S_1) се нарича факторна или ***концентрационна площ***, а останалата площ (S_2) на триъгълника ABC под линията AC се нарича остатъчна или неконцентрационна площ. Следователно колкото по-голяма част от площта на триъгълника ABC е затворена между емпиричната крива и диагонала, толкова неравномерността е по-голяма. Това дава основание отношението на концентрационната площ (S_1) към площта на целия триъгълник ABC ($S = S_1 + S_2$) да се приеме като коефициент на неравномерността (концентрацията, диференциацията), означаван с G_R. Той може да се запише най-общо така:

$$G = \frac{S_1}{S} = \frac{S_1}{S_1 + S_2}. \qquad (9.23)$$

Ако площта на целия квадрат се приеме за 1, площта на триъгълника ABC е $S = \frac{1}{2}$. Следователно

(9.24) $$G = \frac{S_1}{\frac{1}{2}} = 2S_1 .$$

И тъй като $S_1 = S - S_2 = \frac{1}{2} - S_2$,

(9.25) $$G = 2\left(\frac{1}{2} - S_2\right) = 1 - 2S_2 .$$

Очевидно е, че при пълна равномерност G_R ще бъде равен на 0. При неравномерност (концентрация) той ще бъде по-голям от 0 и теоретичната му горна граница е 1 ($0 \le G_R \le 1$).

Проблемът при изчисляването на G_R се свежда до намирането на S_2. Съществуват различни начини за това. Ще изложим някои от тях без подробни извеждания и доказателства.

На кумулативните относителни дялове C_i и C_j отговорят съответни точки в диаграмата на Лоренц и именно кривата, която съединява тези точки, е кривата на неравномерността. Ако всеки две точки се съединят с прави, те и съответните им отсечки от абсцисата и ординатите на точките образуват трапеци. (На фиг. 9.6 тези прави са показани пунктирано). Търсената площ S_2 може да се представи като сбор от лицата на тези трапеци. Основата на всеки трапец е равна на съответния относителен дял v_i, а страните съответстват на кумулативните относителни дялове C_j. Тъй като лицето на всеки трапец се намира като произведение от основата и полусбора на двете му страни, сборът на лицата на всички трапеци, който е търсеното S_2, ще бъде:

(9.26) $$S_2 = \frac{1}{2}\sum_{1}^{k} v_i \left(C_j + C_{j-1}\right).$$

Като се замести с този израз, S_2 във формула 9.25, ще се получи:

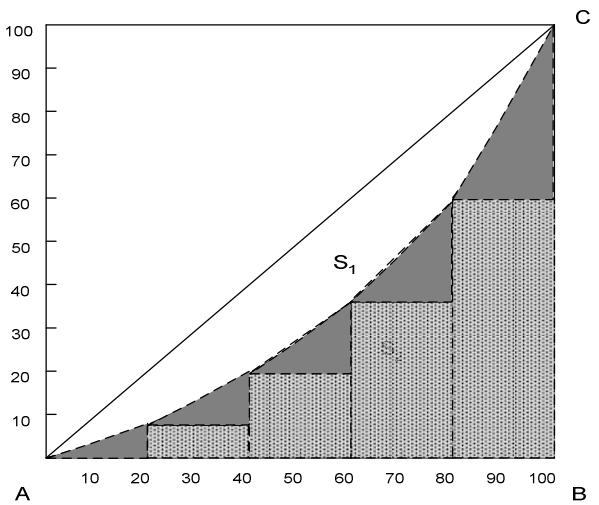
(9.27) $$G_R = 1 - \sum_{1}^{k} v_i \left(C_j + C_{j-1}\right).$$

Но тъй като v_i може да се запише като разлика между всеки кумулативен относителен дял C_j и предходния кумулативен относителен дял C_{j-1}, горната формула може да приеме вида:

$$G_R = 1 - \sum_{1}^{k} \left[(C_i - C_{i-1})(C_j + C_{j-1}) \right]. \qquad (9.28)$$

Изложеният подход е приложим в този му вид при разгледания първи вариант на групировка, когато груповите интервали обхващат различен брой единици на съвкупността. При втория вариант, когато се формират групи с еднакъв брой единици (еднакви относителни дялове), е възможен и друг подход за изчисляване на G_R.

Да приемем в ***примера***, че фирмите са разпределени по равен брой във всяка група, следвайки възходящо подреждане според броя на заетите, например първите 10%, вторите 10 % и т.н. (децилни групи). В общия случай при k групи относителният дял на броя на фирмите във всяка група ще бъде $\frac{1}{k}$ (или $\frac{100}{k}$). На всяка група фирми съответстват определен относителен дял на заетите (v_j). На тях ще съответстват определени кумулативни относителни дялове C_j. При съставянето на диаграмата на Лоренц могат да се образуват триъгълници и четириъгълници, сборът от лицата на които ще бъде равен на S_2. Това е показано на фиг. 9.7., където условно е прието $k = 5$.



Фиг. 9.7

Основата на всеки триъгълник и на всеки четириъгълник е равна на $\frac{1}{k}$ (или $\frac{100}{k}$). Височините на триъгълниците са равни на относителните дялове на заетите (v_j), а височините на четириъгълниците - на кумулативните им относителни дялове (C_j).

Лицето на всеки триъгълник ще бъде равно на $\frac{1}{k}.\frac{1}{2}v_j$. Сумата от лицата на всички триъгълници ще бъде

(9.29) $$\frac{0{,}5}{k}\sum_{j=1}^{k} v_j = \frac{0{,}5}{k}$$ (тъй като $v_j = 1$).

Лицето на всеки четириъгълник е равно на $\frac{1}{k}C_j$. Броят на четириъгълниците обаче е с един по-малък от броя на триъгълниците, т.е. броят им е k - 1 и височината на последния е C_{k-1}.

Сумата от лицата на всички четириъгълници ще бъде

(9.30) $$\frac{1}{k}\sum_{j=1}^{k-1} C_j .$$

Търсената площ S_2 ще бъде сбор от лицата на всички триъгълници и всички четириъгълници:

(9.31) $$S_2 = \frac{0{,}5}{k} + \frac{1}{k}\sum_{j=1}^{k-1} C_j .$$

Като се замести с този израз S_2 във формула 9.25,

(9.32) $$G_R = 1 - 2\left(\frac{0{,}5}{k} + \frac{1}{k}\sum_{j=1}^{k-1} C_j\right).$$

След малка преработка тя приема вида:

(9.33) $$G_R = 1 - \frac{1 + 2\sum_{j=1}^{k-1} C_j}{k}.$$

Тъй като последната кумулативна честота (C_k) е равна на единица, то $\sum_{j=1}^{k-1} C_j = \sum_{j=1}^{k} C_j - 1$. Затова формула 9.33, след съответна преработка, може да се представи така:

$$G_R = 1 - \frac{2\sum_{j=1}^{k} C_j - 1}{k}. \qquad (9.34)$$

Изчислителната работа по тази формула е значително опростена, защото се свежда само до намирането на общия сбор на кумулативните относителни дялове C_j.

Ще илюстрираме изчисляването на коефициента на Лоренц - Джини по данните от табл. 9.2. Прилага се формула 9.27. За целта е съставена табл. 9.3

Таблица 9.3

**Разпределение на фирмите в отрасъл ..... към ..... 2007 г.
по брой на заетите лица и изчисляване на
коефициента на Лоренц - Джини**

Групови интервали по брой на заетите лица	Относителни дялове		Кумулативни относителни дялове		$C_j + C_{j-1}$	$v_i(C_j + C_{j-1})$
	фирми	заети лица	фирми	заети лица		
	v_i	v_j	C_i	C_j		
до 10	0,003	0,001	0,003	0,001	0,001	0,000003
11-30	0,025	0,001	0,028	0,002	0,003	0,000075
31-50	0,048	0,003	0,076	0,005	0,007	0,000336
51-80	0,074	0,009	0,150	0,014	0,019	0,001406
81-140	0,108	0,035	0,258	0,049	0,063	0,006804
141-200	0,136	0,028	0,394	0,77	0,126	0,017136
201-500	0,318	0,198	0,712	0,275	0,352	0,111936
501-1000	0,157	0,213	0,869	0,488	0,736	0,119791
1001-2001	0,092	0,243	0,961	0,731	1,219	0,112148
над 2000	0,039	0,269	1,000	1,000	1,731	0,067509
	1,000	1,000				0,437144

$$G_R = 1 - \sum_{1}^{k} v_i \left(C_j + C_{j-1}\right) = 1 - 0{,}44 = 0{,}56.$$

Изчисленият коефициент на Лоренц-Джини е твърде близък до интегралния коефициент на сравнителната неравномерност (K_{PR} = 0,55) и също показва значителна концентрация. В други случаи разликата може да бъде по-голяма. Не е доказано по теоретичен път наличие на строго определено съотношение между двата коефициента. Те са изградени върху различни изходни положения и принципи. В предходното изложение бяха посочени особеностите на интегралния коефициент на структурните изменения (K_S), модификация на който е интегралният коефициент на сравнителната неравномерност на структурите (K_{PR}). Относно коефициента на Лоренц-Джини трябва да се има предвид, че той може да приема еднакви числови стойности при различни конфигурации на структурите, т.е. при различно положение на кривата на Лоренц щом не се изменя съотношението между площите S_1 и S_2.

Както интегралният коефициент, така и коефициентът на Лоренц-Джини, зависи от броя на групите (k), на които е разчленена съвкупността. Затова могат да се сравняват само коефициенти, изчислени при еднакъв брой групи (еднакъв брой структурни елементи).

Освен разгледаните коефициенти в литературата има предложени редица други. Една част от тях се основават също на диаграмата на Лоренц, останалите са изградени върху други изходни положения. Те обаче се прилагат сравнително по-рядко и затова не се разглеждат в тази книга.

9.6. Практикум

9.6.1. Въпроси за самопроверка

1. Какво е съдържанието на понятието статистическа структура?
2. Какво значи свързани едномерни структури?
3. Какви основни задачи се решават чрез структурния статистически анализ?

4. В какво се състои съществената разлика между квадратичния коефициент и интегралния коефициент на структурните изменения?
5. Какво значи абсолютна и сравнителна неравномерност на структурите?
6. Как се изчислява интегралния коефициент на сравнителната неравномерност?
7. Какво представлява диаграмата на Лоренц?

9.6.2. Задачи за упражнение

Задача 1. Дадена е структурата на дълготрайните активи на една фирма през 1995 и 2005 г., посочена в следващата таблица.

Таблица 9.4

Структура на дълготрайните активи на фирма "Н" през 2000 и 2007 г.

Вид на активите	Относителни дялове		Разлики между относи-телните дялове	Квадрати на разли-ките	Квадрати на относителните дялове	
	2000 г.	2007 г.			2000 г.	2007 г.
	v_0	v_t	$(v_t - v_0)$	$(v_t - v_0)^2$	v_0^2	v_t^2
Сгради и съоръжения	0,43	0,22	-0,21	0,0441	0,1849	0,0484
Машини и устройства	0,50	0,70	0,20	0,0400	0,2500	0,4900
Транспортни средства	0,04	0,06	0,02	0,0004	0,0016	0,0036
Други	0,03	0,02	-0,01	0,0001	0,0009	0,0004
Сума	1,00	1,00	0,00	0,0846	0,4374	0,5424
$\sum\lvert v_t - v_0\rvert$			0,60			

Иска се:

1. Да се представят чрез кръгова диаграма структурите през двете години.

2. Да се измери общото изменение на структурата на дълготрайните активи през периода 1995 - 2005 г. чрез:

а) индекса на различие

б) квадратичния коефициент на абсолютните структурни изменения

в) интегралния коефициент на структурните изменения

Отговори:

2. $I_S^* = 0{,}22$; $\sigma_{\Delta v}^* = 0{,}21$; $K_S = 0{,}30$.

Задача 2. Съвкупност от домакинства през дадена година са разпределени в 10 групи от по 10 %, подредени възходящо по размер на доходите. Изчислени са относителните дялове и на доходите на отделните групи домакинства. Данните са примерни и са посочени в следващата таблица.

Таблица 9.5

Структура на доходите на домакинствата по подоходни групи през ... г. и изчисляване на коефициентите на подоходната диференциация

Пореден номер на подоходните групи от по 10 % от домакинствата	Относителен дял на доходите %	Квадрати на относителните дялове	Кумулативни относителни дялове
	v_j	v_j^2	C_j
1	2,7	7,29	2,7
2	4,5	20,25	7,2
3	5,6	31,36	12,8
4	6,5	42,25	19,3
5	7,5	56,25	26,8
6	8,7	75,69	35,5
7	10,0	100,00	45,5
8	11,9	141,61	57,4
9	15,0	225,00	72,4
10	27,6	761,76	100,0
	100,0	1461,46	379,6

Иска се:

1. Да се състави диаграмата на Лоренц.

2. Да се изчисли интегралният коефициент на сравнителната неравномерност на структурата (на подоходната диференциация).

3. Да се изчисли коефициентът на Лоренц-Джини.

Отговори:

2. K_{PR} = 0,434
3. G_R = 0,341

10. ДИНАМИЧЕН СТАТИСТИЧЕСКИ АНАЛИЗ

"Времето е в нас и ние сме във времето"

Васил Левски

Съдържанието на тази глава е със силно подчертано практико-приложно значение. То се предопределя от безспорната необходимост явленията да се разглеждат в тяхното развитие във времето. Запознатият с това съдържание не само ще разбере смисъла на съответните статистически характеристики, но ще може лесно да ги прилага. Ще може да измерва скоростта на развитието на интересуващите го явления, заложените в тях тенденции (трендови модели), сезонността и цикличността в протичането на определени процеси. Ще разбере особеностите и условията за коректно прилагане на различните методи. Тези знания и умения са необходими на почти всеки специалист, анализатор на реални процеси, независимо от неговата конкретна професия.

10.1. Обща постановка

Необходимостта от изследване на явленията в тяхното развитие - от миналото през настоящето към бъдещето, не подлежи на съмнение. Всяко явление съществува и протича във времето. Не се налага следователно да се доказва и потребността от методи за динамичен статистически анализ.

Конкретните познавателни задачи определят насоките, обхвата и приложимите методи на анализа. Интересът е ориентиран най-често към изследване на: 1) скоростта на развитието на явленията; 2) основната тенденция (тренда), обусловена от систематично действуващи относително трайни фактори; 3) периодично повтарящи се сезонни вариации в рамките на годишни периоди, които се дължат на специфични за определени части (сезони) от годината климатични или други особености;

4) цикличните колебания (ако има такива) в дългосрочното развитие на явленията.

Основа на емпиричния анализ в посочените направления са динамичните (хронологичните) статистически редове, съдържащи данни за периодни или моментни съвкупности (вж. т. 9 от гл. 2). Затова вместо динамичен анализ се използва и терминът анализ на динамични редове.

Безусловно изискване за коректност на динамичния анализ, както и изобщо за статистическия анализ, е редовете да се състоят от данни, които са сравними и съпоставими по същество, по териториален и времеви обхват, по използвана мярка и др. Освен това в редица случаи е необходимо редовете да са достатъчно дълги, за да могат да се изявят изучаваните тенденции и закономерности.В методологичен аспект основополагащо значение имат две изходни положения.

Първо, развитието се разглежда като функция от времето - $Y = F(t)$.

Второ, развитието на изследваното явление, представено в динамичния ред, може да се разложи на съставни компоненти. Размерът му (Y) може да съдържа в себе си тренд ($\hat{Y}$) и случаен компонент, т.е. $Y = \hat{Y} + \varepsilon$. В определени случаи се съдържа сезонен компонент (S), т.е. $Y = \hat{Y} + S + \varepsilon$, а в по-дълги периоди - цикличен компонент (C), т.е. $Y = \hat{Y} + C + \varepsilon$.

При това връзката може да бъде адитивна или мултипликативна. Ако приемем в общия случай, че в развитието се съдържат тренд, циклични колебания, сезонни колебания и случайни колебания, адитивният модел ще има вида: $Y = \hat{Y} + C + S + \varepsilon$.

Мултипликативният модел съответно може да бъде: $Y = \hat{Y}.C.S.\varepsilon$.

Моделът в някои случаи може да бъде и смесен, например от вида: $Y = \hat{Y}.C.S + \varepsilon$.

Посочените модели, представящи развитието като композиция от елементи, предлагат възможност да се измери всеки един от тях и да се проследи поведението им в динамика.

10.2. Графично представяне на динамичните редове

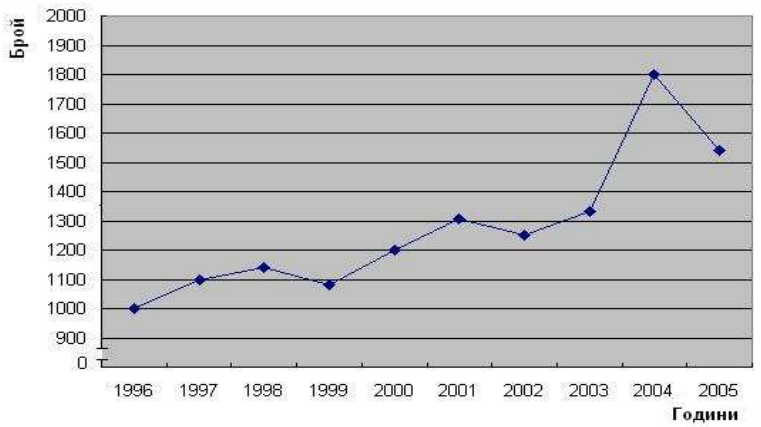
Развитието (динамиката) на изучаваните явления, което се съдържа в динамичните (хронологичните) редове, се представя графично най-често с линейни и стъпаловидни плоскостни диаграми.

Линейните диаграми са много подходящи когато се представят сравнително дълги редове, когато се сравнява развитието едновременно на 2 и повече явления или се очертава основната тенденция (тренда) на развитието. Те се строят обикновено в първия квадрант на ортогоналната координатна система чрез две скали - по абсцисната ос за времето и по ординатната ос за размера на явлението (за членовете на динамичния ред). Най-съществения момент при тези диаграми е съставянето на скалите така, че да се изрази реално скоростта на изменението. Строги предписания в това отношение са невъзможни, но все пак съществуват някои правила (изисквания), които подпомагат правилното съставяне на скалите.

Необходимо е винаги, когато е възможно, ширината (основата) на диаграмата да бъде по-голяма от нейната височина в съотношение 1 към $\sqrt{2}$, т.е. ако височината е примерно 5 см., ширината трябва да бъде 5.1,414 = 7 см.

Когато чрез мащабите на скалите се очертава общо полето на диаграмата, това трябва да стане така, че графичният образ на линията да не заема само долната или само горната половина на полето, т.е. да не остава излишна голяма част от скалата по ординатната ос. По принцип нулевата точка на тази скала трябва да се съдържа във всяка диаграма. Ако при представяните данни част от скалата се окаже излишна, тя се прекъсва, например по начин, показан на фиг. 10.1. (данните са примерни). Виж също и фиг. 10.5.

**Производство на продукт "А" във фирма "Н"
през периода 1996 - 2005 г.**



Фиг. 10.1

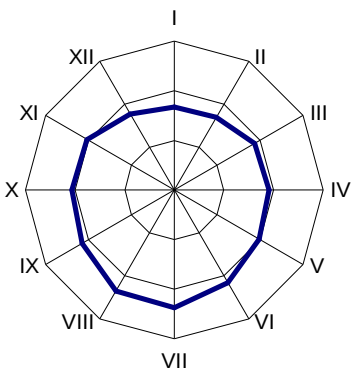
Линейните диаграми може да имат различна конкретна форма, според това дали се представя развитието на едно или повече явления, каква е целта на графичното изобразяване и т.н.

Линейната диаграма, съставена с аритметична скала (с еднакви деления) на ординатната ос дава визуална представа за ***абсолютните изменения*** на явлението от един подпериод до друг. Ако едновременно са представени две явления (два реда), възприемат се визуално различията в абсолютните изменения. Ако целта е да се покажат и сравнят ***относителните изменения***, аритметичната скала е неподходяща. Тогава трябва да се състави ***логаритмична*** линейна диаграма. Тя може да се състави по два начина: 1) на скалата с еднакви деления се отчитат логаритмите на представяните величини и по тях се определят точките, през които ще се прекара линията (графичният образ); 2) съставя се логаритмична скала, т.е. с различни поделения, съответстващи на логаритмите и по-нататък се оперира с дадените действителни величини, съдържащи се в динамичния ред. Вторият начин е практически по-удобен, защото не изисква предварително логаритмуване на представяните величини.

Когато на диаграмата е логаритмична само скалата по ординатната ос, тя обикновено се нарича ***полулогаритмична***, а когато са логаритмични двете скали - ***двойно логаритмична***.

При статистическия анализ се налага съставянето на диаграми, представящи нагледно сезонни колебания. За тази цел са подходящи ***радиалните диаграми***. Те се строят по полярната координатна система и затова се наричат още ***полярни диаграми***. Те могат да бъдат ***затворени*** и ***спирални***. Затворени са, когато се представят осреднени величини по месеци за 2 и повече години (вж. фиг. 10.2). По отдалечеността на линията от центъра за всеки месец (всеки лъч) се съди за степента и посоката на сезонните колебания. Спиралните диаграми се съставят, когато се изобразяват сезонните колебания за всеки месец поотделно за 2 или повече години. Тогава линията не е затворена, както във фигура 10.2, а образува спирала.

Сезонни колебания на продажбите
на стока "А" в град "Н" през 2003 - 2007 г.

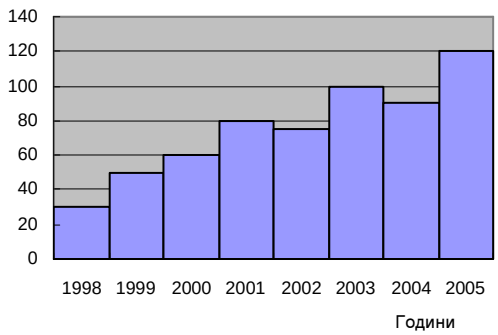


Фиг. 10.2

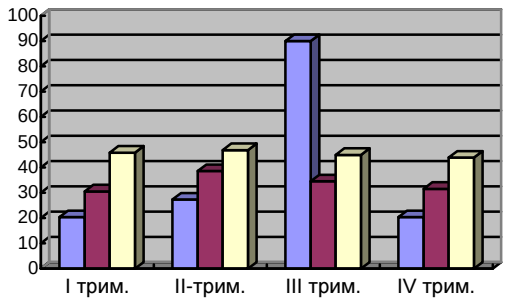
За представяне на сезонните колебания може да се състави и линейна диаграма в първия квадрант на координатната система, в която се показва сезонната вълна около определена базова линия (вж. фиг. 10.6)

Развитието на явленията може да се представи графично чрез плоскостни диаграми, от които най-подходящи са ***стъпаловидните диаграми***. Наричат се така, защото графичният образ се състои от

правоъгълници с различни височини, подредени като стълбове един до друг. Те могат да се съставят в различни варианти в зависимост от това, дали динамичният ред е моментен или периоден, дали той обхваща всички последователни подпериоди, или само избрани подпериоди, дали се представя само развитието на размера на явлението, или едновременно и неговата структура и т.н. На фиг. 10.3 и фиг. 10.4 са показани образци на два вида такива диаграми.



Фиг. 10.3



Фиг. 10.4

Популярна форма за нагледно представяне на развитие са ***картинните диаграми***. Обикновено това са подредени стилизирани фигури, които по асоциация се свързват с явленията. Приема се определен мащаб, според който всяка фигура отговаря на определен размер на представяните величини. Възможно е голямо разнообразие във формата на графичните изображения, особено когато се съставят с помощта на софтуерни приложения.[1]

10.3. Анализ на скоростта на развитието

Да се измери и анализира скоростта на развитието означава да се установи колко бързо (интензивно) се изменя абсолютният или средният размер на явлението от един период или момент до друг.

[1] Разнообразни по форма и предназначение графични изображения могат да се видят в публикациите на Националния статистически институт и в **Манов, А.** Статистика със SPSS, С., 2001, с. 353 и сл.

Общият абсолютен размер $\left(\sum Y\right)$ при периодни динамични редове е сума от абсолютните размери по подпериоди (Y_i), или

$$\sum Y = Y_1 + Y_2 + ... + Y_N \ . \qquad (10.1)$$

Средният абсолютен размер ($\overline{Y}$) се изчислява като хронологична средна аритметична величина.

В гл. 3 беше дадена обща характеристика на средните величини, в това число на същността, изчисляването и свойствата на аритметична средна величина. Когато обаче тази средна се прилага за изчисляване на среден размер при динамичния статистически анализ, възникват някои особености, които се отнасят по-конкретно за моментните динамични редове.

Ако редовете са периодни и ако абсолютните размери (членовете на редовете) за отделни подпериоди се означат с $Y_1, Y_2, ..., Y_N$, то средният размер, изчислен като ***хронологична аритметична средна*** ($\overline{Y}$), може да се представи с формулата:

$$\overline{Y} = \frac{\sum Y}{N}. \qquad (10.2)$$

Това е непретеглената форма на средната. Така например ще се изчисли средната месечна печалба на фирмата за дадена година от размера на печалбата (Y) през отделните месеци.

Ако абсолютният размер (в примера печалбата) се повтаря през някои от месеците и те се означат с t, средният размер може да се изчисли като претеглена средна по формулата:

$$\overline{Y} = \frac{\sum Yt}{\sum t}. \qquad (10.3)$$

Ако динамичният ред обаче е моментен, горните формули са неприложими. Нека са дадени например стоковите запаси на фирмата през четвъртото тримесечие на 2005 г.:

1 октомври 2005 г. - 600 хил. лв.

1 ноември 2007 г. - 1000 хил. лв.

1 декември 2007 г. - 1400 хил. лв.

1 януари 2007 г. - 1200 хил. лв.

При изчисляването на средния стоков запас за тримесечието не е възможно да се сумират запасите към определени моменти. Такова сумиране би било безсмислено. Необходим е друг подход. Възможно е, първо, да се изчислят средните за отделните месеци като полусбор от запасите в началото и в края на месеците. Тогава получените средни могат да се сумират и сумата да се раздели на месеците.

$$\overline{Y}_X = \frac{600+1000}{2} = 800 \text{ хил. лв.}; \qquad \overline{Y}_{XI} = \frac{1000+1400}{2} = 1200 \text{ хил. лв.};$$

$$\overline{Y}_{XII} = \frac{1400+1200}{2} = 1300 \text{ хил. лв.}$$

$$\overline{Y} = \frac{800+1200+1300}{3} = \frac{3300}{3} = 1100 \text{ хил. лв.}$$

Ако средните по месеци се заместят с моментните размери, от които са изчислени,

$$\overline{Y} = \frac{\frac{600+1000}{2} + \frac{1000+1400}{2} + \frac{1400+1200}{2}}{3} = \frac{\frac{600+2.1000+2.1400+1200}{2}}{3} =$$

$$= \frac{\frac{600}{2}+1000+1400+\frac{1200}{2}}{3} = \frac{3300}{3} = 1100 \text{ хил.лв.}$$

От описания подход, може да се изведе обща формула при N моментни величини:

$$\overline{Y} = \frac{\frac{Y_1+Y_N}{2} + \sum_{i=2}^{N-1} Y_i}{N-1}. \qquad (10.4)$$

По този начин се избягва предварителното изчисляване на средни за подпериодите.

Това е ***непретеглена хронологична аритметична средна*** при моментни редове, по която се изчислява ***средният абсолютен размер***, ако подпериодите са еднакви по продължителност (в примера месец).

Примерът може да се видоизмени:

1 януари 2007 г. - 600 хил. лв.

1 април 2007 г. - 1000 хил. лв.

1 септември 2007 г. - 1400 хил. лв.

1 януари 2007 г. - 1200 хил. лв.

Както в предходния пример, могат да се изчислят, първо, средни по подпериоди, които ще бъдат също 800, 1200 и 1300. При изчисляване на общата средна обаче изчислените средни по подпериоди трябва да се умножат (претеглят) по продължителността на подпериодите (в примера месеците) от един момент до друг:

$$\overline{Y} = \frac{800.3 + 1200.5 + 1300.4}{3 + 5 + 4} = \frac{13600}{12} = 1133 \text{ хил. лв.}$$

Ако се замести с приетите символи (Y и t) и се направят възможните заместявания, ще се получи формулата на ***претеглената хронологична аритметична средна***, по която се изчислява среден размер от моментен динамичен ред с различна продължителност на периодите от един момент до друг:

$$\overline{Y} = \frac{\sum_{i=1}^{N-1} Y_i t_i + \sum_{i=2}^{N} Y_i t_{i-1}}{2\sum_{i=1}^{N} t_i}. \qquad (10.5)$$

По данните от примера:

600.3 = 1800 хил. лв.	1000.3 = 3000 хил. лв.
1000.5 = 5000 хил. лв.	400.5 = 7000 хил. лв.
1400.4 = 5600 хил. лв.	1200.4 = 4800 хил. лв.
12400 хил. лв.	14800 хил. лв.

$$\bar{Y} = \frac{12400 + 14800}{2.12} = \frac{27200}{24} = 1133 \text{ хил. лв.}$$

Както се вижда, по този начин също се избягва предварителното изчисляване на средни по подпериоди.

10.3.1. Абсолютен прираст и среден абсолютен прираст

Скоростта на измененията на абсолютния, както и на средния размер, от един период (момент) до друг се характеризира чрез прирасти и темпове на растеж.

Абсолютният прираст (ΔY) е разликата между размера на явлението през даден период и друг период, приет за база. Той може да бъде положителна величина и да показва увеличение или отрицателна, показваща намаление (отрицателен прираст). Трябва при това да се има предвид, че терминът "абсолютен" не означава, че прирастът се отнася винаги само за абсолютни величини. Той може да характеризира изменението и на средни и относителни величини, например на средната производителност на труда.

При изчисляване на абсолютния прираст за базов период (момент) може да се приеме всеки предходен (Y_{i-1}) или първият в реда (Y_1). В някои случаи е целесъобразно за базов да се приеме характерен в някакво отношение период (момент), намиращ се извън дадения ред (Y_0).

Когато за базов се приеме всеки предходен период в реда, абсолютните прирасти (първите последователни разлики) се наричат ***верижни*** абсолютни прирасти и могат да се представят в обща формула:

$$\Delta Y_{i/t-1} = Y_t - Y_{t-1}. \quad (10.6)$$

Когато за базов се приема първият период в реда (Y_1), или намиращ се извън реда (Y_0), получават се абсолютни прирасти ***с постоянна база***:

$$\Delta Y_{t/1} = Y_t - Y_1. \quad (10.7)$$

$$\Delta Y_{t/0} = Y_t - Y_0. \quad (10.8)$$

Последният абсолютен прираст при приет за постоянна база първият член на реда ($\Delta Y_{N/1}$) изразява общия прираст за целия период ($Y_N - Y_1$), обхванат от динамичния ред. Той може да се получи и като сума от верижните абсолютни прирасти:

$$\Delta Y_{N/1} = Y_N - Y_1 = (Y_2 - Y_1) + (Y_3 - Y_1) + ... + (Y_N - Y_{N-1}) = \sum_{t=2}^{N} (Y_t - Y_{t-1}). \quad (10.9)$$

От друга страна, като разлика между даден и предходен прираст с постоянна база се получава съответен верижен прираст:

$$\Delta Y_{t/t-1} = (Y_t - Y_1) - (Y_{t-1} - Y_1). \quad (10.10)$$

Разликите между последователните верижни прирасти изразяват ***абсолютното ускорение*** (вторите последователни разлики) – ($\Delta^{(2)}Y_t$):

$$\Delta^{(2)}Y_t = \Delta Y_t - \Delta Y_{t-1} \ . \quad (10.11)$$

По смисъла на това определение ускорението може да бъде и отрицателно, т.е. да показва забавяне на развитието (на прирастите).

Средният абсолютен прираст за целия период, обхванат от динамичния ред, е средна аритметична величина на прирастите за отделните подпериоди:

$$\overline{\Delta Y} = \frac{\Delta Y_2 + \Delta Y_3 + ... + \Delta Y_N}{N-1}. \quad (10.12)$$

(Знаменателят е N - 1 тъй като броят на прирастите е с един по-малко от броя на абсолютните размери, от които са изчислени прирастите).

И тъй като сумата на прирастите за отделните подпериоди изразява общия прираст за целия период,

$$\overline{\Delta Y} = \frac{Y_N - Y_1}{N-1}. \quad (10.13)$$

За изчисляването на общия среден прираст по тази формула не е необходимо предварителното изчисляване на отделните прирасти по подпериоди.

10.3.2. Темпове на растеж и среден темп

Скоростта на развитието се характеризира най-често с темпове на растеж. Те са по форма динамични относителни величини, представяни в коефициент или процент.

Темпът на растеж (*T*) се получава като отношение на абсолютния размер в даден период (момент) към абсолютния размер в друг период (момент), приет за база. За редица последователни периоди темповете могат да бъдат с верижна база или с постоянна база.

Темповете с верижна база, наричани още верижни темпове ($T_{t/t\text{-}1}$), се изчисляват като размерът на явлението във всеки подпериод (Y_t) се раздели на размера в предходния подпериод ($Y_{t\text{-}1}$):

$$T_{t/t-1} = \frac{Y_t}{Y_{t-1}}. \qquad (10.14)$$

Темпът, представен по този начин, показва ***колко пъти*** размерът в дадения подпериод е по-голям или по-малък от размера през предходния подпериод, приет за единица. Ако се умножи по сто, темпът се представя в процент, т.е. базата се приема за 100.

Темповете с постоянна база се изчисляват по същия начин, но за база се приема или размерът на явлението през първия подпериод (Y_1), или през друг период извън реда (Y_0), който в някакво отношение е характерен и е целесъобразно да се приеме за базов:

$$T_{t/1} = \frac{Y_t}{Y_1}; \qquad (10.15)$$

$$T_{t/0} = \frac{Y_t}{Y_0}. \qquad (10.16)$$

Между темповете с верижна и с постоянна база има връзка, която позволява да се преминава от едните към другите.

От темпове с постоянна база се преминава към верижни, като всеки темп с постоянна база се дели на предходния темп с постоянна база.

От верижни темпове се преминава към темпове с постоянна база, като всеки верижен темп се умножава на всички предходни верижни темпове.

Скоростта на развитието може да се характеризира и с темпове на прираст.

Темповете на прираста (*T**) изразяват относителните прирасти на размерите на явленията през дадени подпериоди спрямо други, приети за база. Базата също може да бъде верижна или постоянна.

Темпът на прираста може да се изчисли по два начина.

а) като се раздели абсолютният прираст от единия период до другия на абсолютния размер през базовия период;

б) като се извади от темпа на растеж единица, респ.100:

$$(10.17) \qquad T^*_{t/t-1} = \frac{Y_t - Y_{t-1}}{Y_{t-1}} = \frac{Y_t}{Y_{t-1}} - 1;$$

$$(10.18) \qquad T^*_{t/1} = \frac{Y_t - Y_1}{Y_1} = \frac{Y_t}{Y_1} - 1;$$

$$(10.19) \qquad T^*_{t/0} = \frac{Y_t - Y_0}{Y_0} = \frac{Y_t}{Y_0} - 1.$$

Средната скорост на развитието на изследваните явления за определен период, съставен от подпериоди, се характеризира със среден темп на растеж, респ. среден темп на прираст.

Средният темп на растеж ($\overline{T}$) се изчислява обикновено като средна геометрична величина от верижните темпове за отделните подпериоди ($T_{t/t-1}$).

Известно е от гл. 3, че формулите на средната геометрична са

$\overline{x}_g = \sqrt[N]{\Pi x}$ (непретеглена) и

$\overline{x}_g = \sqrt[\sum f]{\Pi x^f}$ (претеглена),

в които x са осредняваните величини, N - броят на тези величини, а f - съответните тегла. Като се заместят x и f с приетите в тази глава символи, средният темп може да се запише с формулата:

$$\overline{T} = \sqrt[N-1]{\Pi T_{t/t-1}} \quad (10.20)$$

Коренният показател (N-1) е броят на темповете, който е с единица по-малък от броя на членовете на реда (подпериодите).

Тъй като произведението на верижните темпове дава темпа за целия период при база размерът на явлението през първия подпериод (първия член на реда), съгласно посочената връзка между верижни темпове и темпове с постоянна база, то

$$\overline{T} = \sqrt[N-1]{\Pi T_{t/t-1}} = \sqrt[N-1]{T_{N/1}} = \sqrt[N-1]{\frac{Y_N}{Y_1}} \quad (10.21)$$

Средният темп на прираста се намира като от средния темп на растеж се извади единица, респ. 100, когато е изразен в процент.

Когато са изчислени средни темпове за различни подпериоди (k на брой), които имат различна продължителност (t_i), общият среден темп ($\overline{T}$) трябва да се изчисли като претеглена средна геометрична величина по формулата:

$$\overline{T} = \sqrt[\sum t]{\overline{T}_1^{t_1}.\overline{T}_2^{t_2}.....\overline{T}_k^{t_k}} \quad (10.22)$$

Разглежданият ***средногеометричен темп*** отговаря на развитие по геометрична прогресия. Той по-конкретно показва с какъв среден темп явлението достига от един начален размер в първия подпериод до размера си в крайния подпериод, независимо от конфигурацията на развитието в останалите подпериоди. Очевидно е от формула 10.21, че

$$Y_N = Y_1.\overline{T}^{N-1} \quad (10.23)$$

Обикновено при динамичния анализ е точно такава постановката на познавателната задача. Затова средногеометричният темп намира широко практическо приложение. Може обаче в определени случаи да възникне необходимост от изчисляване на средни темпове при друга

изходна постановка. Например при определящо свойство общият (сумарният) абсолютен размер на явлението за целия период, състоящ се от подпериоди, или среден темп, който да отразява изменението на размера на явлението през всички отделни подпериоди. Затова в литературата има опити за обосноваване и на други подходи, например за изчисляване на среднокумулативен (среднопараболичен, средно-полиномен) и на средноекспоненциален темп.[1]

10.4. Анализ на тенденцията (тренда) на развитието

Беше посочено в т. 10.1, че в методологичен план развитието може да се разглежда като функция от времето и че размерът на явлението във всеки даден отрязък от време може условно да се разлага на съставни компоненти. Когато ни интересува основната, трайната тенденция, наричана ***тренд***, трябва да се елиминират останалите компоненти. За целта се прилага процедура, наречена ***изравняване*** или ***изглаждане*** на динамичните редове. Чрез изравняването се преодоляват колебанията и графичният образ на развитието се превръща от начупена в гладка линия с определена форма, описваща тренда.

Трендът изразява действието на трайни, систематични причини, определящи основната закономерност в развитието. Да се разложи това развитие и да се опише тренда означава то ***да се моделира***, като се представи такова, каквото би било без отклоненията около основната тенденция. Като всяко моделиране, то допуска определена условност. Не може да се твърди, че след като в резултат на изравняването вместо емпиричните стойности (членове) на динамичния ред се получават нови, изравнени ($\hat{Y}$), че те са точният реален резултат от трайно действуващите причини. Те са ***оценки***, описващи основната тенденция, общата закономерност. Освен това тези оценки зависят от възприетия метод за изравняване.

Необходим е внимателен подход при изравняването и особено при интерпретирането на резултатите от него. Съществува риск чрез

[1] Вж. **Кандиларов, Г.** и **А. Димитров,** Методология и таблици за изчисляване на средни темпове на икономически процеси. С., 1984; **Казинец , Л.С.,** Темпы роста и структурные сдвиги в экономике, М., 1986.

неподходящо изравняване да се представят в недействителна, изопачена форма истинските тенденции или да се третират като случайни такива изменения, които в действителност са закономерни и вътрешно присъщи на явлението. Поради това и в тази област се налага с особена сила необходимостта от качествен анализ, от предварителни знания относно същността и вътрешните закономерности на интересуващите ни явления.

Има разработени различни методи за изравняване на динамични редове - от елементарни, служещи за най-общо ориентиране относно трайната тенденция, до прецизни, изградени върху строги принципи и изискан математически апарат. Тук ще бъдат разгледани някои от тях.

10.4.1. Изравняване посредством среден прираст и среден темп

Когато има достатъчно основание да се твърди (главно въз основа на предварително съставена диаграма), че основното направление на развитието има характер на изменение в аритметична прогресия, редът може да се изравни посредством средния прираст, който беше разгледан в т. 10.3.1. При такова развитие първите последователни разлики са постоянни, т.е. еднакво е нарастването (или намаляването) от един подпериод до друг и то е равно на средния прираст.

Ако размерът на явлението в отделните подпериоди е $Y_1, Y_2, ..., Y_N$ и средният прираст е $\overline{\Delta}Y$, изравнените размери (членове на реда) ще бъдат $\hat{Y}_1, \hat{Y}_2, ..., \hat{Y}_N$ и ще се намерят така:

$$
\left|\begin{array}{l}
\hat{Y}_1 = Y_1 \\
\hat{Y}_2 = \hat{Y}_1 + \overline{\Delta}Y \\
\hat{Y}_3 = \hat{Y}_2 + \overline{\Delta}Y = \hat{Y}_1 + 2\overline{\Delta}Y \\
\ldots\ldots\ldots \\
\hat{Y}_N = \hat{Y}_{N-1} + \overline{\Delta}Y = \hat{Y}_1 + (N-1)\overline{\Delta}Y .
\end{array}\right. \qquad (10.24)
$$

Ако развитието в основната си тенденция протича като геометрична прогресия, изравняването може да се направи чрез средния геометричен темп ($\overline{T} = \sqrt[N-1]{\frac{Y_N}{Y_1}}$) по следния начин:

$$(10.25) \quad \left| \begin{array}{l} \hat{Y}_1 = Y_1 \\ \hat{Y}_2 = \hat{Y}_1.\overline{T} \\ \hat{Y}_3 = \hat{Y}_2.\overline{T} = \hat{Y}_1.\overline{T}^2 \\ ..... \\ \hat{Y}_N = \hat{Y}_{N-1}.\overline{T} = \hat{Y}_1.\overline{T}^{N-1}. \end{array} \right.$$

При достатъчно основания да се предполага, че и в бъдеще (след последния подпериод в изравнявания ред) развитието ще следва същата тенденция, неговият размер в следващите подпериоди може да се прогнозира чрез средния прираст или средния темп като се продължава динамичния ред до желания подпериод в бъдеще.

Трябва да се имат предвид отбелязаните вече особености на средния прираст и на средния геометричен темп, а именно, че зависят от двата крайни члена на реда - Y_1 и Y_N. От тях следователно ще зависи и всеки от изравнението членове ($\hat{Y}$). Поради това този метод не гарантира надеждни резултати, когато крайните членове се отклоняват значително от общата тенденция или по-общо казано, когато тази тенденция не съответствува по своята форма на развитие по аритметичната или геометричната прогресия.

10.4.2. Изравняване по метода на верижните средни

Изравняването по този метод се извършва посредством изчисляване на средни аритметични величини от определен брой членове на изходния динамичен ред, като последователно се изпуска един и се прибавя следващият. Основание за такъв подход е обстоятелството, че чрез осредняването се елиминира в задоволителна степен влиянието на случайните фактори, предизвикващи колебания около тренда. Ако напри-

мер предварително е установено, че в рамките на по-голям период, обхващащ три от по-малките подпериоди, се получава задоволително неутрализиране на случайните колебания, могат да се изчислят средни аритметични от по три стойности на *Y*:

(10.26)

$$\begin{array}{lll} t_1 & Y_1 & - \\ t_2 & Y_2 & \hat{Y}_2 = \dfrac{Y_1 + Y_2 + Y_3}{3} \\ t_3 & Y_3 & \hat{Y}_3 = \dfrac{Y_2 + Y_3 + Y_4}{3} \\ \ldots\ldots\ldots\ldots & & \\ t_{N-1} & Y_{N-1} & \hat{Y}_{N-1} = \dfrac{Y_{N-2} + Y_{N-1} + Y_N}{3} \\ t_N & Y_N & - \end{array}$$

При изравняване по този метод практически е целесъобразно верижните средни да се изчисляват за нечетен брой членове, тъй като всяка средна се отнася за определен подпериод от първоначалния ред. Това правило не винаги може да се спази, тъй като други съображения могат да наложат верижните средни да се изчислят за четен брой членове на реда. В такъв случай те попадат между периодите. Тогава се налага от получените средни да се изчислят нови от по 2 верижни средни, за да се центрират към съответните подпериоди. Често този въпрос се решава и по друг начин: взема се половината от първия и от последния член за периода и се прибавят останалите членове при изчисляването на всяка верижна средна. Така средните се центрират за подпериода, намиращ се в средата.

Центрираните подвижни средни от 4 члена например ще бъдат:

(10.27)
$$\left|\begin{array}{l} \hat{Y}_3 = \dfrac{\dfrac{Y_1}{2} + Y_2 + Y_3 + Y_4 + \dfrac{Y_5}{2}}{4} \\ \hat{Y}_4 = \dfrac{\dfrac{Y_2}{2} + Y_3 + Y_4 + Y_5 + \dfrac{Y_6}{2}}{4} \\ \ldots\ldots\ldots\ldots \\ \hat{Y}_{N-2} = \dfrac{\dfrac{Y_{N-4}}{2} + Y_{N-3} + Y_{N-2} + Y_{N-1} + \dfrac{Y_N}{2}}{4}. \end{array}\right.$$

Когато изчислените верижни средни не представят достатъчно добре трайната тенденция, може да се повтори същата процедура с получените верижни средни, т.е. да продължи изравняването. Това всъщност означава верижните средни да се изчислят като претеглени със симетрично разположени тегла по такъв начин, че да се придава по-голямо тегло на стойностите, които са по-близо до средата, а по малки тегла - на крайните стойности. Ако например от тричленните верижни средни се изчислят нови тричленни верижни средни, ще се получи:

(10.28)
$$\hat{\hat{Y}} = \frac{\hat{Y}_2 + \hat{Y}_3 + \hat{Y}_4}{3} = \frac{\dfrac{Y_1 + Y_2 + Y_3}{3} + \dfrac{Y_2 + Y_3 + Y_4}{3} + \dfrac{Y_3 + Y_4 + Y_5}{3}}{3} =$$
$$= \frac{Y_1 + 2Y_2 + 3Y_3 + 2Y_4 + Y_5}{9}.$$

Може да се предполага, че когато верижните средни обхващат повече подпериоди, редът се изравнява по-добре. С увеличаване на броя на подпериодите обаче се скъсява изравненият ред. Поради това методът на верижните средни се прилага обикновено при достатъчно дълги редове, например такива, които съдържат месечни данни за две или повече години. Освен това, от описанието на метода се вижда, че той се прилага без оглед на формата на трайната тенденция (тренда) и поради това не винаги е достатъчно прецизен. Не е пригоден и за прогнозиране на развитието на изследваните явления. Въпреки това, когато не се

стремим към точно моделиране на развитието, а към приблизително елиминиране на колебанията и очертаване на тенденцията, методът е приемлив и практически удобен. Такъв е например случаят при измерване на сезонните колебания.

10.4.3. Графичен метод на изравняване

Този метод, наричан още метод на свободната ръка, се основава на линейна диаграма, представяща развитието на изследваното явление. Линията, която изобразява емпиричния динамичен ред, обикновено е начупена, но показва определено направление на развитието. Свободно, по субективна преценка, се прекарва права или гладка крива линия, която минава между емпиричните точки така, че положителните отклонения да бъдат приблизително равни на отрицателните (по окомерна преценка). След това емпиричните точки се пренасят по перпендикулярите върху прекараната гладка линия. Оттам те се проектират на ординатната скала, където се отчитат числовите стойности на изравнения ред ($\hat{Y}$). За да се получат по-точни резултати, трябва да се използва милиметрова хартия за чертежи.

Изравняването по графичния метод не изисква изчисления и се извършва лесно. Освен това едновременно с намирането на изравнения ред се получава и диаграмата, която онагледява фактическото развитие и основната му тенденция. Негов недостатък е субективното определяне на мястото на линията, описваща тенденцията, от което зависят и числовите стойности на изравнения ред.

10.4.4. Изравняване по метода на най-малките квадрати

Посоченият по-горе недостатък на графичния метод се преодолява с приложението на ***аналитичния метод***, основаващ се върху математическия ***метод на най-малките квадрати***.

Ако разглеждаме развитието като функция от времето, можем да изберем съответен математически израз на функцията, по която да се опише основната тенденция. Методът на най-малките квадрати

удовлетворява изискването сумата от квадратите на разликите между първоначалните и изравнените стойности на реда да е минимум, т.е. $\sum\left(Y-\hat{Y}\right)^2=\min$. Тук подходът е аналогичен на този, прилаган при регресионния анализ, като вместо факторния признак *x* се включва времето *t*.

Ако се приеме, че трайната тенденция (трендът) се описва добре от права линия, трябва да се приложи уравнение на права (линейна функция), която в случая ще приеме вида:

(10.29) $$Y=a+bt\,.$$

За да се намерят *a* и *b*, съставя се система от две нормални уравнения:

(10.30) $$\left|\begin{array}{l}\sum Y=Na+b\sum t\\ \sum Yt=a\sum t+b\sum t^2\,.\end{array}\right.$$

Изчисляването на параметрите *a* и *b* е възможно и без съставяне на система от уравнения, като се приложат следните формули (изведени от системата):

(10.31) $$a=\frac{\sum Y\sum t^2-\sum Yt\sum t}{N\sum t^2-\left(\sum t\right)^2};$$

(10.32) $$b=\frac{N\sum Yt-\sum Y\sum t}{N\sum t^2-\left(\sum t\right)^2}\,.$$

След като *a* и *b* са известни, те се заместват в изходното уравнение (10.29) и се получава уравнението на тренда (линейният трендов модел):

(10.33) $$\hat{Y}=a+bt\,.$$

По това уравнение може за всяка стойност на *t* да се намерят изравнените стойности на реда ($\hat{Y}$). Тези стойности трябва да се разглеждат като ***оценки*** на *Y*, т.е. такива стойности, каквито би имал динамичният ред, ако явлението се развива плавно, без колебания.

Коефициентът *b* е аналогичен на *b* в регресионния модел. Той показва с колко единици (според приетата мярка) се изменя *Y* при

изменение на времето с една единица (година, месец и др.). Казано по друг начин, това е постоянният прираст на Y за единица време. Като правим тази аналогия с регресионния анализ, това не означава, че времето само по себе си е фактор за развитието на изследваните явления. При изравняването на динамичните редове не се интересуваме от факторите, които обуславят една или друга динамика на явленията. Интересува ни само каква е формата на трайната тенденция, обусловена от онези вътрешни причини, които са двигатели на развитието.

Трендовите модели могат да служат за прогнозиране на вероятните бъдещи изменения на даденото явление, ако има основание да се предполага, че ще се запази в бъдеще моделираната тенденция. Получените чрез тези модели прогнози се наричат ***екстраполационни***, тъй като се екстраполира бъдещето въз основа на миналото.

Изчислителните операции за намиране на a и b на трендовия модел могат значително да се опростят, като се приложи ***съкратен метод***, изведен от система (10.30).

Ако се положи $\sum t = 0$, първото нормално уравнение ще приеме вида $\sum Y = Na$, а следователно

$$a = \frac{\sum Y}{N}, \qquad (10.34)$$

т.е. a е средната аритметична величина на Y.

Второто уравнение съответно приема вида $\sum Yt = a\sum t^2$, откъдето следва, че

$$b = \frac{\sum Yt}{\sum t^2}. \qquad (10.35)$$

За да се получи $\sum t = 0$, необходимо е подпериодите (t) да се номерират от средата на реда в двете посоки от 0 до N и от 0 до -N. Това всъщност означава да се измести координатното начало в средата на реда, при което половината от значенията на t са положителни, а другата половина - отрицателни. Когато редът се състои от нечетно число членове, t ще приема стойност 0 в средата, а в двете посоки 1, 2, 3 и т.н и -1, -2, -3 и т.н. Когато редът е с четно число членове, координатното

начало попада между двата члена, намиращи се в средата. Тогава *t* ще приема значения 1, 3, 5 и т.н. и съответно -1, -3, -5 и т.н.

Ще илюстрираме изравняването по метода на най-малките квадрати (съкратен вариант) с ***примера***, съдържащ се в табл. 10.1.

Таблица 10.1

Производство на стока "А" през 2007 г.
във фирма "Н" по месеци
(изравняване по метода на най-малките квадрати)

Месеци	t	Y (хил. бр.)	Yt	t^2	$\hat{Y}$ (хил. бр.)
януари	-11	170,4	-1874,4	121	175,2
февруари	-9	171,0	-1539,0	81	173,1
март	-7	177,2	-1240,4	49	170,9
април	-5	171,6	-858,0	25	168,8
май	-3	172,0	-516,0	9	166,7
юни	-1	165,6	-165,6	1	164,6
юли	1	158,8	158,8	1	162,5
август	3	160,8	482,4	9	160,3
септември	5	154,8	774,0	25	158,2
октомври	7	145,8	1020,6	49	156,1
ноември	9	153,2	1378,8	81	154,0
декември	11	161,0	1771,0	121	151,8
		1962,2	-607,8	572	1962,2

$$a = \frac{\sum Y}{N} = \frac{1962,2}{12} = 163,52;$$

$$b = \frac{\sum Yt}{\sum t^2} = \frac{-607,8}{572} = -1,06.$$

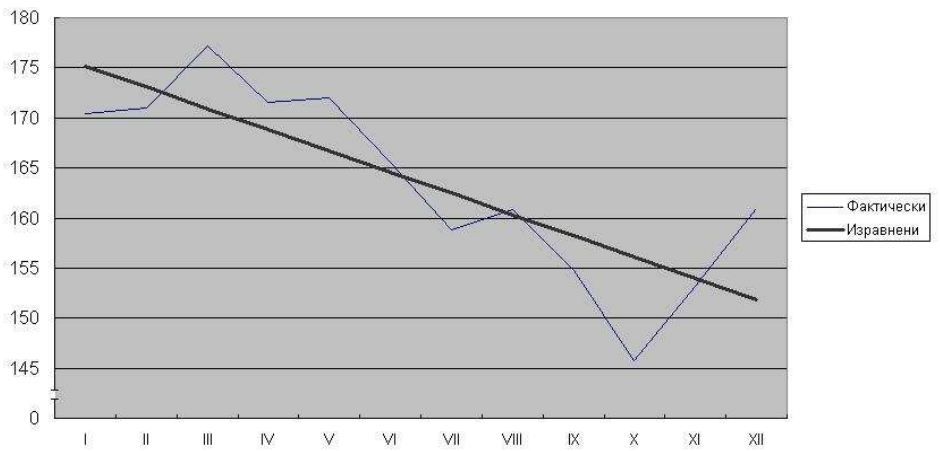
$$\hat{Y}_i = 163,52 + (-1,06)t;$$

$$\hat{Y}_1 = 163{,}52 + (-1{,}06)(-11) = 163{,}52 + 11{,}66 = 175{,}18;$$
$$\hat{Y}_2 = 163{,}52 + (-1{,}06)(-9) = 163{,}52 + 9{,}54 = 173{,}06,$$

и т.н.

Изравнените стойности на всички членове на динамичния ред са записани в последната колона на табл. 10.1. Очевидно е, че $\sum Y = \sum \hat{Y}$, както трябва да бъде съгласно метода на най-малките квадрати. Ако изравненият ред се представи графично чрез линейна диаграма, тя ще има формата на права линия, минаваща между стойностите на изходния (неизравнения) ред така, че се получава $\sum (Y - \hat{Y})^2 = \min$ (вж. фиг. 10.5).

Производство на стока "А" през 2007 г. във фирма "Н" по месеци – фактически и изравнени стойности по метода на най-малките квадрати



Фиг. 10.5

Дотук приемахме, че трайната тенденция (трендът) се описва от права линия. Действително много явления проявяват такава тенденция и за тях линейният трендов модел е адекватен. Има обаче явления, чиято трайна тенденция има друга форма и изискват моделиране чрез други,

нелинейни функции. В някои случаи трендът има графичен образ парабола от втора степен

и неговият аналитичен модел е

(10.36) $\hat{Y} = a + bt + ct^2$.

В други случаи адекватният модел има формата на полулогаритмична функция

(10.37) $log\,\hat{Y} = a + bt$,

или полулогаритмична по отношение на *t*:

(10.38) $\hat{Y} = a + b\,log\,t$.

Възможни са още различни други модели, като

(10.39) $log\,\hat{Y} = a + b\,log\,t$;

(10.40) $\hat{Y} = ab^t$;

(10.41) $\hat{Y} = \frac{1}{a+bt}$ и др.

При всеки конкретен случай трябва да се избере подходяща функция. Изхождайки от своя опит и от някои общи правила, специалистът трябва да направи избора.

Първоначално ориентиране относно характера на линията, която описва тренда, може да се получи от линейната диаграма, изобразяваща емпиричните (неизравнените) данни. Освен това трябва да се имат предвид особеностите на отделните възможни функции.

Правата линия (линейният тренд) изразява нарастване по аритметична прогресия. Това означава, че прирастите (първите последователни разлики) са постоянни величини. При развитие във формата на парабола от втора степен прирастите на прирастите (вторите последователни разлики) са постоянни величини. При парабола от трета степен постоянни са третите последователни разлики и т.н. Изобщо при

тренд, представен с полином от *k*-та степен ($\hat{Y} = a + bt + ct^2 + ... + zt^k$ постоянни са *k*-тите последователни разлики.

Полулогаритмичният модел $log\,\hat{Y} = a + bt$ има графичен образ права линия, ако значенията на Y са представени в диаграма с логаритмична скала. По същия начин $log\,\hat{Y} = a + b\,log\,t$ приема формата на права при логаритмична скала за *t*.

Степенната функция $\hat{Y} = ab^t$ съответства на развитие по геометрична прогресия.

Има случаи, при които е трудно да се избере категорично като адекватна само една функция. Тогава могат да се "проиграят" две или повече, а след това въз основа на получените резултати да се прецени коя от тях е подходяща. При функции с еднакъв брой параметри, като критерий може да служи стандартната грешка на оценката $S_Y = \sqrt{\frac{\sum\left(Y - \hat{Y}\right)^2}{N}}$. Най-подходящ ще бъде този модел, при който S_Y е най-малко число.[1]

В предходното изложение се предполагаше, че наличието на тренд е безспорно. Когато той е силно подчертан, лесно може да се констатира с помощта на линейна диаграма. Тя, както беше посочено, служи и като помощно средство за ориентиране относно формата на тренда. Могат обаче да възникнат случаи, при които наличието или отсъствието на тренд не може да се установи визуално и е необходимо да се провери посредством определени критерии. За такива обикновено се използуват коефициентите на автокорелацията, сериите от отклонения около средната величина и др.

В развитието на някои явления през по-дълъг период могат да се появят обстоятелства и фактори, които да предизвикат съществени изменения в посоката и интензивността на тяхното изменение. В такива случаи не е целесъобразно да се изравнява целия динамичен ред по една функция. Полученият трендов модел не би изразил характера и направлението на трайната тенденция, щом тази тенденция се е

[1] Относно възможностите за използване на други критерии, за избор на модела и за проверка относно наличието на тренд вж.: **Величкова, Н.** Цит. съч., с.62 и сл.

променяла през отделните части на периода. По-целесъобразно е да се обособят характерните подпериоди с присъщите им тенденции. Известни възможности предлагат и разработените в математиката сплайн функции, които отчитат преходите от една тенденция към друга в рамките на един достатъчно дълъг период.

10.5. Измерване на сезонни колебания

Много явления, които се подлагат на статистическо изучаване, се намират под въздействието на условия и фактори със сезонен характер. Поради това в развитието им се наблюдават ежегодно повтарящи се колебания през отделни месеци или други подпериоди в рамките на годината. Те се наричат ***сезонни колебания***. Това не значи, че винаги са обусловени от смяната на годишните времена, от сезоните в обикновения смисъл на думата. Общо под сезонни колебания се разбират закономерно повтарящи се в определено време през годината изменения в размера на явленията. Такива са например колебанията по месеци в размера на някои производства на хранително-вкусовата промишленост, в продажбите и потреблението на някои стоки, в строителството, в раждания и др.

Измерването на сезонните колебания е необходимо не само за да се опознаят закономерностите и компонентите на развитието на явленията, но и от чисто практическа гледна точка с оглед предприемането на мерки за преодоляването им, когато е възможно и целесъобразно, или за съобразяване с тях при организирането на стопанската и друга дейност. В някои случаи при анализа е необходимо сезонните колебания да бъдат изучени с оглед да се елиминират като компонент на развитието, за да се проявят и измерят другите компоненти.

Разработени са различни методи за измерване на сезонните колебания, някои от които се разглеждат в следващото изложение.

10.5.1. Метод на обикновените средни

Този метод е приложим когато в динамичния ред се съдържат сезонни колебания, но по години не се проявява трайна тенденция (тренд)

на изменение в общия размер на явлението, т.е. когато редът е стационарен. В метода е заложена следната логика.

Ако в развитието на явлението няма възходяща или низходяща тенденция (тренд), няма и сезонни колебания, то неговият размер през отделни подпериоди ще се колебае около даден постоянен размер само под действието на случайни фактори (случайни колебания). Този постоянен размер може да се изчисли като средна аритметична величина, която елиминира случайните колебания в двете посоки. Ако обаче има влияние на сезонни фактори, те ще предизвикват повтарящи се през определени месеци (или други отрязъци от време) отклонения от постоянния среден размер. Ето защо разликите между фактическите размери по месеци (или други подпериоди) и общата средна аритметична ще се дължат на сезонните фактори. За да се отстранят случайните колебания, трябва да се образува динамичен ред с данни за няколко години, минимум три.

Тази логика подсказва конкретните процедури, които трябва да се изпълнят, за да се измерят сезонните колебания по разглеждания метод.

1. От данните за едноименните месеци за всички години се изчисляват помесечни средни аритметични величини ($\overline{Y_i}$). Ако например означим размерите на явлението през януари за *k* години с $Y_1^{(1)}, Y_1^{(2)}, ..., Y_1^k$, януарската средна величина за *k* години ще бъде

$$\overline{Y_I} = \frac{Y_1^{(1)} + Y_1^{(2)} + ... + Y_1^{(k)}}{k}.$$

По същия начин ще се намерят средните за февруари, март и т.н. до декември.

Може общо да се запише средната за *i*-тия месец с формулата

$$\overline{Y_i} = \frac{\sum_{i=1}^{k} Y_i}{k}. \qquad (10.42)$$

2. Изчислява се общата средна ($\overline{\overline{Y}}$) от помесечните средни:

$$(10.43) \qquad \overline{\overline{Y}} = \frac{\overline{Y}_1 + \overline{Y}_2 + ... + \overline{Y}_{12}}{12} = \frac{\sum_{i=1}^{12} \overline{Y}_i}{12}.$$

Тя може да се изчисли и от помесечните данни за отделните години, като общата им сума се раздели на общия брой на месеците за всички години (k.12).

3. Намират се разликите между помесечните средни ($\overline{Y}_i$) и общата средна ($\overline{\overline{Y}}$), които измерват ***абсолютните сезонни колебания***:

$$(10.44) \qquad S_i = \overline{Y}_i - \overline{\overline{Y}}$$

4. Като се отнесат тези разлики (абсолютни сезонни колебания) към общата средна и се умножат по 100, получават се ***относителните сезонни колебания*** в процент:

$$(10.45) \qquad S_{i(\%)} = \frac{\overline{Y}_i - \overline{\overline{Y}}}{\overline{\overline{Y}}} 100 = \frac{\overline{Y}_i}{\overline{\overline{Y}}} 100 - 100.$$

Те показват с колко процента се отклоняват в едната или в другата посока помесечните средни от общата средна под влияние на сезонните фактори.

5. Ако помесечните средни се разделят на общата средна, ще се получат ***индексите*** на сезонните колебания (също могат да се представят в процент):

$$(10.46) \qquad I_i = \frac{\overline{Y}_i}{\overline{\overline{Y}}} 100.$$

Въпрос на предпочитание е използуването на относителните сезонни колебания ($S_{i(\%)}$) или индексите на сезонните колебания (I_i). Разликата е само в това, че първите измерват колебанията спрямо база нула, а вторите - при база единица (или 100).

6. За да се представи нагледно ***сезонната вълна***, съставя се линейна диаграма.

Ще илюстрираме метода с ***пример***. Да приемем, че ни интересува сезонността в продажбите на стока А в един град. Дадени са данни за

продадените количества по месеци общо за 2005, 2006 и 2007 г. Тези данни, както и изчисленията по описания начин се съдържат в табл. 10.2.

Таблица 10.2

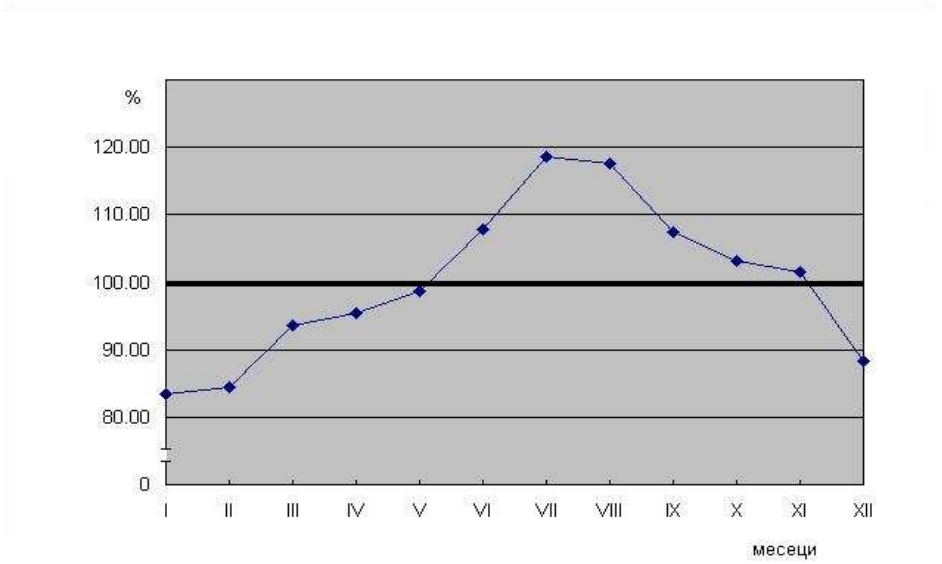
Сезонни колебания в продажбите на стока "А" в град "Н" през 2005-2007 г.

Месеци	Прод. кол. през трите години, хил. бр.	Поме-сечни средни хил. бр.	Сезонни колебания		
			абсолютни хил. бр.	относителни %	индекси %
	$\sum Y_i$	$\overline{Y_i} = \frac{\sum Y_i}{3}$	$S_i = \overline{Y_i} - \overline{\overline{Y}}$	$S_{i(\%)} = \frac{\overline{Y_i} - \overline{\overline{Y}}}{\overline{\overline{Y}}} 100$	$I_i = \frac{\overline{Y_i}}{\overline{\overline{Y}}} 100$
I	2010	670	-132	-16,5	83,5
II	2031	677	-125	-15,6	84,4
III	2253	751	-51	-6,4	93,6
IV	2295	765	-37	-4,6	95,4
V	2373	791	-11	-1,4	98,6
VI	2595	865	63	7,9	107,9
VII	2853	951	149	18,6	118,6
VIII	2829	943	141	17,6	117,6
IX	2583	961	59	7,4	107,4
X	2484	828	26	3,2	103,2
XI	2442	814	12	1,5	101,5
XII	2124	708	-94	-11,7	88,3
Обща средна	28872	9624			

В примера приемаме, че е налице условието за приложимост на метода -стационарен динамичен ред за целия период, т.е липса на трайна тенденция на възходящо или низходящо развитие по години.

Изчислените характеристики на сезонните колебания и диаграмата (фиг. 10.6.) показват, че в продажбите (потребителското търсене) на дадената стока има подчертана сезонност. Най-голямо положително сезонно отклонение има през юли (18,6 %), а най-голямо отрицателно отклонение - през януари (-16,5 %).

Индекси на сезонните колебания в продажбите на стока "А" през 2005 – 2007 г.



Фиг. 10.6.

Ще допълним, че разглежданият метод е неприложим когато освен сезонни колебания в развитието на явлението по години се съдържа трайна тенденция (тренд). Не е приложим, защото в такъв случай разликите между помесечните средни и общата средна ще съдържат не само сезонни колебания, но и съответна част от нарастването от година на година.

10.5.2. Метод на отношенията на фактическите към изравнените стойности

Когато има тенденция на нарастване или намаляване на размера на явлението по години (нестационарен динамичен ред), трябва да се приложи метод, който елиминира влиянието на тази тенденция. Логично е да се твърди, че влиянието на сезонността ще се проявява като колебание не около постоянната средна величина ($\bar{\bar{Y}}$), а около тренда. Необходимо е следователно фактическият размер по месеци (или други подпериоди) да се сравнява с оня размер, който явлението би имало, следвайки основната тенденция без колебания (с оценките $\hat{Y}$).

Приложението на метода се свежда до изпълнението на следните процедури:

1. Динамичният ред, съдържащ данни по месеци или други подпериоди за 3, 4 или повече години се изравнява и по този начин се получават оценките $\hat{Y}$. Изравняването се извършва най-често по метода на подвижните (плъзгащи се) средни, обхващащи по 12 месеца , като се центрират по описания вече начин. При данни по месеци първата средна ще се отнася за юли и ще се намери по формулата:

$$\hat{Y}_7 = \frac{\frac{Y_1}{2} + Y_2 + ... + Y_{12} + \frac{Y_{13}}{2}}{12}. \qquad (10.47)$$

По същия начин се намират всички следващи подвижни средни.

(Изравняването може да се направи и по метода на най-малките квадрати).

2. Намират се по месеци разликите между фактическите и изравнените стойности ($Y_i - \hat{Y}_i$). Така се получават абсолютните отклонения.

3. Като се отнесат получените разлики към изравнените стойности, ще се получат относителни отклонения $\left(\frac{Y_i - \hat{Y}_i}{\hat{Y}_i}\right)$, които могат да се представят в процент, като се умножат по 100. Ако фактическите

стойности се разделят на изравнените, ще се получат индекси $\left(\frac{Y_i}{\hat{Y}_i}\right)$, които също могат да се изразят в процент.

4. За да се елиминират случайните колебания и се получат 12 величини (по месеци), измерващи сезонните колебания средно за всички разглеждани години, следва абсолютните и относителните отклонения и съответно индексите да се осреднят по едноименни месеци за всичките години по формулите:

$$S_i = \frac{\sum_{i=1}^{k}\left(Y_i - \hat{Y}_i\right)}{k}; \qquad (10.48)$$

$$S_{i(\%)} = \frac{\sum_{i=1}^{k}\left(\frac{Y_i - \hat{Y}_i}{\hat{Y}_i}\right)}{k}; \qquad (10.49)$$

$$I_i = \frac{\sum_{i=1}^{k}\frac{Y_i}{\hat{Y}_i}}{k}. \qquad (10.50)$$

Ще илюстрираме метода с ***пример*** (табл. 10.3). Тъй като процедурите по изравняването са известни, дадени са разликите между фактическите и изравнените стойности, от които са изчислени относителните сезонни колебания.

Таблица 10.3

Сезонни колебания в продажбите на стока "А"
в град "Н" през 2005 - 2007 г.

Месеци	Разлики между фактическите и изравнените стойности, хил. бр. $(Y_i - \hat{Y}_i)$			Относителни разлики между фактическите и изравнените стойности, % $\left(\frac{Y_i - \hat{Y}_i}{\hat{Y}_i}\right)100$			Относителни сезонни колебания $S_{i(\%)} = \frac{\sum_{i=1}^{k}\left(\frac{Y_i - \hat{Y}_i}{\hat{Y}_i}\right)}{k}100$
	2005	2006	2007	2005	2006	2007	
I	-135	-109	-113	-17,65	-13,81	-13,90	-15,1
II	-125	-118	-99	-16,30	-14,92	-12,15	-14,5
III	-19	-41	-66	-2,47	-5,47	-8,08	-5,3
IV	-23	-22	-45	-2,98	-2,77	-5,50	-3,8
V	7	-4	-21	0,91	-0,50	-2,56	-0,7
VI	75	71	52	9,68	8,89	6,32	8,3
VII	166	153	131	21,36	19,10	15,88	18,8
VIII	157	138	125	20,15	17,19	15,12	17,6
IX	39	75	54	4,99	9,32	6,51	6,9
X	29	29	5	3,70	3,59	0,60	2,6
XI	-15	-25	55	-1,91	-3,09	6,60	-6,9
XII	-67	-111	-131	-8,51	-13,69	15,69	-10,1

Получените резултати от изчисленията показват, че най-голямо е положителното сезонно отклонение през юли (18,8%), а най-голямото отрицателно отклонение - през януари (-15,1%). Представени на линейна диаграма, месечните сезонни колебания ще опишат общо формата на сезонната вълна.

10.5.3. Обобщаващи измерители на сезонността

Сезонните колебания, изчислени по разгледаните или по други методи, характеризират силата на влияние на сезонните фактори през всеки месец (или друг подпериод). Те обаче не дават обобщена характеристика на сезонността, т.е. не показват в каква степен изучаваното явление се намира под въздействието на сезонни фактори. Когато липсва такава обобщаваща характеристика трудно могат да се преценят измененията, които настъпват в сезонността във времето или различията в пространството (в териториален разрез).

Един възможен обобщаващ измерител на сезонността е ***коефициентът на вариацията на помесечните индекси на сезонните колебания*** (V_s), изчислен по средното аритметично (линейно) отклонение:[1]

$$V_S = \frac{\sum_{i=1}^{12} \left| I_i - \bar{I} \right|}{12}, \qquad (10.51)$$

където $\bar{I}$ е средноаритметичният (средномесечен) индекс на сезонните колебания.

По величината на този коефициент, изчислен за различни периоди, може да се съди дали се засилва или затихва сезонността (проявлението на сезонните фактори).

Възможен е и друг подход. Има основание разпределението по месеци на общия размер ($\sum Y$) на изследваното явление да се разглежда като хронологична структура, изразена с относителни дялове на месечните размери в общия размер за целия изследван период. Ако явлението не показва трайна тенденция (няма тренд) по години, но няма и сезонни колебания, то всеки месец би имало $^1/_{12}$ част от общия размер, т.е. би била налице равномерна хронологична структура. Сезонните фактори (ако съществуват) ще предизвикват неравномерно разпределение по месеци (неравномерна структура). Тази логика дава основание да се

[1] Вж. **Венецкий, И. Г.** Статистические методы в демографии. М., 1977, с. 28

конструира ***интегрален коефициент на сезонността*** по аналогия на интегралния коефициент на неравномерността на структурите (вж. гл. 9).

Ако фактическите относителни дялове на помесечните средни в общата им сума се означат с v_i, а относителните дялове на равномерната структура са $^1/_{12}$, може да се съставят следните формули на коефициента на сезонността (K_{sv}) по аналогия на формули 9.17 и 9.18 от гл. 9.

$$(10.52) \qquad K_{SW} = \sqrt{1 - \frac{2}{1 + 12\sum_{i=1}^{12} v_i^2}}.$$

Ако v_i е в процент, горната формула ще има следния вид:

$$(10.53) \qquad K_{SW} = \sqrt{1 - \frac{20000}{10000 + 12\sum_{i=1}^{12} v_i^2}}.$$

При положение, че случайните колебания са елиминирани чрез помесечното осредняване за 3, 4 или повече години, ако няма сезонни колебания $K_{SW} = 0$. Колкото по-силно е изразена сезонността, толкова повече K_{SW} ще бъде по-близък до единица. (Решен пример се съдържа в т. 10.7).

Сезонните колебания не се проявяват винаги с постоянна амплитуда. Освен това с течение на времето може да настъпва изместване на техния минимум и максимум. Това налага при емпиричните изследвания, обхващащи повече и по-продължителни периоди, не само да се измерват сезонните колебания, но да се проследят и измененията в амплитудата им и в общата конфигурация на сезонната вълна.

10.6. Измерване на циклични колебания

Разгледаните сезонни колебания са краткотрайни, проявяват се в рамките на годината, с постоянна или променяща се амплитуда през отделните години. Разглеждани обаче в по-дълъг период по години, редица явления протичат с повтарящи се цикли. В средно срочен обхват

те имат характер на ***конюнктурни колебания*** около тренда, а в дългосрочен - като ***"дълги вълни"*** ("Кондратиеви вълни"; "*К*-вълни")[1].

Конюнктурните циклични колебания (бизнес цикли) обхващат обикновено 3-5 годишни периоди. Те се обуславят от различни причини и поради това не са еднакви по форма и интензитет.

Дългите вълни имат 50-60 годишен обхват. Обуславят се от крупни инвестиции и технологични промени, характеризирани като "технологични революции". Присъщи са не толкова на отделни страни, колкото общо на световната икономика. Един от основните проблеми при изследването им е свързан с осигуряването на сравнима информация за много дълги периоди.

По отношение на методологията за измерване на цикличните колебания, има разработени прецизни като математически апарат методи, наречени ***хармоничен (периодограмен) анализ*** и ***спектрален анализ***. Те обаче съдържат теоретични изисквания, които като правило не се удовлетворяват от реалното икономическо развитие. Поради това практически не намират широко приложение.[2]

Един възможен подход, който се прилага обикновено при измерването на конюнктурните циклични колебания (бизнес циклите), се извежда от постановката за разлагане на развитието на съставни компоненти. Ако приемем, че в динамичния ред от данни по години се съдържа тренд ($\hat{Y}$) и циклични колебания (C), то тези колебания ще се проявяват като отклонения от тренда на фактическите размери на явлението. Това означава, че при мултипликативния модел например $Y_i = \hat{Y}_i C_i$. Оттук следва, че $C_i = \frac{Y_i}{\hat{Y}_i}$. Това са относителни величини, наречени ***индекси на цикличните колебания*** (I_C). Те могат да се представят в процент и да се запишат с обща формула:

[1] Описани са от руския учен **Николай Кондратиев** (1892-1938г.). Вж. **Кондратьев, Н.** Большие циклы коньюнктуры. Във: Вопросы коньюнктуры. М., 1925, **Кондратьев, Н.** Проблемы экономической динамики. М., 1989; **Schumpeter, J.** Business Cycles. New York, 1939; **Rostow, W.** The World Economy, 1979.

[2] Относно същността и прилагането на тези методи вж. ***Величкова, Н***., Цит. съч., с. 184 и сл.

(10.54) $$I_C = \frac{Y_i}{\hat{Y}_i}100.$$

При адитивен модел на развитието $Y_i = \hat{Y}_i + C_i$. Оттук следва, че $C_i = Y_i - \hat{Y}_i$. Това са цикличните колебания в абсолютни величини. Като относителни величини спрямо тренда, те могат да се изчислят по формулата:

(10.55) $$C_{i(\%)} = \frac{Y_i - \hat{Y}_i}{\hat{Y}_i}100 = \left(\frac{Y_i}{\hat{Y}_i} - 1\right)100.$$

Както се вижда, описаният метод за измерване на цикличните колебания е аналогичен на прилагания при измерване на сезонните колебания в рамките на годишни периоди като отношение (или разлика) между фактическите и изравнените членове на динамичния ред.

Чрез графично представяне на цикличните колебания се получава визуална представа за тяхната сила, за формата и обхвата на циклите. Когато се представят едновременно цикличните колебания в развитието на две или повече зависими помежду си явления, може да се види как цикличността в протичането на едно от тях влияе за формирането на цикли в другите.

10.7. Практикум

10.7.1. Въпроси за самопроверка

1. Какъв смисъл има при анализа разглеждането на развитието като функция от времето и условното му разлагане на компоненти?
2. Каква е разликата между прираст и темп?
3. Какъв е познавателния смисъл на средно геометричния темп?
4. Какво се разбира под тренд и какъв е основният принцип при неговото разкриване и моделиране?

5. При какво условие са приложими средният прираст и средният темп за изравняване на динамичните редове?
6. Как се изравняват редовете по метода на подвижните (верижните) средни?
7. Какво е основното изискване на метода на най-малките квадрати за моделиране на тренда?
8. Как се прави избора на адекватен трендов модел при "конкуриращи се" функции?
9. При какво условие е приложим методът на обикновените средни за измерване на сезонните колебания?
10. Как се измерват сезонните колебания при нестационарни динамични редове?
11. Как може да се изчисли обобщаващ измерител на сезонността?
12. Как могат да се измерят конюнктурните циклични колебания?

10.7.2. Задачи за упражнение

Задача 1. Дадени са данни за производството на продукт "А" във фирма "Н" през периода 2002 - 2007 г. Необходимо е да се изчислят:

а) средногодишното производство

б) верижните прирасти и прирастите спрямо 2000 г.

в) средният абсолютен прираст за целия период

г) верижните темпове на растеж и темповете с база 2000 г.

д) средногодишният темп на нарастване за периода

Изходните данни и част от резултатите от анализа се съдържат в следващата таблица.

Таблица 10.4

Производство на продукт “А” във фирма “Н” през 2002 – 2007 г.

Години	Произве-дена продукция-хил. бр.	Абсолютни прирасти-хил. бр.		Темпове на растеж - %	
		верижни	при база 2002 г.	верижни	при база 2002 г.
	Y_t	$\Delta Y_{t/t-1}$	$\Delta Y_{t/1}$	$T_{t/t-1}=\frac{Y_t}{Y_{t-1}}100$	$T_{t/1}=\frac{Y_t}{Y_1}100$
2002	250	-	-	-	100,0
2003	250	0	0	100,0	100,0
2004	260	10	10	104,0	104,0
2005	264	4	14	101,5	105,6
2006	262	-2	12	99,2	104,8
2007	264	2	14	100,8	105,6
	1550				

Решение:

а) Средногодишният размер на произведената продукция е

$$\overline{Y}=\frac{\sum Y}{N}=\frac{1550}{6}=258{,}3 \text{ хил. бр.}$$

б) Абсолютните прирасти са посочени в таблицата.

в) Средният абсолютен прираст е $\overline{\Delta Y}=\frac{Y_N-Y_1}{N-1}=\frac{14}{5}=2{,}8$ хил.бр.

г) Темповете на растеж могат да се изчислят от изходните данни (посочени в таблицата). Ако са изчислени единият ред темпове, другият може да се изчисли по правилото за преминаване от темпове с постоянна база към верижни и обратно.

От темповете с база предходната година (верижни) ще се изчислят темповете с база 2002 г., като всеки темп се умножава на всички предходни:

1,04.1,00 = 1,04 или 104,0% ;

1,015.1,04.1,0 = 1,056 или 105,6% ;

0,992.1,015.1,04.1,0 = 1,048 или 104,8% ;

1,008.0,992.1,015.1,04.1,0 = 1,056 или 105,6% .

От темповете с постоянна база 2002 г. ще се изчислят верижните, като всеки темп с постоянна база се дели на предходния:

$$\frac{104,0}{100,0}100 = 104\% ; \quad \frac{105,6}{104,0}100 = 101,5\% ;$$

$$\frac{104,8}{105,6}100 = 99,2\% ; \quad \frac{105,6}{104,8}100 = 100,8\% .$$

д) Средногодишният темп за целия период е:

$$\overline{T} = \sqrt[N-1]{\frac{Y_N}{Y_1}} = \sqrt[5]{1,056} = 1,011 \text{ или } 100,8 \% .$$

Следователно средногодишно производството е нараствало 1,011 пъти или с 1,1 %.

Задача 2. Дадени са данни за средните добиви от слънчоглед в една област през 1997-2007 г., съдържащи се в следващата таблица.

Таблица 10.5

Средни добиви от слънчоглед в област "А"
през периода 1997 – 2007 г.

Години	Среден добив - кг.
1997	160
1998	148
1999	172
2000	155
2001	185
2002	176
2003	190
2004	194
2005	200
2006	183
2007	197

Иска се:

1. Да се състави линейна диаграма за развитието на добивите.
2. Да се изчислят темповете на нарастване на добивите-верижни и при база 1997 г.
3. Да се изравни редът по метода на подвижните средни-чрез 4-годишни центрирани средни.
4. Да се състави линеен трендов модел.

Отговор по т. 4: $\hat{Y} = 178 + 4{,}5t$

Задача 3. Дадени са данни в следващата таблица.

Таблица 10.6

Относителен дял на средния брой на живородените деца по месеци в % от общия им брой в град "Б" през 2005 - 2007 г.

Месеци	Относителен дял - %	Квадрати на относителните дялове
	v_i	v_i^2
I	8,7	79,69
II	7,9	62,41
III	8,5	72,25
IV	8,3	68,89
V	8,6	73,96
VI	8,5	72,25
VII	9,1	82,81
VIII	8,8	77,44
IX	8,5	72,25
X	8,2	67,24
XI	7,4	54,76
XII	7,5	56,25
	100,0	836,20

Иска се да се изчисли интегрален коефициент на сезонността.

Решение:

$$K_{SW} = \sqrt{1 - \frac{20000}{10000 + 12\sum_{i=1}^{12} v_i^2}} = \sqrt{1 - \frac{20000}{10000 + 12.836{,}26}} = 0{,}045.$$

10.7.3. Легенда за екстраполацията на Диоген

В много случаи въз основа на определената при динамичния анализ скорост или основна тенденция на развитието на изследваното явление се предвижда (прогнозира) неговото бъдещо развитие, т.е. прави се екстраполационна прогноза. Без да си е служил с формули, древногръцкият философ ***Диоген*** очевидно е предвиждал чрез екстраполация.

Според легендата, Диоген се оттеглил далеч от Атина, за да се отдаде на своите размишления, като превърнал в жилище изоставена голяма бъчва. Един ден към бъчвата приближил пътник и се обърнал към Диоген с думите: "Кажи ми, мъдри старче, ще стигна ли до Атина преди залеза на слънцето?".

"Върви"-отвърнал Диоген. Пътникът не разбрал отговора и повторил въпроса си. "Върви"-още по-силно извикал Диоген. Тогава пътникът си тръгнал, а Диоген гледал известно време след него и извикал: "Върни се". Пътникът се върнал и озадачен попитал: "Защо преди малко ме изгони, а сега ме призова да се върна?"

Диоген отговорил: "А как бих могъл да ти кажа дали ще стигнеш до Атина преди залеза, докато не съм видял колко бързо ходиш? А сега ще ти кажа, че няма да стигнеш до Атина преди залеза на слънцето".[1]

[1] Цит. по: Занимательная статистика. М., 1980, с. 65

11. ИНДЕКСИ

> "По същество приложението на статистическите методи се заключава в използването на здравия смисъл при анализа на едни или други данни."
>
> Д. Вайнберг и Д. Шумекер

В тази глава се разглеждат същността, функциите и конкретните характеристики на индексния метод в статистиката, чиято практическа приложимост е доказана и безспорна, а областта на приложение е извънредно широка. Тя предлага не само теоретични знания, но и практически ориентири относно различните видове множествени (сложни, съвкупностни) индекси за характеризиране на динамиката на различни величини. Всеки, който ги владее, ще разбира и коректно ще използва различните форми на индекси на равнища (цени, заплати, производителност и др.) и ще интерпретира различията между тези индекси. Правилно ще си обяснява конструкцията на индексите на обеми и маси и връзките между тях. Би разбирал различието между индексите при постоянен и при променлив състав, връзките между тях и причините за възникването на привидни статистически парадокси. Той лесно би разбрал особеностите на териториалните индекси и някои възможни подходи при международни и други междурегионални сравнения.

11.1. Същност и функции на индексния метод

При изучаване на динамиката и на териториалните различия, както и при измерването на факторните ефекти, широко се използува ***индексният метод***. Обширна област на приложение има той при изучаване на икономическия растеж, на динамиката на цените, на производителността на труда, на рентабилността, на доходите и др., както и при измерване влиянието на факторите, обуславящи тази динамика.

Под ***индексен метод*** в статистиката трябва да се разбира системата от принципи, правила, схеми, формули и модели за измерване и логическо тълкуване на относителните различия по време, по място,

между фактически резултати и нормативи и на факторно-резултативни връзки, проявяващи се в динамиката на явленията.

Под ***индекси*** в широк статистически смисъл трябва да се разбират специфично конструирани измерители на посочените относителни различия и факторни ефекти.

Според това, дали измерваните относителни различия се отнасят за отделни единици или общо за цели съвкупности, индексите биват ***единични*** (индивидуални, прости) и ***множествени*** (сложни, съвкупностни). Единични са например индексите, измерващи относителното изменение за даден период на цените на отделни видове стоки, а множествени са тези, които измерват средното изменение на цените на съвкупностите от потребителски стоки и услуги.

Индексният метод изпълнява две функции - синтетична и аналитична. Съобразно с това и множествените индекси по своите функции биват два основни вида: ***синтетични*** и ***аналитични***.

Синтетичните индекси дават обобщена числова характеристика на относителни изменения или различия (на цени, продажби и др.) за съвкупности, вътре в които се проявяват единични различия. Тези единични различия се ***синтезират*** и изразяват в едно число. Съществен момент в конструкцията и вътрешното съдържание на синтетичните индекси е именно синтезът, т.е. обобщаването на единичните относителни различия, за да се измери ***средното относително различие***, характерно за цялата съвкупност.

Аналитичните индекси измерват влиянието на определени фактори върху изменението (различието) на някакво явление - резултат. Чрез тях се осъществява ***анализ***, като се разчленява изменението на явлението-резултат на отделни негови ***факторни компоненти***.

Различията във функциите, които изпълняват синтетичните и аналитичните индекси, обуславят и различия в тяхното конструиране и интерпретиране по същество.

В следващото изложение се разглеждат синтетичните индекси. Аналитичните индекси (индексният факторен анализ) не са включени в съдържанието на тази книга.[1]

Според характера на различията, които измерват, синтетичните индекси биват динамични (хронологични), териториални и планови (нормативни).

Динамичните индекси измерват относителните изменения на явленията във времето. Исторически това са първите индекси, намерили приложение в теорията и практиката.

Териториалните индекси се използуват при изучаване на относителните различия в териториален разрез, т.е. за измерване на относителните различия в интересуващите ни еднородни явления между отделни териториални единици.

Плановите (нормативните) индекси се използуват за измерване на планирани относителни изменения или за характеризиране на степента на изпълнение на планове. По своя строеж, начин на изчисляване и познавателна същност те не се различават от динамичните индекси. Основните методологически положения относно строежа и съдържанието на динамичните индекси важат напълно и за плановите индекси.

[1] Относно някои принципни постановки, формули и модели в областта на теорията на индексите и в частност на индексния факторен анализ вж. **Адриенко, В.**, Статистические индексы в экономических исследованиях, киев, 1985; **Аллен, Р.**, Экономические индексы, Москва, 1980; **Въжаров, Е.**, Опит за еднозначно решение на някои "индексни" икономически задачи, *Статистика*, 1984, кн. 1; **Гатев, К.**, Въведение в статистиката, И-во "ЛИА", София, 1995; **Гатев, К.**, Индексни модели за анализ на факторни ефекти, И-во "Стопанство", София, 1991; **Казинец, Л.**, Теория индексов, Москва, 1963; **Къналиев, Т.**, Относно построяването на индексни формули чрез последователно съблюдаване на съвкупностния подход, *Статистика*, 1978, кн.1; **Минасян, Г.**, Структури и ефекти, *Икономическа мисъл*, 1985, кн. 5; **Станев, С.**, Инфлационни процеси – същност и динамика, И-во на ВФСИ "Д. А. Ценов", Свищов, 1994; **Христов, Е.**, Оценяване на структурни и неструктурни ефекти в икономиката, *Икономическа мисъл*, 1987, кн. 8; **Шкодрев, Е.**, Логическата противоречивост на индексния метод и неговите познавателни възможности, *Икономика*, 1989, кн. 7; **Цонев, В.**, Традиционны и новый алгоритм построения индексных формул, Научни трудове на ВИИ "К. Маркс", Ф-т "Икономическа информация", т. 1, София, 1983 (с приложена библиография); **Цонев, В.**, Теорията на индексите и нейната статистическа алтернатива, *Статистика*, 1997, кн. 6 (с приложена библиография).

При индексния факторен анализ аналитичните индекси, които измерват изменението на явленията-резултат се наричат ***резултативни*** (общи, тотални) индекси, а индексите, измерващи факторните ефекти - ***факторни субиндекси***.

При изчисляване на всеки индекс независимо от това, дали той е синтетичен или аналитичен, дали е динамичен, териториален или планов, се съпоставят две величини: едната делимо, за която искаме да установим различието, а другата - делител, по отношение на която се измерва различието. Тези две величини се наричат ***индексни величини***. Отношението им се нарича ***индексно отношение*** или просто ***индекс***. Резултатите, които се получават от съпоставянето на индексните величини, се наричат ***индексни числа***. Те се изразяват или в коефициент и показват ***колко пъти*** едната величина е по-голяма или по-малка от другата, или в проценти, като се умножат на 100 (много рядко се изразяват в промили).

Периодът или териториалната единица, за които се установява различието, се нарича ***индексиран период***, респективно ***индексиран район*** (под район се разбира всякаква териториална единица - държава, окръг и др.). Периодът или териториалната единица, спрямо които се установява различието, се нарича ***базов период***, респективно ***базов район***.

Индексите могат да се отнасят за равнища, обеми и маси.[1] Затова те биват още: индекси на равнища, индекси на обеми и индекси на маси.

Индексите на равнища измерват изменението (или териториалните и др. различия) на значения или средни значения на признаци на единиците на някакви съвкупности или части от съвкупности. Такива са например индексите на цените на определени съвкупности от стоки, индексите на себестойността на съвкупност от произведени изделия и др.

Индексите на обеми измерват изменението (или териториалните и др. различия) на обемите на някакви съвкупности или части от съвкупности. Такива са например индексите на физическия обем на брутния вътрешен продукт, индексите на физическия обем на промишлената продукция и др.

[1] В литературата се употребяват и други термини, които са синоними на равнище, обем и маса, като например количествен, качествен и производен показател и др.

Индексите на маси измерват изменението (или териториалните и др. различия) едновременно на равнища и на обеми. Такъв е например индексът на оборота на фирмата. Оборотът е такава величина (маса), която може да се представи като произведение от количеството продадени стоки и техните цени. Затова индексът съдържа както изменението на количеството продадени стоки, така и на цените.

11.2. Динамични индекси на равнища

Исторически най-стари индекси на равнища са индексите на цените. Всички останали индекси от този вид по своя строеж малко или много са аналогични на тях. Поради това за изясняване на общите въпроси относно строежа и изчисляването на индексите на равнища ще използваме за пример индексите на цените. Преди това обаче, трябва да изтъкнем, че при изчисляването на динамичните индекси можем да съпоставяме както само две величини, отнасящи се за два периода, така и редица величини, образуващи цял динамичен ред. Във втория случай множествените (сложните) индекси, тъй както и единичните, могат да бъдат изчислени при постоянна база и при верижна база. Засега ще приемем, че имаме само два периода - индексиран (наричан още текущ или отчетен) и базов. Последователното индексиране при цял динамичен ред ще разгледаме отделно.

Цената на отделните стоки (отделните равнища) ще означаваме с p (от лат. *pretium* – цена), а количествата на отделните видове стоки (обемите) - с q (от лат. *quantitas* – количество). За да не обременяваме изложението по-нататък с нови символи, условно ще приемем, че p означава изобщо равнище, а q - обем, независимо за какви конкретни икономически или други показатели се отнасят. Индексираният период ще означаваме с малка единица, записана ниско в дясно към съответния символ (p_1, q_1), а базовия период - с малка нула, записана по същия начин (p_o, q_o). Като знак за множествен индекс се използува латинската буква I, а за единичен индекс - i. Към знака за индекс се прибавя съответният символ на индексираната величина, а до нея в скоби - знакът (символът) на величината, която е приета за тегло или съизмерител.

За ***пример*** нека приемем, че са дадени продадените количества и цените на една група стоки през 2005 и 2007 г. За опростяване на примера се вземат само три вида стоки. Въз основа на разполагаемите данни е съставена табл. 11.1.

Таблица 11.1

Количества и цени на продадените стоки в град "Н" през 2005 и 2007 година

Вид на стоките	Мярка	2005 г. (базов период)		2007 г. (индексиран период)		Единични индекси	
		коли-чества	цени (лв.)	коли-чества	цени	на коли-чествата	на цените
		q_0	p_0	q_1	p_1	$i_q = \frac{q_1}{q_0}$	$i_p = \frac{p_1}{p_0}$
а	бр.	2000	40	1400	48	0,70	1,20
б	кг.	8000	25	8000	35	1,00	1,40
в	бр.	10000	50	18000	45	1,80	0,90

При тези данни може да се постави задачата да се установи какво е ***средното относително изменение*** на цените общо на всички видове стоки от дадената стокова група през 2007 г. в сравнение с базовата 2005 г.

Ако проследим исторически решаването на подобни задачи, ще намерим различни решения.

Френският икономист ***Шарл Дюто*** например е изчислил (1738 г.) индекс на цените, като сумата на цените на стоките през индексирания период е разделил на сумата на цените на стоките, продадени през базовия период.

Италианецът ***Джан Карли*** е постъпил (1750 г.) при подобен случай по друг начин, изчислявайки индекса като непретеглена средна аритметична величина от предварително изчислени единични индекси на цените на отделните стоки.

Англичанинът ***Стенли Джевънс*** стига до извода (1863 г.), че индексът трябва да се изчисли като геометрична средна величина от единичните индекси.

Общата особеност на посочените и други подобни индекси се състои в това, че не се вземат под внимание количествата на продадените стоки. Цената на всяка стока участва един път в изчисленията, независимо от това колко единици (кг., бр. и др.) стоки са продадени по такава цена. Поради това тези индексни конструкции са история.

Цените се отнасят винаги за определени съвкупности от продадени стоки през базовия и през индексирания (отчетния) период. Затова индексът следва да отрази ***средното изменение*** на цените при дадено съотношение между продадените стоки, т.е. на дадена съвкупност, но измененията в обема на съвкупността не трябва да влияят върху индекса на цените. Това ще рече, че индексът на средното изменение на цените следва да се изчисли при еднаква съвкупност от стоки в числителя и знаменателя на индексното отношение.

Немският икономист ***Етиен Ласпер (1834 – 1913)*** е предложил (1864 г.) индексът на цените да се изчислява по формулата:

$$(11.1) \qquad I_{p(q_0)} = \frac{\sum p_1 q_0}{\sum p_0 q_0}.$$

Това е ***претеглен агрегатен индекс***, който има за числител условна величина - стойността на стоките, пресметната по цени на индексирания период, и количества от базовия период, а за знаменател - фактическата стойност на продадените стоки през базовия период.

В случая индексираната величина е цената, която за числителя и знаменателя е различна, а количествата, които са еднакви за числителя и знаменателя, придават на съответните цени тегло, съответстващо на дела на отделните стоки в тяхната обща съвкупност през базовия период. Затова количествата от базовия период изпълняват функцията на тегла.

Немският икономист ***Херман Пааше (1851 – 1925)*** е предложил (1874 г.) аналогична формула, но с тегла количествата стоки от индексирания период.

(11.2) $$I_{p(q_1)} = \frac{\sum p_1 q_1}{\sum p_0 q_1}.$$

Числителят е действителната продажна стойност на стоките през индексирания период, а знаменателят - условна величина, получена при цени от базовия период и количества от индексирания.

Тези две формули не са само история. Те в наше време се използват както при изчисляване на индексите на цените, така и при характеризиране на динамиката на равнища изобщо.

За да се изчислят двата индекса по данните от конкретния ***пример***, необходимо е от цените и количествата предварително да се изчисли продажната стойност на отделните стоки за базовия и за индексирания период - фактическа и условна (вж. табл. 11.2).

Таблица 11.2

Стойност на продадените стоки в град "Н" през 2005 и 2007 г.

Вид на стоките	Продажна стойност на стоките през 2005 г. (лв.)		Продажна стойност на стоките през 2007 г. (лв.)	
	по цени от 2005 г.	по цени от 2007 г.	по цени от 2005 г.	по цени от 2007 г.
	$q_0 p_0$	$q_0 p_1$	$q_1 p_0$	$q_1 p_1$
а	80000	96000	56000	67200
б	200000	280000	200000	280000
в	500000	450000	900000	810000
	780000	826000	1156000	1157200

$$I_{p(q_0)} = \frac{\sum p_1 q_0}{\sum p_0 q_0} = \frac{826000}{780000} = 1{,}059 \text{ или } 105{,}9\ \%\ ;$$

$$I_{p(q_1)} = \frac{\sum p_1 q_1}{\sum p_0 q_1} = \frac{1157200}{1156000} = 1{,}001 \text{ или } 100{,}1\ \%\ .$$

Както се вижда, индексът, изчислен по формулата на Ласпер, показва, че цените са се увеличили средно с 5,9 на сто, а по формулата на Пааше - само с 0,1 на сто.

Разликата между двата индекса по същество се състои в това, че първият (на Ласпер) измерва средното относително изменение на цените (общо казано - на равнищата) при структура на стоките от базовия период, докато вторият индекс (на Пааше) измерва средното изменение на цените (на равнищата) при структура от индексирания период. Иначе казано, двата индекса измерват средното относително изменение на цените на две различни съвкупности, чиято вътрешна структура (съотношение между количествата) е различна.

Агрегатната форма е основна форма на множествените индекси. Преди да се извърши индексирането, частите на цялата съвкупност се обединяват (агрегират). Получените две суми от произведенията на цените (равнищата) и количествата (обемите) са абсолютни величини (агрегати), които се съпоставят като индексно отношение. Този метод за изчисляване на множествени индекси се нарича ***агрегатен метод***, а индексите - ***агрегатни индекси***.

В редица случаи практически не е удобно индексите да се изчислят по агрегатния метод. По тази причина често се прилага ***осреднителният метод***. Той се състои в това, че предварително се изчисляват единични индекси, които се осредняват и се получава претеглена средна като формула, по която се определя и начинът на осредняването - дали ще бъде аритметично или хармонично и какви да бъдат теглата. Следователно средноаритметичният и среднохармоничният индекс трябва да бъдат тъждествени на агрегатния индекс.

Да вземем агрегатния индекс на цените на Ласпер $I_{p(q_0)} = \frac{\sum p_1 q_0}{\sum p_0 q_0}$ и да изведем тъждествените му ***средноаритметичен*** и ***среднохармоничен*** индекси.

Преди всичко трябва да се намерят единичните индекси. При приетите символи те ще бъдат

$$\frac{p_1'}{p_0'}, \frac{p_1''}{p_0''}, \frac{p_1'''}{p_0'''}, \ldots, \frac{p_1^{(n)}}{p_0^{(n)}}.$$

Известно е, че всяка средна аритметична претеглена има за знаменател сумата от теглата, с които са умножени осредняваните

величини в числителя. При това правило и при условие, че ***средноаритметичният индекс*** трябва да бъде тъждествен на агрегатния, лесно е да се установи, че теглата при средноаритметичния индекс ще бъдат онези величини, чиято сума образува знаменателя на агрегатната формула. При агрегатната формула на Ласпер това са: $p_0'q_0', p_0''q_0'', ..., p_0^{(n)}q_0^{(n)}$. Или аритметичното осредняване на единичните индекси ще се извърши така:

(11.3)

$$I_{p(q_0)} = \frac{\frac{p_1'}{p_0'}p_0'q_0' + \frac{p_1''}{p_0''}p_0''q_0'' + ... + \frac{p_1^{(n)}}{p_0^{(n)}}p_0^{(n)}q_0^{(n)}}{p_0'q_0' + p_0''q'' + ... + p_0^{(n)}q_0^{(n)}} = \frac{\sum \frac{p_1}{p_0}p_0q_0}{\sum p_0q_0}.$$

Не е трудно да се установи при съответни съкращения в числителя, че получената средноаритметична формула се трансформира в агрегатна. Изпълнено е следователно условието

$$\frac{\sum p_1q_0}{\sum p_0q_0} = \frac{\sum \frac{p_1}{p_0}p_0q_0}{\sum p_0q_0}.$$

Известно е, че между средната аритметична и средната хармонична в статистиката съществува определена връзка. Ако се опрем на тази връзка, можем от средноаритметичния индекс да изведем среднохармоничния.

Осреднявани величини при средноаритметичния индекс са единичните индекси $\frac{p_1'}{p_0'}, \frac{p_1''}{p_0''}, \frac{p_1'''}{p_0'''}, ..., \frac{p_1^{(n)}}{p_0^{(n)}}$, а теглата - $p_0'q_0', p_0''q_0'', ..., p_0^{(n)}q_0^{(n)}$. Очевидно е, че произведенията на единичните индекси и посочените тегла дават условните величини

$$\frac{p_1'}{p_0'}p_0'q_0' = p_1'q_0'; \frac{p_1''}{p_0''}p_0''q_0'' = p_1''q_0''; ...; \frac{p_1^{(n)}}{p_0^{(n)}}p_0^{(n)}q_0^{(n)} = p_1^{(n)}q_0^{(n)}$$ и т.н.,

чиято сума образува числителя на агрегатния индекс. Следователно те са необходимите тегла за изчисляване на ***среднохармоничния индекс***.

(11.4)

$$I_{p(q_0)} = \frac{p_1'q_0' + p_1''q_0'' + ... + p_1^{(n)}q_o^{(n)}}{\frac{1}{\frac{p_1'}{p_0'}} p_1'q_0' + \frac{1}{\frac{p_1''}{p_0''}} p_1''q_0'' + ... + \frac{1}{\frac{p_1^{(n)}}{p_0^{(n)}}} p_1^{(n)}q_0^{(n)}} = \frac{\sum p_1q_0}{\sum \frac{1}{\frac{p_1}{p_0}} p_1q_0}.$$

От практическа гледна точка среднохармоничната формула в дадения случай няма особено практическо значение, тъй като тя изисква наличие на условните величини $p_1'q_0', p_1''q_0''$ и т.н. Ако разполагаме с тях, възможно е да се изчисли агрегатният индекс. Необходима и удобна форма тук е средноаритметичната форма на индекса, тъй като при нея теглата са фактически величини от базовия период.

Като се имат предвид установените вече връзки между агрегатната, средноаритметичната и среднохармоничната формула, когато имаме за основа агрегатния индекс на Ласпер, по аналогия можем да установим средноаритметичната и среднохармоничната формула, съответстващи на агрегатната формула на Пааше.

Агрегатната формула е

$$I_{p(q_1)} = \frac{\sum p_1q_1}{\sum p_0q_1}.$$

Средноаритметичната ще има за тегла условните величини, чиято сума образува знаменателя на агрегатната, или

(11.5) $$I_{p(q_1)} = \frac{\sum \frac{p_1}{p_0} p_0q_1}{\sum p_0q_1}.$$

Среднохармоничната, обратно, ще има за тегла онези величини, чиято сума образува числителя на агрегатната, или

(11.6) $$I_{p(q_1)} = \frac{\sum p_1q_1}{\sum \frac{1}{\frac{p_1}{p_0}} p_1q_1}.$$

Следователно

$$\frac{\sum p_1 q_1}{\sum p_0 q_1} = \frac{\sum \frac{p_1}{p_0} p_0 q_1}{\sum p_0 q_1} = \frac{\sum p_1 q_1}{\sum \frac{1}{\frac{p_1}{p_0}} p_1 q_1}.$$

Вижда се, че в този случай, когато агрегатният индекс е с тегла количествата от индексирания период (индекс на Пааше), практически удобна е среднохармоничната формула, при която теглата са известни реални величини от индексирания период, докато средноаритметичната изисква наличие на условни величини.

Средноаритметичният индекс, тъждествен на агрегатния индекс на Ласпер, и среднохармоничният индекс, тъждествен на агрегатния индекс на Пааше, ще се изчислят по данните от примера по следния начин:

$$I_{p(q_0)} = \frac{\sum \frac{p_1}{p_0} p_0 q_0}{\sum p_0 q_0} = \frac{1{,}20 \cdot 80000 + 1{,}40 \cdot 200000 + 0{,}90 \cdot 50000}{80000 + 200000 + 50000} = \frac{826000}{780000} = 1{,}059$$

или 105,9 % ;

$$I_{p(q_1)} = \frac{\sum p_1 q_1}{\sum \frac{1}{\frac{p_1}{p_0}} p_1 q_1} = \frac{67200 + 280000 + 810000}{\frac{1}{1{,}20} 67200 + \frac{1}{1{,}40} 280000 + \frac{1}{0{,}90} 810000} = \frac{1157210}{1156000} = 1{,}001$$

или 100,1 % .

Вижда се, че получените резултати по осреднителния метод не се различават от тези, получени по агрегатния. Не може и да бъде иначе, тъй като разликата между агрегатните, средноаритметичните и средно-хармоничните индекси е само във формата, обуславяща различна техника на изчисляване.

Нека се върнем по същество към индексите на Ласпер и на Пааше. Както се вижда от конструкцията на техните формули, изборът се свежда до избор на системата на претеглянето, т.е. дали индексът да се изчисли

при структура от базовия или при структура от индексирания период. Този избор зависи от поставената конкретна задача.

Не може следователно да се каже, че едната от двете формули дава верен, а другата - неверен резултат. В една или друга степен те се използват в световната статистическа практика и при икономическите анализи. При конкретните икономически индекси, конкретно се обосновава индексната формула и в теорията на икономическата статистика тези индекси са предмет на обстойно разглеждане. Необходимо е обаче да се знаят общите условия, от които зависи количествената разлика между двата индекса.

Владислав Борткиевич (1868 – 1931) е доказал (1923 г.), че като отношение между двата индекса - на Пааше и Ласпер, може да се получи математически израз, въз основа на който могат да се обяснят условията, при които те са еднакви или различни. Формулата на това отношение всъщност се основава на известната вече формула за отношението на две средни аритметични с различни тегла.

$$\frac{\sum p_1 q_1}{\sum p_o q_1} : \frac{\sum p_1 q_o}{\sum p_o q_o} = \frac{\sum \frac{p_1}{p_o} p_o q_1}{\sum p_o q_1} : \frac{\sum \frac{p_1}{p_o} p_o q_o}{\sum p_o q_o} = 1 + V_{\frac{p_1}{p_o}} \cdot V_{\frac{q_1}{q_o}} \cdot r_{\frac{p_1}{p_o}, \frac{q_1}{q_o}}, \quad (11.7)$$

където:

$V_{\frac{p_1}{p_o}}$ е коефициент на вариацията на единичните индекси на цените (равнищата)[1];

[1] $V_{\frac{p_1}{p_o}} = \sqrt{\frac{\sum \left(\frac{p_1}{p_o} - \frac{\Sigma\, p_1 q_o}{\Sigma\, p_o q_o} \right)^2 p_o q_o}{\sum p_o q_o}} : \frac{\sum p_1 q_o}{\sum p_o q_o}$.

$V_{\frac{q_1}{q_o}}$ - коефициент на вариацията на единичните индекси на количествата (обемите)[1];

$r_{\frac{p_1}{p_o},\frac{q_1}{q_o}}$ - коефициент на линейната корелация между двете групи единични индекси[2].

От тази формула се вижда, че двата индекса са еднакви само при наличието поне на едно от следните три условия:

1. Ако единичните индекси на цените (или общо на равнищата) са еднакви, т.е. $\frac{p_1'}{p_0'} = \frac{p_1''}{p_0''} = ... = \frac{p_1^{(n)}}{p_o^{(n)}} = const;$ тогава $V_{\frac{p_1}{p_o}} = 0;$
2. Ако единичните индекси на количествата (или общо на обемите) са еднакви, т.е. $\frac{q_1'}{q_0'} = \frac{q_1''}{q_0''} = ... = \frac{q_1^{(n)}}{q_o^{(n)}} = const;$ тогава $V_{\frac{q_1}{q_o}} = 0;$
3. Ако няма линейна зависимост между двете групи единични индекси, т.е. ако коефициентът на линейната корелация $r_{\frac{p_1}{p_o},\frac{q_1}{q_o}} = 0.$

Ако не е налице нито едно от условията, двата индекса ще се различават по числовата си стойност. Индексът на равнище, изчислен по формулата на Пааше, ще бъде по-голям от индекса, изчислен по формулата на Ласпер, когато коефициентът на линейната корелация е положителен (тъй като коефициентите на вариацията са винаги

---

[1] $V_{\frac{q_1}{q_o}} = \sqrt{\frac{\sum\left(\frac{q_1}{q_o} - \frac{\Sigma q_1 p_o}{\Sigma q_o p_o}\right)^2 p_o q_o}{\sum p_o q_o}} : \frac{\sum q_1 p_o}{\sum q_o p_o}.$

[2] $r_{\frac{p_1}{p_o},\frac{q_1}{q_o}} = \frac{\sum\left[\left(\frac{p_1}{p_o} - \frac{\sum p_1 q_o}{\sum p_o q_o}\right)\left(\frac{q_1}{q_o} - \frac{\sum q_1 p_o}{\sum q_o p_o}\right)\right] p_o q_o}{\sqrt{\sum\left(\frac{p_1}{p_o} - \frac{\sum p_1 q_o}{\sum p_o q_o}\right)^2 p_o q_o \sum\left(\frac{q_1}{q_o} - \frac{\sum q_1 p_o}{\sum q_o p_o}\right)^2 p_o q_o}}$

положителни величини). Това ще се получи, когато на по-големите единични индекси на цените съответстват общо взето по-големи индекси на количествата, а на по-малките индекси на цените съответстват общо взето по-малки индекси на количествата. Обратно, ако съчетанието между единичните индекси на цените и количествата е такова, че срещу по-големите индекси на цените стоят по-малки индекси на количествата, а срещу по-малките индекси на цените стоят по-големи индекси на количествата, коефициентът на линейната корелация е отрицателен, а следователно и индексът, изчислен по формулата на Пааше, ще бъде по-малък от индекса, изчислен по формулата на Ласпер.

Тази зависимост на относителната разлика между двата индекса от вариацията на единичните индекси и линейната корелация между тях има съществено значение, тъй като дава възможност да се обясни и характерът на настъпващите изменения в структурата на съвкупността от базовия до индексирания период.

Очевидно е от конкретните изчисления въз основа на данните от примера, че индексът с тегла от индексирания период е по-малък от индекса с тегла от базовия период. Това е указание, че през индексирания период се е увеличил относителният дял на онези стоки, при които има най-малко увеличение или има намаление на цените. Наистина стока "в" има намаление на цената и паралелно с това най-голямо увеличение на количествата (с 80 на сто). Стока "б" има най-голямо увеличение на цената, а количествата са останали непроменени. Стока "а" има значително увеличение на цената, а количествата са намалени.

Както ще стане известно по-нататък, от съпоставянето на двата индекса (на Пааше и на Ласпер) се получава допълнителна информация за силата и посоката на влияние на структурните изменения.

Разгледаните два индекса на равнища - на Ласпер и на Пааше, не са единствените в литературата и практиката, които съдържат съотношението в количествата (обемите). В стремежа да се преодолее двузначното решение (с тегла от базовия или от индексирания период), са конструирани "компромисни" формули, които съдържат някаква средна структура. Такива са например формулите на ***Френсис Еджуърт (1845-1926)*** и на ***Ирвинг Фишер (1867-1947)***.

Формулата на ***Еджуърт*** (1896 г.) има следния вид:

$$(11.8) \qquad I_{p(q_0+q_1)} = \frac{\sum p_1\left(\frac{q_0+q_1}{2}\right)}{\sum p_0\left(\frac{q_0+q_1}{2}\right)} = \frac{\sum p_1(q_0+q_1)}{\sum p_0(q_0+q_1)}.$$

Както се вижда, за тегла се използват средните количества, установени като полусбор от количествата през базовия и през индексирания период. Индексът на цените, изчислен по тази формула, би показал средното изменение на цените на условна, осреднена от базовия и индексирания период съвкупност от стоки. Главно поради това, че не може да се адресира към реална съвкупност от определен период, този индекс почти не се използва в световната практика.

Формулата на ***Фишер*** (1922 г.), наречена от него "идеална", представлява геометрична средна от индексите на Ласпер и Пааше.

$$(11.9) \qquad I_p = \sqrt{\frac{\sum p_1 q_1}{\sum p_0 q_1} \cdot \frac{\sum p_1 q_0}{\sum p_0 q_0}}.$$

Както се вижда, индексът на Фишер, макар и под друга форма, е адресиран към осреднена съвкупност, а не към фактическата съвкупност от базовия или индексирания период. Това трябва да се има предвид при неговото съдържателно тълкуване. Поради тази особеност той се използва сравнително рядко и то главно при международни сравнения.

Има и други опити за конструиране на формули, с които да се преодолее алтернативата по отношение на избора на теглата, но не са общоприети от теорията и практиката.

По-горе беше коментиран по принцип въпросът за теглата при конструирането на индексите на равнища. В статистическата практика, особено при изчисляването на индекса на потребителските цени и характеризиране на инфлационните процеси, възникват редица други въпроси относно системата на теглата, с оглед да се отрази структурата на потреблението на домакинствата, да се създаде възможност за изчисляване на индексите за кратки периоди (тримесечия, месеци и др.) и т.н. Тези въпроси са предмет на разглеждане в икономическата статистика.

11.3. Динамични индекси на обеми

Докато разгледаните индекси дават синтетичен израз на средното относително изменение на равнища, т.е. на определени признаци на единиците на определена по състав (или съответно осреднена) съвкупност, динамичните индекси на обеми характеризират относителните изменения (различия) в обема на сравняваните съвкупности.

Конструирането на тези индекси и начинът на изчисляване зависят в голяма степен от съизмеримостта на единиците на съвкупностите. Ако единиците на съпоставяните съвкупности са съизмерими в тяхното натурално изражение, индексът може да има формата на агрегатен индекс без съизмерители. Ако съпоставим например общата сума на количествата продадени ябълки на всички градски пазари в страната през 2007 г. ($\sum q_1$) с продадените ябълки на същите пазари през 2006 г. ($\sum q_0$), бихме получили следният агрегатен индекс за натурален обем:

$$I_q = \frac{\sum q_1}{\sum q_0}. \qquad (11.10)$$

По същия начин може да се изчисли индекс на броя на заетите лица във всички отрасли на промишлеността, индекс на реколтираните площи от дадена култура в страната и т.н. Както се вижда, в такива случаи не възникват трудности.

Когато обаче единиците на съвкупността са несъизмерими непосредствено в натурално изражение, изчисляването на агрегатен индекс по посочената формула е невъзможно, тъй като е невъзможно сумирането на всички единици. Съвършено очевидно е например, че е невъзможно да се сумира в натура продукцията на цялата промишленост, изразяваща се в разнообразни по потребителна стойност, по натурално-веществена форма и по мярка изделия. В такъв случай е необходимо преди да се съпоставят двете съвкупности, да се приведат в съизмерим вид чрез някакъв общ съизмерител. Това може да бъде например някакъв коефициент, чрез който единиците на съвкупността се привеждат в условни единици. Най-често обаче при икономическите индекси за общ съизмерител служи цената. Във всички случаи съизмерителите трябва да бъдат конкретно обосновани. Едно задължително общо изискване е

съизмерителят да бъде единен за двата периода не само по същество, но и по отношение на времето и пространството, т.е. да се отнася за един и същ период както в числителя, така и в знаменателя на индексното отношение и да е еднакъв за всички обхванати териториални и други единици.

Ако си послужим с приетите вече символи, множественият ***агрегатен индекс на обем*** ще има следната обща формула:

$$(11.11) \qquad I_{q(p_c)} = \frac{\sum q_1 p_c}{\sum q_0 p_c}.$$

Тук p_c означава съизмерител (в случая цена), който е постоянен (бележи се със c - от лат. *constantis* - постоянен, неизменен). Съизмерителят може да не бъде нито от базовия, нито от индексирания период, а от друг някакъв период. Такъв съизмерител могат да бъдат например т.нар. съпоставими цени, отнасящи се за определена година. В редица случаи обаче се налага да се изчислят индекси на обеми, при които съизмерителят се отнася за базовия или за индексирания период. Агрегатните формули в тези два случая ще имат следния вид:

$$(11.12) \qquad I_{q(p_0)} = \frac{\sum q_1 p_0}{\sum q_0 p_0};$$

$$(11.13) \qquad I_{q(p_1)} = \frac{\sum q_1 p_1}{\sum q_0 p_1}.$$

Във всеки от посочените три индекса се сравняват не натурални обеми, а агрегирани величини, получени чрез някакви съизмерители. Следователно те не измерват пряко изменението на обемите, а изменението на маси, получени при еднакви съизмерители за двата периода. По такъв начин се характеризира косвено изменението на физическия обем. При това обаче има значение за кой период се отнася съизмерителят.

Индексите на обеми могат да се изчислят и по осреднителния метод, като се спазват същите изисквания, които бяха изложени по отношение на индексите на равнища.

На агрегатния индекс $I_{q(p_c)} = \frac{\sum q_1 p_c}{\sum q_0 p_c}$ съответствуват:

средноаритметичен индекс

$$I_{q(p_c)} = \frac{\sum \frac{q_1}{q_0} q_0 p_c}{\sum q_0 p_c}; \qquad (11.14)$$

среднохармоничен индекс

$$I_{q(p_c)} = \frac{\sum q_1 p_c}{\sum \frac{1}{\frac{q_1}{q_0}} q_1 p_c}. \qquad (11.15)$$

По аналогичен начин се конструират средноаритметичните и среднохармоничните индекси, тъждествени на агрегатните индекси на обеми със съизмерители от базовия и от индексирания период:

$$I_{q(p_0)} = \frac{\sum q_1 p_0}{\sum q_0 p_0} = \frac{\sum \frac{q_1}{q_0} q_0 p_0}{\sum q_0 p_0} = \frac{\sum q_1 p_0}{\sum \frac{1}{\frac{q_1}{q_0}} q_1 p_0}; \qquad (11.16)$$

$$I_{q(p_1)} = \frac{\sum q_1 p_1}{\sum q_0 p_1} = \frac{\sum \frac{q_1}{q_0} q_0 p_1}{\sum q_0 p_1} = \frac{\sum q_1 p_1}{\sum \frac{1}{\frac{q_1}{q_0}} q_1 p_1}. \qquad (11.17)$$

Ще илюстрираме изчисляването на агрегатния и на тъждествения му средноаритметичен и среднохармоничен индекс при съизмерители от базовия период по данните от ***примера***, съдържащ се в табл. 11.1 и табл. 11.2.

$$I_{q(p_0)} = \frac{\sum q_1 p_0}{\sum q_0 p_0} = \frac{1156000}{780000} = 1{,}482 \text{, или } 148{,}2\ \%;$$

$$I_{q(p_0)} = \frac{\sum \frac{q_1}{q_0} q_0 p_0}{\sum q_0 p_0} = \frac{0{,}70 \cdot 80000 + 1{,}00 \cdot 200000 + 1{,}8 \cdot 500000}{80000 + 200000 + 500000} = \frac{1156000}{780000} = 1{,}482,$$

или 148,2 % ;

$$I_{q(p_0)} = \frac{\sum q_1 p_0}{\sum \frac{1}{\frac{q_1}{q_0}} q_1 p_0} = \frac{56000 + 200000 + 900000}{\frac{1}{0{,}70} 56000 + \frac{1}{1{,}00} 200000 + \frac{1}{1{,}8} 900000} = \frac{1156000}{780000} = 1{,}482,$$

или 148,2 %.

Получените индекси показват, че физическият обем на продажбите (физическият обем на стокооборота) се е увеличил през 2007 г. спрямо 2005 г. с 48,2 на сто по цени от 2005 г.

Изборът между индексите на обеми с различни съизмерители (от базовия или индексирания период) трябва да бъде направен съобразно с характера на икономическия показател и конкретната цел, за която се изчислява индексът. Когато изборът трябва да се направи между последните две формули, предпочитание има тази със съизмерител от базовия период, тъй като обикновено си поставяме задача да измерим именението на физическия обем от базовия до индексирания период при условие изменението на съизмерителя да не влияе върху величината на индекса.

Относителната разлика между двата индекса (със съизмерител от индексирания и със съизмерител от базовия период) се определя от онези условия, които важат при индексите на равнища:

$$(11.18) \qquad \frac{\sum q_1 p_1}{\sum q_0 p_1} : \frac{\sum q_1 p_0}{\sum q_0 p_0} = \frac{\sum \frac{q_1}{q_0} q_0 p_1}{\sum q_0 p_1} : \frac{\sum \frac{q_1}{q_0} q_0 p_0}{\sum q_0 p_0} = 1 + V_{\frac{p_1}{p_0}} . V_{\frac{q_1}{q_0}} . r_{\frac{p_1}{p_0}, \frac{q_1}{q_0}}$$

Както се вижда, относителната разлика между двата агрегатни индекса на обеми се определя от онези коефициенти, от които се определя и относителната разлика между двата агрегатни индекса на

равнища (на Пааше и на Ласпер). Това се обяснява с ***инвариантността на индексното отношение*** при промяна на местата на q и p:

$$\left(\frac{\sum q_1 p_1}{\sum q_0 p_1} : \frac{\sum q_1 p_0}{\sum q_0 p_0}\right) = \left(\frac{\sum p_1 q_1}{\sum p_o q_1} : \frac{\sum p_1 q_o}{\sum p_o q_o}\right). \quad (11.19)$$

Тази инвариантност позволява при необходимост да се използва едното или другото отношение с увереност, че ще се получават еднакви резултати (например при индексния факторен анализ).

11.4. Динамични индекси на маси

Динамичните индекси на маси характеризират относителните изменения на стойностни и други маси, които са сума от произведения на обеми и равнища и чието изменение следователно е резултат на едновременното изменение на обемите и на равнищата. Изменението например на стокооборота в дадения пример в табл. 11.1 е резултат от изменението на количествата на продадените стоки и на изменението на цените.

Индексът на маса има вида

$$I_{qp} = \frac{\sum q_1 p_1}{\sum q_0 p_0}. \quad (11.20)$$

По данните от примера индексът на продажната стойност (стокооборота) на стоките е

$$I_{qp} = \frac{1157200}{780000} = 1{,}483, \text{ или } 148{,}3\ \%.$$

Посочената агрегатна формула на индекса на маса може да се транформира в ***средноаритметична*** и в ***среднохармонична***.

$$I_{qp} = \frac{\sum\left(\frac{q_1}{q_0} \cdot \frac{p_1}{p_0} q_0 p_0\right)}{\sum q_0 p_0}; \quad (11.21)$$

$$(11.22) \qquad I_{qp} = \frac{\sum q_1 p_1}{\sum \left(\frac{1}{\frac{q_1}{q_0} \cdot \frac{p_1}{p_0}} q_1 p_1 \right)}.$$

Както се вижда, изчисляването на индекса на маса по тези две формули изисква да са изчислени предварително единичните (или груповите) индекси на обемите и равнищата. Формулите имат практически смисъл в случаите, когато от предварително изчислени индекси за определени подсъвкупности трябва да се изчисли общ индекс за цялата съвкупност.

11.5. Връзка между индекси на равнище, обем и маса

Между индексите на равнище, обем и маса, изчислени от едни и същи данни, съществува определена връзка. Индексът на маса е равен на произведението на индекса на равнище и индекса на обем при условие, че единият от тях е изчислен с тегла (респ. съизмерители) от базовия период, а другия - с тегла (респ. съизмерители) от индексирания период:

$$(11.23) \qquad \frac{\sum q_1 p_1}{\sum q_0 p_0} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \cdot \frac{\sum q_1 p_0}{\sum q_0 p_0};$$

$$(11.24) \qquad \frac{\sum q_1 p_1}{\sum q_0 p_0} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \cdot \frac{\sum q_1 p_1}{\sum q_0 p_1}.$$

По данните от примера 1,483 = 1,001 **.** 1,482.

От това следва, че може да се намери всеки един от трите индекса при дадени другите два. Например:

$$(11.25) \qquad \frac{\sum p_1 q_1}{\sum p_0 q_1} = \frac{\sum q_1 p_1}{\sum q_0 p_0} : \frac{\sum q_1 p_0}{\sum q_0 p_0};$$

$$(11.26) \qquad \frac{\sum q_1 p_0}{\sum q_0 p_0} = \frac{\sum q_1 p_1}{\sum q_0 p_0} : \frac{\sum p_1 q_1}{\sum p_0 q_1}.$$

В случай, че индексът на равнище и индексът на обем са с тегла, респ. съизмерители, от единия период (базов или индексиран), посочената връзка не се осъществява. Например:

$$(11.27) \qquad \frac{\sum q_1 p_1}{\sum q_0 p_0} \neq \frac{\sum p_1 q_0}{\sum p_0 q_0} . \frac{\sum q_1 p_0}{\sum q_0 p_0} .$$

Равенство ще се получи, ако се прибави още един множител, изразяващ съотношението между съответните индекси с тегла (съизмерители) от индексирания и базовия период. Например:

$$(11.28) \qquad \frac{\sum q_1 p_1}{\sum q_0 p_0} = \frac{\sum p_1 q_0}{\sum p_0 q_0} . \frac{\sum q_1 p_0}{\sum q_0 p_0} \left(\frac{\sum p_1 q_1}{\sum p_0 q_1} : \frac{\sum p_1 q_0}{\sum p_0 q_0} \right).$$

Посочените връзки имат съществено значение при индексния факторен анализ.

11.6. Динамични индекси на средни равнища

Динамичните индекси на равнища, които бяха разгледани, измерват ***средното относително изменение*** на две или повече единични равнища, всяко от които се изменя в различна степен. Те са конструирани при ***постоянен състав*** (структура) на съвкупността посредством теглата от базовия или от индексирания период. Затова се наричат още ***индекси при постоянен състав***. Често обаче се налага да се измери ***изменението на средното равнище*** за съвкупността от еднородни единици, които могат да се сумират в натурално изражение. Необходим е например индекс на средната производителност на труда на работниците от едно предприятие, индекс на средната работна заплата, на средната цена на една стока, продавана в различни количества на различни пазари (в различни фирми и др.) и т.н.

Цената на отделен продукт, например продаван на няколко пазара, е различна не само за отделните периоди (базов и индексиран), но и за различните пазари за даден период. Общо за всички пазари средната цена през базовия период ($\overline{p}_0$) и през индексирания период ($\overline{p}_1$), ще бъде средна аритметична, претеглена със съответните количества:

$$\overline{p}_0 = \frac{\sum p_0 q_0}{\sum q_0}; \qquad \overline{p}_1 = \frac{\sum p_1 q_1}{\sum q_1}.$$

Изменението на средната цена (на средното равнище) ще се намери чрез индекс, който се получава като отношение на двете средни:

$$(11.29) \qquad I_{\overline{p}} = \frac{\overline{p}_1}{\overline{p}_0} = \frac{\sum p_1 q_1}{\sum q_1} : \frac{\sum p_0 q_0}{\sum q_0}.$$

Числовата стойност на претеглената средна аритметична, както е известно, се определя не само от осредняваните величини, но и от съотношението на теглата. Следователно отношението на двете средни претеглени изразява не ***средното изменение*** на отделните равнища, а ***изменението на средното равнище***. Това са две различни, макар и свързани постановки.

Изменението на средното равнище съдържа освен средното изменение на единичните равнища (в примера - цените на отделните пазари), още и изменението в ***състава*** (в структурата) на съвкупността (в примера - съотношението между продадените количества на отделните пазари). Затова индексът на средното равнище се нарича още ***индекс при променлив състав***. Тогава, когато измененията в структурата са значителни, възможно е индексът на средното равнище да бъде по-голям от най-големия единичен индекс или по-малък от най-малкия единичен индекс на отделните равнища. Това е известният ***статистически парадокс*** при индексите с променлив състав (на средни равнища). Този парадокс, разбира се, е привиден, тъй като индексът на средното равнище по силата на своята природа го допуска като нещо напълно нормално и обяснимо.

При изчисляването на индекси на средни равнища (индекси при променлив състав) се поставя познавателна задача, различна от тази, на която отговорят индексите на средното изменение на равнищата (индексите при постоянен състав). Те синтезират в едно число по-широк кръг фактори - както тези, които определят средното изменение на равнищата на определена по състав съвкупност, така и изменението на структурата на съвкупността от базовия до индексирания период.

Ако двете средни равнища - за базовия ($\bar{p}_0$) и за индексирания период ($\bar{p}_1$), се изчислят с еднакви за двата периода тегла (например от базовия период), тогава тяхното отношение престава да бъде индекс на средно равнище и става индекс на средно изменение на равнищата (индекс при постоянен състав). Това се вижда от елементарната преработка:

$$(11.30) \qquad \frac{\sum p_1 q_0}{\sum q_0} : \frac{\sum p_0 q_0}{\sum q_0} = \frac{\sum p_1 q_0}{\sum p_0 q_0}.$$

Това показва, че разликата между индекса на средното равнище ($I_{\bar{p}}$) и индекса на средното изменение на равнищата ($I_{p(q_0)}$) ще се дължи на изменението на съотношението между теглата (структурата на съвкупността) от базовия до индексирания период. Следователно като отношение между двата индекса ще се получи индекс, измерващ влиянието на структурните изменения (I_{str}). Той по-конкретно показва в какво направление и с каква сила влияят измененията в структурата върху изменението на средното равнище във връзка с различията в единичните равнища през индексирания период.

$$(11.31) \qquad I_{str} = \left(\frac{\sum p_1 q_1}{\sum q_1} : \frac{\sum p_0 q_0}{\sum q_0} \right) : \frac{\sum p_1 q_0}{\sum p_0 q_0} = \frac{\sum q_1 p_1}{\sum q_0 p_1} : \frac{\sum q_1}{\sum q_0}.$$

Този структурен индекс има съществено значение при индексния факторен анализ, при който трябва да се измерва влиянието и на структурните изменения като фактор.

11.7. Индекси с постоянна и с верижна база

При статистическия анализ често се налага изчисляването на динамични индекси не само за отделно взет период, а за поредица от подпериоди, обхванати от цял динамичен ред. Това е така нареченото ***последователно индексиране***, при което се получават редове от индексни числа. При него се поставя въпросът за базата, спрямо която ще се характеризират измененията, както и за системата на теглата (респ. съизмерителите) при съответно избрана база.

Според познавателния интерес за база на всички индекси може да бъде приет един от подпериодите или всеки предходен подпериод. В първия случай се получават ***индекси с постоянна база***, а във втория - ***верижни индекси***. Теглата (респ. съизмерителите) могат също да бъдат постоянни или променливи. От съчетанието на базата и на вида на теглата се получават четири системи от индекси: 1) индекси с постоянна база с постоянни тегла; 2) индекси с постоянна база с променливи тегла; 3) верижни индекси с постоянни тегла; 4) верижни индекси с променливи тегла. В последния случай теглата могат да бъдат от базовия или от индексирания период.

Тези комбинации са представени в табл. 11.3. (с $t = 1, 2, 3, ..., n$ са означени подпериодите).

Таблица 11.3

Системи от индекси

Тегла / База	Постоянни	Променливи
Постоянна	$I_{p_{t/0}(q_c)} = \frac{\sum p_t q_c}{\sum p_0 q_c}$	$I_{p_{t/0}(q_t)} = \frac{\sum p_t q_t}{\sum p_0 q_t}$
Верижна	$I_{p_{t/t-1}(q_c)} = \frac{\sum p_t q_c}{\sum p_{t-1} q_c}$	$I_{p_{t/t-1}(q_t)} = \frac{\sum p_t q_t}{\sum p_{t-1} q_t}$ $I_{p_{t/t-1}(q_{t-1})} = \frac{\sum p_t q_{t-1}}{\sum p_{t-1} q_{t-1}}$

Всяка от изложените системи индекси има определен смисъл и намира приложение в практиката. Съществуват, разбира се, и ограничителни условия при тяхното използване, отнасящи се до сравнимостта, тъй като с течение на времето се изменя и кръгът на единиците, включвани в съпоставяните съвкупности.

Очевидно е, че при индексите с постоянни тегла не се обхващат измененията, настъпващи в състава, тъй като измерваните различия се

отнасят за една и съща съвкупност от единици (имат се предвид индексите на равнища). Верижните индекси с постоянни тегла дават възможност да се включи за всеки подпериод по-широк кръг единици, за които съществува сравнимост с предходния период. При тях обаче се нарушава сравнимостта на самите индексни числа, когато трябва да се разглеждат като цял динамичен ред. Ето защо изборът на едно или друго от възможните решения по отношение на теглата може да бъде обосновано само с конкретни за дадена изследователска задача съображения.

При индексите с постоянна и верижна база и с различни тегла възниква още един съществен въпрос: ***възможно ли е преминаване от индекси с постоянна база към верижни индекси и обратно***?

Такава възможност има само за индексите с постоянни тегла.

Да приемем, че са дадени индексите с постоянна база с постоянни тегла:

$$\frac{\sum p_1 q_c}{\sum p_0 q_c}, \frac{\sum p_2 q_c}{\sum p_0 q_c}, ..., \frac{\sum p_{n-1} q_c}{\sum p_0 q_c}, \frac{\sum p_n q_c}{\sum p_0 q_c}.$$

Ако се раздели всеки от индексите на предхождащия го индекс, ще се получат верижните индекси с постоянни тегла:

(11.32)
$$\frac{\sum p_2 q_c}{\sum p_0 q_c} : \frac{\sum p_1 q_c}{\sum p_0 q_c} = \frac{\sum p_2 q_c}{\sum p_1 q_c};$$

$$\frac{\sum p_3 q_c}{\sum p_0 q_c} : \frac{\sum p_2 q_c}{\sum p_0 q_c} = \frac{\sum p_3 q_c}{\sum p_2 q_c}; ...;$$

$$\frac{\sum p_n q_c}{\sum p_0 q_c} : \frac{\sum p_{n-1} q_c}{\sum p_0 q_c} = \frac{\sum p_n q_c}{\sum p_{n-1} q_c}.$$

Ако са дадени верижните индекси с постоянни тегла, като се умножи всеки от тях на всички предходни, ще се получат индексите при постоянна база с постоянни тегла:

(11.33) $$\frac{\sum p_2 q_c}{\sum p_1 q_c} \cdot \frac{\sum p_1 q_c}{\sum p_0 q_c} = \frac{\sum p_2 q_c}{\sum p_0 q_c};$$

$$\frac{\sum p_3 q_c}{\sum p_2 q_c} \cdot \frac{\sum p_2 q_c}{\sum p_1 q_c} \cdot \frac{\sum p_1 q_c}{\sum p_0 q_c} = \frac{\sum p_3 q_c}{\sum p_0 q_c}; \ ...;$$

$$\frac{\sum p_n q_c}{\sum p_{n-1} q_c} \cdot \frac{\sum p_{n-1} q_c}{\sum p_{n-2} q_c} ..... \frac{\sum p_2 q_c}{\sum p_1 q_c} \cdot \frac{\sum p_1 q_c}{\sum p_0 q_c} = \frac{\sum p_n q_c}{\sum p_0 q_c}.$$

Очевидно е, че този косвен метод, приложен по отношение на индексите с постоянни тегла, дава резултати, напълно идентични с тези, които се получават по прекия метод.

При индексите с променливи тегла прекият и косвеният метод дават по правило различни резултати.

Това се дължи на обстоятелството, че връзката между верижните индекси и индексите с постоянна база се прекъсва при промяна на теглата, тъй като те за всеки отделен подпериод съдържат различен кръг единици, т.е. чрез тях се обхващат различни съвкупности:

(11.34) $$\frac{\sum p_2 q_2}{\sum p_0 q_2} : \frac{\sum p_1 q_1}{\sum p_0 q_1} \neq \frac{\sum p_2 q_2}{\sum p_1 q_2} \neq \frac{\sum p_2 q_1}{\sum p_1 q_1};$$

(11.35) $$\frac{\sum p_3 q_3}{\sum p_0 q_3} : \frac{\sum p_2 q_2}{\sum p_0 q_2} \neq \frac{\sum p_3 q_3}{\sum p_2 q_3} \neq \frac{\sum p_3 q_2}{\sum p_2 q_2}.$$

Въпреки това, в известни случаи е възможно поредица от верижни индекси с променливи тегла да се умножат и да се получи нов индекс. Неговото съдържание и икономически смисъл обаче ще бъдат различни от съдържанието и смисъла на индекса с постоянна база, изчислен пряко за същия период, тъй като ще съдържа влиянието на промените в теглата между подпериодите. Затова той трябва да се интерпретира не просто като индекс на средното изменение на равнищата през целия период за определена по състав съвкупност, а като характеристика на изменението на равнищата при съответно изменение и на състава на съвкупността от един подпериод до друг.

11.8. Обща характеристика на териториалните индекси

Необходимостта от териториални индекси се налага от интереса към различията, които съществуват между обособени териториални единици по определени икономически и други показатели. Този интерес се засили особено в наше време във връзка с международните сравнения, сравненията по икономически райони, по административно-териториални единици, по стопански организации и др.

Главната особеност на териториалните индекси, която поражда и особености в тяхната методология, е това, че те измерват ***различия по място*** (пространствени различия) за явления, взети в ***статично*** състояние за един период или момент. При динамичните индекси винаги има един изходен (базов) период или момент, след който явлението се изменя и достига в някой следващ период или момент друго равнище или друг обем. При териториалните индекси няма изменение във времето. Има два района и с еднакво основание всеки от тях може да бъде базов или индексиран район. В един случай може например да се сравнява производителността на труда (при определено производство) на Р България с това на Австрия, а в друг случай - обратно.

Във връзка с това възниква основният проблем за системата на претеглянето, който все още не е получил цялостно и еднозначно решение.

И при териториалните индекси на средните различия основна форма е агрегатната. Но при построяването на множествените териториални индекси на равнища и обем трябва съответно да се изберат теглата или съизмерителите. Трудността възниква при съвкупности, единиците на които не могат да се сумират в натурално изражение и следователно не е възможно изчисляването на средна величина (средно равнище).

Тук ще изложим някои възможни решения при построяването на териториални индекси на равнища. Принципно по същия начин могат да се построят и териториалните индекси на обеми.

Съществува възможност за тегла да се вземат количествата на един от районите. Бихме имали две възможни решения, при които формулите ще имат следния вид:

(11.36) $$I_{p_{A/B}(q_A)} = \frac{\sum p_A q_A}{\sum p_B q_A};$$ (11.37) $$I_{p_{B/A}(q_B)} = \frac{\sum p_B q_B}{\sum p_A q_B};$$

(11.38) $$I_{p_{A/B}(q_B)} = \frac{\sum p_A q_B}{\sum p_B q_B};$$ (11.39) $$I_{p_{B/A}(q_A)} = \frac{\sum p_B q_A}{\sum p_A q_A}.$$

(Тук с A и B са означени първият и вторият от съпоставяните райони).

Териториалните индекси, изчислени по този начин, имат смисъл тогава, когато е необходимо да се измерят различията в равнищата на даден район в сравнение с друг, но при структура на индексирания район. Когато например една страна съпоставя производителността на труда в един свой отрасъл на промишлеността с производителността на труда в същия отрасъл на друга страна, тя може да се интересува от различията при нейната структура на производството в отрасъла. В такъв случай индексът на производителността на труда се изчислява при национални тегла. Индексите, изчислени по такъв начин, обаче са сравними само за отделно взет район (респ. страна). България например може да сравнява редица от индекси за своята производителност на труда, изчислени при база всяка друга страна. Сравнимостта на индексите в случая е осигурена от обстоятелството, че всички индекси са с еднакви тегла - националните тегла на България. Ако обаче всеки район (респ. страна) изчислява своите индекси при свои тегла, индексите за отделните райони (страни) са несравними помежду си поради различната структура на теглата.

Индексите по посочените формули, изчислени за два района при взаимно обратни тегла, не са взаимно обратни по своите величини, щом е различна структурата на теглата, т.е.

(11.40) $$I_{p_{A/B}q_A} . I_{p_{B/A}q_B} \neq 1.$$

Поради това могат да се получат противоречиви резултати. Може например индексите, изчислени с национални тегла, да покажат, че

производителността на труда в България е по-висока от тази в същия отрасъл на Австрия и в същото време, че производителността на труда в Австрия е по-висока от тази в България. Такива "парадокси" се предизвикват от различната структура на отрасъла в двете страни, отразена чрез националните тегла. При тълкуване на различията между индексите с различни тегла при териториалните сравнения е недопустима аналогията с динамичните индекси поради посочените вече особености на териториалните индекси.

Друго възможно решение е използването на ***стандартни тегла***, които не са нито от единия, нито от другия район, а са общи за двата района, приети са като стандартни. При стандартни тегла q_S индексите за двата района ще имат следния вид:

$$(11.41) \qquad I_{p_{A/B}(q_S)} = \frac{\sum p_A q_S}{\sum p_B q_S}; \quad (11.42) \qquad I_{p_{B/A}(q_S)} = \frac{\sum p_B q_S}{\sum p_A q_S}.$$

Очевидно е, че при този начин на претегляне може да има само по един индекс за всеки район. Не могат следователно да се получат и противоречиви резултати при съпоставянето на двата района. Съществен момент при тази система на претегляне е обаче определянето на стандартните тегла. В някои случаи, когато се съпоставят само два района, те могат да се получат като сума общо за двата района. В други случаи, особено когато на съпоставяне подлежат повече от два района, стандартните тегла следва да се съставят на по-широка основа при конкретна икономическа обосновка.

При някои международни и други териториални сравнения се прилага средногеометричният индекс, изчислен от два индекса, от които единият е с тегла от района, за който се отнася сравнението, а другият - с тегла от района, приет за база на сравнението (по формулата на И. Фишер). Индексите за всеки от двата района в този случай имат следните общи формули:

$$(11.43) \qquad I_{p_{A/B}} = \sqrt{\frac{\sum p_A q_A}{\sum p_B q_A} \cdot \frac{\sum p_A q_B}{\sum p_B q_B}};$$

$$(11.44) \qquad I_{p_{B/A}} = \sqrt{\frac{\sum p_B q_B}{\sum p_A q_B} \cdot \frac{\sum p_B q_A}{\sum p_A q_A}}.$$

Този начин за изчисляване на териториалните индекси се обяснява главно с това, че чрез геометричното осредняване на двата индекса се неутрализира влиянието на различната специализация на районите, изразено в състава на теглата, и се осигурява ***обратимост в пространството***, при която индексите за двата района са взаимно обратни величини и следователно тяхното произведение е равно на единица:

$$(11.45) \qquad I_{p_{A/B}} . I_{p_{B/A}} = 1.$$

При конкретни териториални индекси на равнища и на обеми възникват специфични методологични проблеми. Широк кръг проблеми се решават и при многостранните международни сравнения, осъществявани по програми на Организацията на обединените нации.

11.9. Практикум

11.9.1. Въпроси за самопроверка

1. Каква е същността на индексния метод и какви са неговите познавателни функции?
2. Кои индекси се наричат единични и кои множествени (сложни, съвкупностни)?
3. Кои индекси се наричат индекси на равнища?
4. Какъв е принципът за трансформиране на агрегатни индекси в средноаритметични и среднохармонични?
5. От какво се обуславя разликата между индексите на Ласпер и Пааше?
6. В какво се състои разликата между индексите на средното изменение на равнищата (индекси при постоянен състав) и

индексите на средните равнища (индекси при променлив състав)?

7. Какво се разбира под инвариантно свойство на индексното отношение?
8. В какво се изразява връзката между индексите на равнище, на обем и на маса?
9. В какво се състои проблемът за теглата при териториалните индекси?
10. В какви случаи може да се използва индексът на И. Фишер и какво представлява критерият “обратимост в пространството”?

11.9.2. Задачи за упражнение

Задача 1. В следващата таблица се съдържат примерни данни на една фирма.

Таблица 11.4

Продадени стоки и продажни цени на фирма “Н” през 2005 г. и 2007 г.

Вид на продукцията	Мярка	2005 г.		2007 г.	
		Продадени количества	Цени за единица-лв.	Продадени количества	Цени за единица-лв.
А	кг.	12000	8,5	20000	10,0
Б	бр.	8000	12,0	25000	20,0
В	бр.	50000	15,0	40000	14,0

Иска се:

а) Да се изчислят единичните индекси на продадените количества стоки и на цените по видове продукция.

б) Да се изчислят индексите на средното изменение на цените по формулите на Ласпер и Пааше:

- агрегатни;

- средноаритметични;
- среднохармонични.

в) Да се изчисли индексът на И. Фишер.

г) Да се изчислят индексите на физическия обем на продажбите по цени на 2005 г. и на постъпленията (оборота) от продажбите.

д) Да се обясни разликата между индексите на цените на Ласпер и Пааше.

е) Да се провери връзката между индексите на равнище, обем и маса.

Отговори:

б) $I_{p(q_0)}$ = 1,034 = 103,4% (Ласпер)

$I_{p(q_1)}$ = 1,178 = 117,8% (Пааше)

в) I_p = 1,104 = 110,4% (Фишер)

г) $I_{q(p_0)}$ = 1,129 = 112,9%

I_{qp} = 1,329 = 132,9%

е) I_{qp} = 1,178.1,129 = 1,329 = 132,9%

Задача 2. Една стока е продавана на 3 пазара по различни цени. През 2007 г. са настъпили значителни изменения спрямо 2005 г. както в цените, така и в продадените количества. Данните се съдържат в следващата таблица.

Таблица 11.5

Продадени количества и продажни цени за стока "А" по пазари през 2005 г. и 2007 г.

Пазари	2005 г.		2007 г.	
	Количества - кг	Цени – лв.	Количества - кг	Цени – лв.
А	2000	2,50	1800	2,00
Б	5000	3,00	4200	3,00
В	3000	2,00	6000	5,00
	10000	х	12000	х

Иска се:

1. Да се изчислят единичните (по пазари) индекси на цените и на продадените количества през 2007 г. спрямо 2005 г.
2. Да се изчисли индексът на средните продажни цени (при променлив състав) общо за трите пазара през 2007 г. спрямо 2005 г.
3. Да се изчисли индексът на средното изменение на цените (при постоянен състав) общо за трите пазара през същия период.
4. Да се обясни на какво се дължи разликата между индексите на средната цена и на средното изменение на цените.