Софийски Университет "Св.Климент Охридски"

Физически факултет

Людмил Цанков

Планиране на експеримента и обработка на експериментални данни (Записки на лекции)

София, 2000г.

Трето издание, 2003г.

Съдържание:

I. Увод

1. Обща представа за физичните измервания

Експерименталното изучаване на закономерностите в природата е свързано с измерване на определени физически величини, които характеризират количествено тези закономерности. Резултатите от тези измервания заедно с всички останали величини, отнасящи се до състоянието на изследваната физична система, се наричат ***първични експериментални данни***.

Обработката на експерименталните данни се изразява в прилагането на различни математични и статистически методи върху данните и има за цел извличането на съдържащата се в тях информация във възможно най-пълен вид.

Физичните експерименти и измервания са толкова разнообразни, че тук е възможно накратко да се спрем само на някои най-общи свойства.

Все пак типовете експерименти според характера на наблюденията и от гледна точка на обработката на данните могат да се разделят на следните крайни категории:

- *"уникални" експерименти и наблюдения*; при тях ние практически нямаме контрол върху условията на експеримента и е много трудно или невъзможно да го повторим (космични явления, междупланетни сонди, ускорители). В тези случаи е необходимо дълго време за предварителна подготовка, евентуално щателно моделиране на явлението с цел да може да бъде събрана възможно най-пълно и бързо нужната експериментална информация. За обработката на резултатите от такива експерименти обикновено има достатъчно време след наблюдението и тя може да бъде извършена много внимателно и от квалифицирани специалисти;
- *масови анализи*; такива са измерванията при процеси, развиващи се във времето, при технологичен контрол и др. В тези случаи постъпващата информация е по-еднообразна по характер, но е в голямо количество и тя трябва да бъде обработена в реално време (до идването на данните от следващото измерване), тъй като е необходима за преценка на по-нататъшните ни действия. Тук в целия процес на обработка трябва да се разчита на напълно автоматични методи за анализ без участието на

експериментатора, още повече че не винаги в тези случаи може да се разчита на достатъчно квалифициран персонал.

Ние ще си представяме, че сме по-скоро в ситуация от втория вид и ще разглеждаме такива методи за обработка на данни, които да позволяват подобен автоматичен анализ.

Измерванията се делят на ***преки*** и ***косвени*** в зависимост от това, дали при измерването ние получаваме някаква стойност на величината, която ни интересува непосредствено, или пък получаваме някаква друга физична величина, която е косвено свързана с нея.

По отношение на числовия си характер физичните величини се делят на ***дискретни*** и ***непрекъснати.*** Всички физични величини, характеризиращи свободното състояние на системите, по принцип се изменят непрекъснато, т.е. могат да заемат стойности в един континуум. Такива са например величините в механиката, в термодинамиката, повечето от електричните и оптичните величини. От друга страна, знаем, че системите на атомно и субатомно ниво имат много характеристики, които се изменят дискретно и на техните стойности може да се съпостави някакво изброимо множество от числа. Такива са стойностите на редица характеристики на елементарните частици (напр. ел.заряд), различните квантови числа и други характеристики на свързаните системи (като напр. енергетични нива, моменти на импулса и др.). Дискретните и непрекъснатите величини имат принципно различни статистически свойства и затова предварителното познаване на природата на измерваните величини е важно.

За да могат експерименталните данни да бъдат обработвани, те трябва да бъдат представени като числа. Този процес на преобразуване на данните трябва да се познава в детайли от експериментатора, тъй като при него могат да се допуснат (и се допускат) допълнителни грешки.

При дискретно изменящите се величини нещата стоят по-просто, тъй като процесът на измерването им всъщност се свежда до еднозначното им съпоставяне с определени стойности (числа) от дискретното множество, към което тези величини принадлежат. Единствената възможност за внасяне на грешки при такива измервания е да се съпостави друга стойност от същото множество (напр. да се разпаднат 10 ядра за една секунда, а ние да отчетем 11 събития). Това може да се случи главно при грешки в регистриращата

апаратура. Самата апаратура в този случай обаче е достатъчно проста и се състои от някакво устройство, което регистрира настъпването на изучаваното събитие и изпраща получената информация в запомнящ блок.

Когато обаче ние измерваме непрекъснато изменящи се величини, ние явно или неявно извършваме още едно преобразование (трансформация) на един аналогов сигнал в цифров. Типичната схема за събиране на данни в един такъв физичен експеримент се състои от датчик, преобразувател на сигнала, аналогово-цифров преобразувател, запомнящо устройство и управление.

Главната част от задачата се свежда, по същество, до конструирането на датчик (детектор, сензор), който да преобразува измеряемата физична величина в електричен потенциал. Изискването за електричен характер на сигнала на изхода от датчика е необходимо, тъй като такива сигнали лесно се преобразуват и обработват по-нататък. Като правило (макар и не винаги), електричният потенциал на изхода на детектора е пропорционален на измеряемата физична величина. Такива прибори например са микрофоните, тензометричните датчици, луксметрите и мн.др.

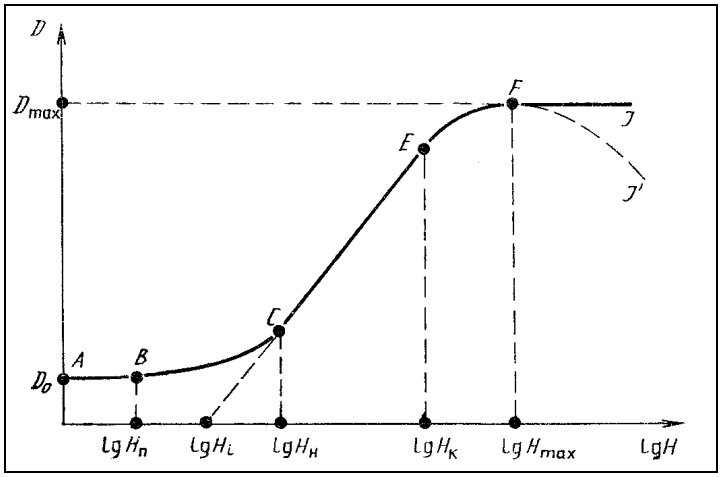
Всеки датчик се характеризира с т.нар. ***предавателна функция*** $R(E,U)$ (нарича се също ***функция на отклика***, response function), която се определя като функция на разпределение на амплитудата на сигнала на изхода на датчика при въздействието върху входа му на физична величина с точно определена големина:

$$S(U) = \int_{-\infty}^{\infty} dE P(E) R(E,U). \qquad (1)$$

Тук $P(E)$ е разпределението по енергии на сигнала на входа, а $S(U)$ е изходният сигнал. Познаването на тази функция е особено необходимо при определянето на влиянието на детектора върху получените наблюдаеми резултати.

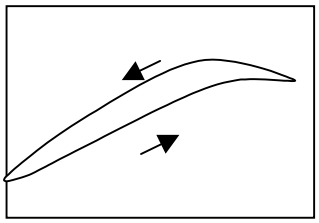
Обикновено датчикът се избира така, че неговата функция на отклика да е линейна в границите на обикновено използуваемия диапазон от стойности на

измерваната величина. Разбира се, невъзможно е да се създаде идеално линейно устройство и този интервал на линеен отклик е ограничен.



Типична функция на отклик (характеристична крива на фотоматериал). Линейната част е между точките C и E.

Друг възможен източник на грешки при работата с датчици е т.нар. *хистерезис (хистерезисна крива)*, т.е. нееднаквата траектория на изходния сигнал при преминаването на входящия сигнал от началната към крайната стойност на измеряемия диапазон и обратно. В такъв случай резултатите на изхода на датчика зависят от предисторията на неговото състояние. За добре конструираните датчици хистерезисните грешки обикновено са пренебрежимо малки.



Хистерезис на датчик

Процесът на преобразуване на аналоговия сигнал в цифров е специфичен и той внася допълнителни смущения в данните, чиято природа трябва да се познава. АЦП (ADC) представлява електронно устройство, на чийто вход се подава аналоговият сигнал (евентуално усилен) от детектора, а на изхода му се генерира цифров код, съответстващ на нивото (амплитудата) на входния сигнал, при което целият диапазон от напрежения на входа се трансформира линейно в целочислов диапазон на изхода (обикновено обемът му е някаква степен на числото 2, напр. 0-1023, 0-4095 и т.н.). По този начин чрез АЦП ние всъщност построяваме една хистограма на разпределението на амплитудата

на входящия сигнал (който може да бъде непрекъснат или импулсен). Данни от такъв тип, които се характеризират с еквидистантна скала на аргумента, ще наричаме ***данни от спектрален тип*** или дори само ***"спектри***", като си даваме сметка, че това название е условно. Равномерната скала на аргумента предоставя много допълнителни удобства при анализа на данните (напр. бързата Фурие (Fourier) трансформация много съществено се основава на това обстоятелство). Понякога дори се налага да преминаваме от неравномерна към равномерна скала на аргумента на данните, които подлежат на обработка.

Тук ние нямаме възможност да се спираме на принципа на действие на този важен прибор (АЦП), а само накратко ще посочим влиянието му върху качеството на получените чрез него експериментални данни. Това влияние се проявява в няколко посоки:

- Поради грешки в преобразуването е възможно за един и същ по големина входен сигнал да се генерира различен цифров код (т.е. той да попадне в различни канали на изхода);
- Ширините на отделните канали може да не са еднакви, което също води до изкривяване на полученото разпределение;
- Коефициентът на преобразуване може да се мени с времето (пълзене) поради температурни и др. влияния, което води до получаване на грешни данни.

Основен принцип при обработката на информацията е задължителното съхраняване на първичните експериментални данни в непроменен вид, независимо от всички последващи действия върху тях. Това позволява да имаме възможност винаги да се връщаме към първоначалните данни при грешки в обработката или при използуване на други методи за анализ.

Литература:

[1]. Р.Отнес, Л.Эноксон: Прикладной анализ временных рядов, Мир, Москва (1982).

II. Първична обработка на данните

Под първична обработка на данните разбираме такива операции върху тях, които да позволяват грубо да се ориентираме в закономерностите, които тези данни изразяват. Обикновено първичната обработка е подготвителен етап към цялостния процес на анализ на данните. Резултатите от първичната обработка може да се използуват като начални приближения към по-нататъшните пресмятания. Още на този етап обаче може да се реши, че данните не представляват по-нататъшен интерес поради лошо поставен експеримент или грешки в работата на апаратурата.

Тук ние ще разгледаме последователно етапите в първичната обработка на данни.

2. Третиране на данните с неправдоподобни стойности

Всички системи за събиране на информация внасят лъжливи (грешни) стойности в данните. Добрите системи правят това все пак по-рядко. Това явление може да се дължи на различни причини като частично загубване на сигнали в линиите за връзка, грешки при дискретизацията на сигнала (АЦП), външни електрически смущения и мн.др. При това измежду данните се получават някои резултати със стойности, значително отличаващи се от очакваната последователност, които нямат нищо общо с изследвания процес, но могат да причинят големи неприятности при следващата обработка, още повече ако тя е автоматична. Особено неприятно е наличието на две или повече съседни неверни стойности на данните.

Поради това, първата задача на системите за обработка на данни (след като те бъдат надеждно съхранени в първоначалния им вид!) е издирването и отстраняването на такива данни. За съжаление, много трудно е да се установят общовалидни критерии за това, дали едно число е с очевидно невярна стойност, особено около екстремумите на изследваните зависимости, когато дори “правилното” поведение на данните е трудно да бъде описано, както и в случаите, когато отклонението от действителната стойност не е очебийно голямо. Основното изискване към процедурите, които ще разглеждаме, е те да могат да действуват напълно автоматично и без човешка намеса (както е при масовите анализи).

Най-напред ще посочим какво може да бъде направено след откриването на невярна стойност. Основните възможности са три:

- Измерването да се повтори точно при тези условия и да се използува резултатът от второто измерване. Това обаче е практически неосъществимо в много от случаите, напр. при данните от спектрален тип, когато всички резултати се получават съвместно. Обикновено е по-лесно да се повтори цялото измерване;
- За стойност на дадения канал да се припише някакво по-достоверно число (напр. усреднената по някакъв начин стойност на неговите съседни);
- Съмнителната стойност да не се използува въобще при обработката, (напр. като й се припише нулево относително тегло). Това е от теоретична гледна точка най-правилният подход, тъй като при него единствено не се внасят нови смущения във вида на изследваната функция (а само се намалява получената от експеримента информация).

Всички методи за отстраняване на данни с неправдоподобни стойности изхождат от по-явно или неявно приеманото предположение, че нормалните данни определят една “гладка” зависимост, а ненормалните нарушават тази картина на благополучие. Затова ние трябва да се занимаем най-напред с въпроса за количествения анализ на понятието “гладкост” в разглеждания контекст.

Нека имаме N на брой измервания с абсциси $\{x_i = i, i = 1,...,n\}$ и ординати $\{y_i = y(i), i = 1,...,n\}$. Еквидистантните стойности на абсцисите, които предполагаме тук за простота, се реализират винаги при данните от спектрален тип, но разглежданията по-долу могат да се пренесат без намаляване на общността и за неравноотстоящи данни (с малко повече писане). Естествено е да търсим количествена мярка за гладкостта на зависимостта $y(x)$ в точката y_i чрез ***локалната кривина*** на $y(x)$ в някаква околност на тази точка. Както е известно от анализа, кривината на една функция $y(x)$ в точката x се дефинира с:

$$c(x) = y''(x) / \sqrt{1 + y'(x)^2} \qquad (1)$$

където с щрихове са показани първата и втората производна на $y(x)$. Тъй като предполагаме, че извън стойностите на измерванията $\{i, i=1,...,n\}$ за функцията y не ни е известно нищо, то оценки за нейните производни могат да бъдат направени само чрез апроксимирането им с крайни разлики:

$$y_i' = y_{i+1} - y_i$$
$$y_i'' = y_{i+1} - 2y_i + y_{i-1} \qquad (2)$$

при условие: $x_{i+1} - x_i = 1.$ (Ако данните са неравноотстоящи, (2а) би имало вида:

$$y_i' = \frac{y_{i+1} - y_i}{x_{i+1} - x_i}; \qquad (2a')$$

това подсказва как биха изглеждали обобщенията на разглежданията по-долу за неравноотстоящи абсциси на измерванията).

Преминавайки към нашата задача, ще разгледаме едно семейство числови филтри, основаващи се на минимизацията на локалната кривина (точно тази минимизация на кривината е формалният начин да построим най-гладката зависимост между дадените точки). Тези филтри ще конструираме така:

- За всяка точка i филтърът $F(k,l)$ използува стойностите на y в точките $\{i-k,...,i+l\}$ без точката i (която предполагаме, че е съмнителна и подлежаща на проверка);
- $F(k,l)$, приложен върху y_i, осигурява $\min \sum_{j=i-k+1}^{i+l-1} y_j''^2$. Това е равносилно на построяването на най-гладката крива през точките $\{i-k,...,i+l\}$ и определянето на една изгладена (очаквана) стойност вместо несигурната y_i от стойността на гладката крива - $\bar{y}_i$. Тук y'' апроксимира кривината в предположение, че измененията на функцията не са много големи $(y' \approx const)$. На свой ред, изразът под сумата е средният квадрат на кривината в разглеждания интервал.

И тъй, $F(k,l)$ се дефинира от изискването:

$$F(k,l): \sum_{j=i-k+1}^{i+l-1} (y_{j-1} - 2y_j + y_{j+1})^2 = \min_{\{y_i\}} \qquad (3)$$

Това е уравнение за $\bar{y}_i$- изгладената стойност, която се получава по стандартната процедура за търсене на минимум от анализа – намиране на производната на $F(k,l)$ спрямо y_i и приравняването ѝ към 0. Полученото равенство разглеждаме като уравнение спрямо търсеното $\bar{y}_i$.

Ето как изглеждат последователните членове на това семейство:

$F(1,1): y_{i-1} - 2\bar{y}_i + y_{i+1} = 0$. Следователно:

$$\bar{y}_i = \frac{1}{2}(y_{i-1} + y_{i+1}) \qquad (4)$$

Това е познатият ни израз за ***линейна интерполация***. С други думи, най-гладката крива, минаваща през две точки - една вляво и една вдясно от разглежданата - е права линия, която в точката y_i има стойността (4).

По-нататък:

$F(2,0): 2(y_{i-2} - 2y_{i-1} + \bar{y}_i = 0$. Или:

$$\bar{y}_i = 2y_{i-1} - y_{i-2}. \qquad (5)$$

Това е ***линейна екстраполация*** в посока напред. Следва:

$$F(2,1): (y_{i-2} - 2y_{i-1} + \bar{y}_i) + (-2)(y_{i-1} - 2\bar{y}_i + y_{i+1}) = 0.$$

Или:

$$\bar{y}_i = \frac{1}{5}(-y_{i-2} + 4y_{i-1} + 2y_{i+1}). \qquad (6)$$

Това е ***квадратична интерполация*** (една точка напред и две назад). Аналогично се пресмятат $F(0,2)$ и $F(1,2)$. Сега идва:

$$F(22): (y_{i-2} - 2y_{i-1} + \bar{y}_i) + (-2)(y_{i-1} - 2\bar{y}_i + y_{i+1}) + (\bar{y}_i - 2y_{i+1} + y_{i+2}) = 0.$$

Оттук:

$$\bar{y}_i = \frac{1}{6}(-y_{i-2} + 4y_{i-1} + 4y_{i+1} - y_{i+2}). \qquad (7)$$

Това е вече ***квадратично изглаждане*** (най-добрата парабола през 4 точки – две напред и две назад).

Тъй като производните от втори ред засягат само съседните точки, можем да се убедим, че включването на повече точки не добавя нови членове в това семейство, така че напр. $F(3,3) = F(2,2)$.

Стратегията за търсене на “изхвърчали” точки се формулира така:

- Намираме две съседни точки, за които считаме, че са правилни (започваме от началните точки y_0, y_1. Ако някоя измежду точките $\{2,3\}$ се

окаже с неправдоподобна стойност по следващия критерий, то считаме y_0, y_1 за съмнителни, отхвърляме ги и караме нататък);

- Образуваме $\bar{y}_i = 2y_{i-1} - y_{i-2}$, (от $F(2,0)$). Ако: $|\bar{y}_i - y| < \lambda\sigma$, приемаме, че точката y_i е нормална и продължаваме нататък. Тук $\lambda\sigma$ е мярка за някакво приемливо отклонение. Тя се определя от очакваната неопределеност на данните σ, която може да се прецени от допълнителни съображения (с този въпрос ние се занимаваме по-късно; засега ще кажем, че λ се избира от порядъка на 3-6). В противен случай следва да направим извода, че поведението на нашата последователност $\{y\}$ в точката y_i не може да се опише с линейна функция. Тогава имаме две възможности:
  - Или y_i е редовен екстремум на функцията (min или max);
  - Или y_i е грешна.

За да преценим коя от двете възможности се реализира, образуваме:

$$\tilde{y}_i = F(2,1) = \frac{1}{5}(-y_{i-2} + 4y_{i-1} + 2y_{i+1}) \qquad (8)$$

Тук надеждата ни е, че y_i и y_{i+1} не са едновременно изхвърчали (или в екстремума). Ако това все пак се окаже така, тази процедура може постепенно да ни отведе далеч от нормалните точки, които в крайна сметка трябва някога да се появят.

Ако $|\tilde{y}_i - y_i| < \lambda_1\sigma$, y_i се счита за нормален екстремум и се оставя. В противен случай y_i се счита за "изхвръкнала" точка и се игнорира.

Съблазнително е да се използува филтърът $F(2,2)$ за апроксимация на y_i, тъй като изглаждането дава сигурност спрямо интерполацията (екстраполацията), но съществуват две опасности:

- При $F(2,2)$ се налага да използуваме две непроверени точки едновременно; това увеличава възможността за грешка;
- Изглаждането може да даде лош резултат около екстремумите (голяма разлика $|\tilde{y}_i - y_i|$, без точката да е с неправдоподобна стойност.

В заключение ще кажем, че противно на правния принцип – “по-добре 10 виновни на свобода, отколкото един невинен в затвора”, за нас е по-добре да

изхвърлим повече точки от необходимото, отколкото да оставим една грешна. Такава стойност може да се намеси на много по-късен етап от обработката и да объркa процеса след като вече са положени значителни усилия за пресмятанията. Във всички случаи трябва да сме много внимателни, ако се окаже, че трябва да се изхвърлят последователни точки, а също и около екстремумите на последователностите от данни, защото истинските екстремуми съдържат много важна и необходима информация за поведението на данните.

3. Интерполация и изглаждане на данните

3.1. Интерполация на данните

След като вече сме премахнали данните с груби грешки в стойностите, следващият етап в първичната обработка на данни е да получим възможност да пресмятаме стойностите на данните в междинните точки между измерванията. Съвкупността от методи, използувани за тази цел, се определя с общото название ***интерполация на данните***.

Такава необходимост възниква винаги, когато искаме да търсим екстремуми (минимуми и максимуми) в последователности от непрекъснато изменящи се данни, които, разбира се, едва ли се намират точно при стойностите на аргумента, за които имаме измервания. Понякога ние искаме да разполагаме не само с някакви стойности на самата последователност, но и с нейните производни в междинни точки или интеграли в някакъв интервал от стойности на аргумента. Във всички такива случаи ние трябва да разполагаме с някакъв алгоритъм за интерполация. Разбира се, интерполацията се използува не само при обработката на експериментални данни, но и при моделирането на данни, когато в някои случаи пресмятането на пълния модел в достатъчен брой точки изисква големи изчислителни усилия. Тогава съчетаването на пълни пресмятания в ограничен брой точки с подходяща интерполационна процедура може да ускори многократно решаването на цялостния проблем.

Формално задачата за интерполация се формулира така:

Нека са дадени една съвкупност от реални абсциси $\{x_i, i=1,...,n\}$ и стойности на ординатите $\{y_i, i=1,...,n\}$. Задачата на едномерната интерполация се състои

в това, да се построи функция $f(x)$ такава, че $f(x_i) = y_i$ за всяко i и $f(x)$ да има някакви приемливи стойности между стойностите x_i .

Разбира се, същината на задачата за интерполация е дефинирането на "разумното" поведение на интерполиращата функция в междинните точки. Очевидно зададените точки могат да бъдат интерполирани по безброй много начини и ние трябва да имаме критерии за избор между тях. Обикновено критериите за избор се формулират в понятия като *простота* и *гладкост* на представянето - напр. $f(x)$ да бъде диференцируема и с непрекъснати производни до определен ред в целия интервал, или да бъде възможно най-гладка в този интервал, или да бъде полином от най-малка степен и т.н.

Класически пример за интерполация е полиномната интерполация, която ще си припомним накратко.

3.1.1. Полиномна интерполация

Множеството от алгебрични полиноми е най-важният клас интерполационни функции поради някои свои особено благоприятни свойства. Знаем, че полиномите лесно се пресмятат, събират, умножават, диференцират и интегрират, полиномите от полиноми също са полиноми и т.н.

Полиномите са подходящи за апроксимация (и в частност за интерполация) на функции поради обстоятелството, че всяка непрекъсната функция може да се апроксимира с произволна точност върху всеки краен затворен интервал чрез полином (теорема на Вайерщрас (Weierstrass) за апроксимация: ако $f(x)$ е непрекъсната функция върху крайния затворен интервал $[a,b]$, за всяко $\varepsilon > 0$ съществува полином $p_n(x)$ от степен n такъв, че $\max_{x\in[a,b]} |f(x) - p_n(x)| < \varepsilon$).

Разбира се, степента на такъв апроксимиращ полином зависи от желаната точност на апроксимацията (измервана с числото ε). Тази степен може да е толкова висока, че полиномът да е практически неизползуваем.

Всеки полином $p_n(x)$ може да се представи в ред по степените на x:

$$y(x) = \sum_{i=0}^{n} a_i x^i . \qquad (1)$$

Този ред има $n+1$ коефициенти. За да бъдат определени тези коефициенти еднозначно, необходимо е да се определят $n+1$ независими условия. При интерполация обикновено се иска полиномът да минава през всичките $n+1$

(различни) точки $\{x_i\}$. Това води до система от $n+1$ линейни алгебрични уравнения:

$$y_i = \sum_{j=0}^{n} a_j x_i^j \quad (i = 0..n).\,, \tag{2}$$

които трябва да бъдат решени относно неизвестните коефициенти $\{a\}$. На нас ни е известно, че такава система има единствено решение. Така че винаги е възможно прекарването на единствен полином от степен $\le n$ през всеки $n+1$ различни абсциси $\{x_i\}$. Такъв подход, при който цялата съвкупност данни се интерполира с една единствена функция, се нарича ***<u>глобална интерполация.</u>***

След като сме решили да интерполираме данните глобално с полином от n степен, ние все още имаме свободата да изберем *<u>базисните функции,</u>* които трябва да принадлежат на пространството на полиномите от степен n. В написаното по-горе използувахме най-простия полиномен базис - от едночлените $\{x^0, x^1, x^2, ..., x^n\}$. Системата линейните уравнения (2) може по принцип да бъде решена с някой от числените методи на линейната алгебра (например чрез Гаусово (Gauss) елиминиране). Проблемът е, че в много случаи тези системи уравнения са твърде *<u>лошо обусловени</u>* (с други думи, решенията им са много чувствителни към грешки от закръгляване). Тази чувствителност е толкова силна, че практически подобни линейни системи се решават с трудности дори за $n > 10$.

Един много по-приемлив начин за построяване на интерполационен полином е за базис да се използуват известните *<u>полиноми на Лагранж</u>* (Lagrange). За разлика от разгледаните едночлени, тези полиноми се построяват специфично за всяко конкретно множество от точки $\{x_i\}$. Те имат свойството:

$$l_j(x_i) = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases}. \tag{3}$$

Вижда се, че полиномът:

$$l_j(x) = \frac{\prod_{i=0, i\neq j}^{n} (x - x_i)}{\prod_{i=0, i\neq j}^{n} (x_j - x_i)} \tag{4}$$

удовлетворява тези условия (всеки множител в числителя анулира полинома за някое $i \neq j$; множителите в знаменателя са за нормировка, за да

осигурят условието $l_j(x_i)=1$). Тъй като има само един такъв полином, ясно е, че именно l_j е полиномът с тези свойства. Интерполиращата функция, представена чрез полиномите на Лагранж, има вида:

$$y_L(x)=\sum_{j=0}^{n} l_j(x)y_j \qquad (5)$$

Тази функция има във всяко x_i стойност точно y_i.

Тук трябва още веднъж да кажем, че полиномът от степен $\le n$, който минава през всичките $n+1$ точки, е *единствен.* Има различни полиномни интерполационни формули, построени върху различни базиси (напр. (2) и (5)), но при еднаква съвкупност от данни те всичките водят до един и същ полином. Различните представяния на този полином обаче могат да имат много различно поведение при пресмятания с ограничена аритметична точност, каквато е всяка машинна аритметика. В този смисъл интерполационният полином на Лагранж (или някой друг подобен) е за предпочитане пред интерполацията, основана на базис от едночлените на аргумента.

В крайна сметка, въпреки своята простота, полиномната интерполация има много недостатъци. Те са свързани основно с:

- необходимостта от пресмятания на полиноми от висока степен, които увеличават изчислителните усилия;
- трудности, свързани с грешките от закръгляване;
- лошо поведение извън краищата на интервала (известно е, че асимптотично полиномите клонят към $\pm\infty$ при неограничено нарастване (както и намаляване) на аргумента, което не е така за реално съществуващите данни от физични измервания. Ето защо интерполационните полиноми в никакъв случай не трябва да се използуват за екстраполация на данните извън интервала;
- Вътре в самия интервал пък полиномите (особено тези от висок ред) могат да имат осцилации, които да са толкова големи, че да се обезсмисля възможността те да бъдат използувани за пресмятане на междинни стойности на данните.

Недостатъците на глобалната полиномна интерполация могат да бъдат значително редуцирани чрез използуването на *локална полиномна интерполация*.

3.1.2. Локална полиномна интерполация. Сплайни

Основната идея на локалната полиномна интерполация е да се използуват полиноми от сравнително ниска (3, 4,... до 10) степен, които да минават през няколко съседни точки; след това част от тези точки се отнемат отляво, а отдясно се прибавят други точки, това множество се интерполира с друг полином от подобна невисока степен и този процес продължава, докато се изчерпят всичките точки от разглежданата последователност. Такива построения се наричат още "плаващи полиноми".

Тъй като *сплайните* представляват по-нататъшно развитие на локалната полиномна апроксимация, което отразява и съвременното състояние на проблема, ние ще разглеждаме именно тях.

Първите сплайн-функции са предложени от Шьонберг (I.J.Schoenberg) (1946) и са били "слепени" от отделни парчета от кубични полиноми. Тази конструкция се е развивала и модифицирала чрез повишаване на степента на полиномите, промяна на условията във възлите и по границите и т.н. Съществен прогрес в развитието на теорията на сплайн-функциите е свързан с името на Холидей (J.C.Holladay)(1957), който е свързал сплайните с решаването на вариационна задача за минимума на потенциалната енергия на огъване на еластичен прът. Това е дало възможност по-нататъшните конструкции на сплайни да се свързват с различни вариационни задачи, отразяващи конкретните физични модели, свързани с произхода на данните.

Ние ще разгледаме по-подробно построяването на класическите кубични сплайни на Шьонберг.

Има смисъл да започнем разглеждането на сплайни именно от аналогията на тези функции с формите на еластичните пръти, каквито са били използувани отдавна от чертожниците за построяване на криви (оттам произлиза и наименованието "spline"). Механичният сплайн се закрепва (с помощта на тежести) в точките на интерполация (възлите) и при това той приема такава форма, че потенциалната му енергия е минимална. От механиката е известно, че тази потенциална енергия е пропорционална на линейния интеграл по дължината на дъгата от квадрата на кривината на сплайна.

Нека сплайнът се представя с функцията $s(x)$. Ако наклонът на тази функция не е голям, ние знаем, че кривината се апроксимира с втората производна на функцията $s''(x)$. Следователно енергията на подобен линеаризиран сплайн е пропорционална на $\int dx\,[s''(x)]^2$.

Формално задачата се дефинира така:
Дадени са абсцисите $\{x_i, i=1,...,n\}$ и ординатите $\{y_i, i=1,...,n\}$. Искаме да построим такава функция $s(x)$, че $s(x_i)=y_i, i=1,...,n$ и $\int_{x_1}^{x_n} dx\,[s''(x)]^2$ да е минимален.

Продължавайки аналогията с механичния сплайн, ние можем да очакваме $s(x)$ и $s'(x)$ да бъдат непрекъснати във възлите $\{x_i\}$ (иначе механичният сплайн ще се счупи). По-нататък от механиката следва, че между всеки два възела $s(x)$ е полином от трета степен и във всеки от възлите преходът между съседните полиноми се осъществява така, че $s(x)$, $s'(x)$ и $s''(x)$ са непрекъснати.

Доказано е, че ако към тези условия се прибавят условията по краищата: $s''(x_1)=s''(x_n)=0$ (такава функция се нарича ***естествен кубичен сплайн***), то тази функция е единствената функция (забележете: не само полином!), която има минимална кривина измежду всички функции, които интерполират данните и имат втора производна с интегрируем квадрат. В този смисъл може да се каже, че естественият кубичен сплайн е най-гладката функция, която интерполира множеството от данните.

Колко са всичките параметри на кубичния сплайн? Между възлите има $n-1$ интервала и в тях има $n-1$ части от кубични полиноми, всяка с по 4 параметъра. Следователно трябва да се наложат общо $4n-4$ условия, за да се определят еднозначно параметрите на сплайна. Условията за непрекъснатост на $s(x)$, $s'(x)$ и $s''(x)$ във вътрешните точки дават $3(n-2)$ условия. Условието $s(x_i)=y_i, i=1,...,n$ дава още n условия за s (дотук общо $4n-6$). Необходими са още две условия за пълното определяне на сплайна. Ако приемем $s''(x_1)=s''(x_n)=0$ (това е равносилно на предположението, че извън множеството от данни сплайнът е линейна функция), то получаваме естествения кубичен сплайн, за който вече стана дума. Разбира се, вместо тези две условия може да се наложат някакви други две условия.

Сега ще видим как се построява кубичен сплайн. Разглеждаме кривата в подинтервала $[x_i, x_{i+1}]$. Общият вид на сплайна в този подинтервал е:

$$s = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3 \qquad (1)$$

Нека да означим: $h_i = x_{i+1} - x_i; \quad w = \frac{x - x_i}{h_i}; \quad \tilde{w} = 1 - w.$ (2)

Удобството на тези означения е, че когато x се мени в този интервал, w се мени от 0 до 1, а $\tilde{w}$ - от 1 до 0. В тези означения е подходящо да представим сплайна във вида:

$$s(x) = wy_{i+1} + \tilde{w}y_i + h_i^2[(w^3 - w)\sigma_{i+1} + (\tilde{w}^3 - \tilde{w})\sigma_i], \qquad (3)$$

където σ_{i+1} и σ_i са константи, подлежащи на определяне. Първите два члена в (3) определят стандартна линейна интерполация, а членът в квадратните скоби е корекция от трета степен, която се анулира и в двата края на подинтервала. Така че (3) осигурява интерполация на данните независимо от избора на σ. Ние обаче ще определим константите $\{\sigma_i\}$ така, че да осигурим необходимата гладкост на сплайна.

Сега ще диференцираме $s(x)$ три пъти по x, за да можем да наложим необходимите условия за гладкост:

$$\begin{aligned} s'(x) &= \frac{y_{i+1} - y_i}{h_i} + h_i[(3w^2 - 1)\sigma_{i+1} - (3\tilde{w}^2 - 1)\sigma_i] \\ s''(x) &= 6w\sigma_{i+1} + 6\tilde{w}\sigma_i \\ s'''(x) &= \frac{6(\sigma_{i+1} - \sigma_i)}{h_i} \end{aligned} \qquad (4)$$

Нека да пресметнем $s'(x)$ в краищата на подинтервала, като вземем предвид, че тъй като $s(x)$ е дефинирана само в затворения подинтервал $[x_i, x_{i+1}]$, то на границите имаме само едностранни производни - отдясно за лявата граница и отляво за дясната граница:

$$\begin{aligned} s'_+(x_i) &= \Delta_{i-1} + h_{i-1}(2\sigma_i + \sigma_{i-1}) \\ s'_-(x_i) &= \Delta_i + h_i(-\sigma_{i+1} - 2\sigma_i) \end{aligned} \qquad (5)$$

Тук: $\Delta_i = \frac{y_{i+1} - y_i}{h_i}$ (6)

За да получим търсената непрекъснатост на първата производна $s'(x)$, необходимо е да бъде изпълнено условието двете едностранни производни на всички вътрешни граници на подинтервалите да са равни:

$$s'_-(x_i) = s'_+(x_i) \qquad (7)$$

(При това непрекъснатостта на $s''(x)$ се получава автоматично). Това условие води до уравнението:

$$\Delta_{i-1} + h_{i-1}(2\sigma_i + \sigma_{i-1}) = \Delta_i - h_i(\sigma_{i+1} + 2\sigma_i),$$
$$\textit{или}: \qquad (8)$$
$$h_{i-1}\sigma_{i-1} + 2(h_{i-1} + h_i)\sigma_i + h_i\sigma_{i+1} = \Delta_i - \Delta_{i-1}, \quad i = 2..n-1$$

Това е система от $n-2$ уравнения за n неизвестни $\{\sigma_i, i = 1,...,n\}$. Необходимо е да се формулират още две допълнителни условия, за да бъде определен интерполационният сплайн $s(x)$ по единствен начин. Тези условия в естествения кубичен сплайн са, както вече писахме:

$$s''(x_1) = s''(x_n) = 0, \quad \textit{или}: \quad \sigma_1 = \sigma_n = 0. \qquad (9)$$

Общият вид на системата линейни уравнения (8) в матричен вид изглежда така:

$$\begin{pmatrix} 2(h_1+h_2) & h_2 & 0 & 0 \\ h_2 & 2(h_2+h_3) & h_3 & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & h_{n-2} & 2(h_{n-2}+h_{n-1}) \end{pmatrix} \begin{pmatrix} \sigma_2 \\ \sigma_3 \\ \dots \\ \sigma_{n-1} \end{pmatrix} = \begin{pmatrix} \Delta_2 - \Delta_1 \\ \Delta_3 - \Delta_2 \\ \dots \\ \Delta_{n-1} - \Delta_{n-2} \end{pmatrix} \qquad (10)$$

Тази система от $n-2$ уравнения лесно може да се реши, макар че тя може да бъде от висок ред. Основание за това ни дава специалният вид на матрицата от коефициентите на системата. Тя има следните особени свойства:

- матрицата е лентова (тридиагонална);
- матрицата е симетрична;
- тази матрица е неособена и е положително дефинитна (= с доминиращи елементи по главния диагонал).

Такива системи линейни уравнения се решават лесно по метода на Холецки (Cholesky), чиято схема накратко привеждаме:

За всяка реална симетрична положително дефинитна матрица A съществува, и при това единствено, представяне:

$$A = LL^T, \qquad (11)$$

където L е лява (долна) триъгълна матрица, а с L^T е означена транспонираната й матрица. Тогава в системата: $Ax = b$ можем да означим:

$y = L^T x$ и да заменим нейното решаване с последователното решаване на двете системи:

$$A = Ly,\ y = L^T x\ . \qquad (12)$$

(първо намираме междинното неизвестно y, а след това и търсеното x). Предимството на тази замяна е обстоятелството, че системи линейни алгебрични уравнения с триъгълни матрици се решават много просто с последователно заместване (отгоре надолу за лява триъгълна матрица и отдолу нагоре за дясна триъгълна матрица).

Връзката между коефициентите $\{a,b,c,d\}$ и коефициентите $\{\sigma\}$ е:

$$\begin{aligned} a_i &= y_i \\ b_i &= \frac{y_{i+1}-y_i}{h_i} - h_i(\sigma_{i+1}+2\sigma_i) \\ c_i &= 3\sigma_i \\ d_i &= \frac{\sigma_{i+1}-\sigma_i}{h_i} \end{aligned} \qquad (13)$$

Представянето на коефициентите във вида (13) дава възможност по-лесно да се пресмятат производните и интегралите на сплайна.

3.2. Изглаждане на данните

При интерполацията ние предполагаме, че стойностите на данните в точките на измерване са точни и нашата задача е да получим възможност да пресмятаме тези данни при междинни стойности на аргумента. В действителност ние много рядко можем да считаме, че измерванията са с толкова висока точност, че да е разумно данните да минават тъкмо през експериментално измерените стойности. Обикновено измерванията са съпроводени със значителни грешки или неопределености. Те понякога са толкова големи (в една наглед шеговита книга се съдържа следният мъдър съвет: “много рядко има измервания с точност, по-добра от 1%, независимо от уверенията на авторите им”), че простата интерполация на такива данни води до решения, които са математически коректни, но нямат никакъв физичен смисъл.

В т акива случаи ние се нуждаем от по-сложни процедури, които да ни позволят да получим една *изгладена последователност* от числови данни, минаваща плавно покрай експерименталните наблюдения.

Защо е нужно такова изглаждане? Има няколко причини за това:

- Първо, то съответствува на интуитивната ни представа, че физичните закономерности са по природата си гладки и ако наблюдаемите зависимости не са такива, то причината е в случайните фактори, влияещи върху измерванията. Те водят до данни, съдържащи грешки от случаен характер, които трябва да бъдат “поправени” с подходяща процедура за изглаждане. Най-често ситуацията е именно такава в смисъл, че измерваните зависимости са наистина гладки. Трябва обаче да знаем, че има някои важни изключения, главно от областта на атомната и ядрената физика, където има множество наблюдаеми величини, чиито стойности се изменят дискретно (например атомните и ядрените спектри на излъчване и поглъщане);
- Второ, изгладените зависимости позволяват да се разграничат по-добре съществените особености на данните от несъществените детайли, които понякога (при големи грешки) дори могат да маскират основната закономерност, която представлява интерес за нас;
- Накрая, изгладените зависимости имат по-добри качества от гледна точка на числения анализ, в частност те позволяват да се получат по-достоверни резултати при числено диференциране, при търсене на минимуми и максимуми и др.

Като математически обекти методите за изглаждане представляват ***операторите,*** т.е. правила, по които на всяка функция (последователност от данни) се съпоставя друга функция (изгладена последователност). Физичната природа на наблюденията обикновено изисква тези оператори да са линейни, т.е.

$$P[\alpha f(x)+\beta g(x)]=\alpha P[f(x)]+\beta P[g(x)]. \qquad (1)$$

(такива са и самите наблюдения, например спектърът на смес от няколко вещества е суперпозиция от спектрите на отделните вещества в първо приближение). Ако означим наблюденията с $\{e_i, i=0,...,n\}$, а изгладената последователност с $\{s_i, i=0,...,n\}$, то най-простият вид на един оператор на изглаждане има вида:

$$s_i=\sum_{k=i-l}^{i+m} c_k e_k\,. \qquad (2)$$

Прости примери за изглаждащи оператори са изглаждането по три точки с еднакви тегла (т.нар. ***пълзящо средно***):

$$s_i = \frac{1}{3}(e_{i-1} + e_i + e_{i+1}) \qquad (3)$$

или пък с различни тегла, например:

$$s_i = \frac{1}{4}(e_{i-1} + 2e_i + e_{i+1}) \qquad (4)$$

Изглаждането, описвано с правилото (3), е по-силно от това с (4), тъй като в последното относителното влияние (тегло) на изглажданата точка i е по-голямо (затова и изглаждането (4) е по-слабо). Очевидно крайната ситуация (най-слабо изглаждане) в това отношение е:

$$s_i = e_i , \qquad (5)$$

което, разбира се, не е никакво изглаждане.

Описаните правила (3) и (4) са прости (и се използуват много често), но техният основен недостатък е, че те третират всички точки в изглажданата последователност по еднообразен начин, без да отчитат различната степен на гладкост на зависимостта в отделните й участъци.

Тук трябва да уточним, че когато делим опитно измерваните зависимости на дискретни и непрекъснати, някои от тези закономерности са по природата си дискретни, но те се наблюдават като непрекъснати благодарение на:

- фона от непрекъснати взаимодействия, които неизбежно присъствуват в измерванията и се наслагват върху полезния сигнал (който може да е дискретен);
- несъвършенствата на детекторите, които пораждат функция на отклика с непрекъснат характер въпреки дискретната природа на въздействието.

От тази гледна точка, основните типове данни се делят на такива, които са по природа непрекъснати и такива, които са квазинепрекъснати в горния смисъл. Истински непрекъснатите зависимости се описват обикновено с толкова гладки функции, че съответните експериментални данни могат да бъдат успешно изгладени с почти всякакви методи за изглаждане, дори и толкова прости като описаните по-горе. Затруднения при изглаждането се проявяват именно при зависимостите, които не са много гладки. Проблемът се състои в това, да се конструира такъв метод за изглаждане, който да осъществява филтриране на шума от статистическия характер на измерването и същевременно да запазва особеностите, присъщи на наблюдаваните закономерности. Обикновено използуваните методи за изглаждане се провалят

именно около участъците с малка гладкост, каквито са около пиковете, минимумите и другите особености на последователностите от данни.

Ние ще разгледаме един такъв общ метод за изглаждане, който е описан най-пълно от В.Б.Злоказов (1980), макар че в различни форми е срещан и по-рано. Той се състои в следното:

Въвежда се мярка за осцилацията на една функция. На визуално наблюдаваната осцилация съответствува бързо изменение на нормата на първата й производна, както и на нейния знак. Ако се сравнят два участъка от една крива, на които първата производна се променя еднакво, ние сме склонни да считаме за по-бързо променящ се този участък, който има по-малка дължина. Така че за мярка на осцилацията на функцията $y(x)$ в един безкрайно малък интервал се приема отношението на изменението на първата й производна към дължината на кривата в този интервал, което е точно известната ни вече кривина:

$$osc(y(x)) = \frac{y''(x)}{\sqrt{1+[y'(x)]^2}} \qquad (6)$$

Интегралната мярка за осцилация на функцията $y(x)$ в даден интервал $[x_1, x_2]$ се определя като:

$$\mu(x_1, x_2) = \int_{x_1}^{x_2} dx\, osc^2(x) = \int_{x_1}^{x_2} dx\, \frac{[y''(x)]^2}{1+[y'(x)]^2} \qquad (7)$$

Така определената интегрална величина μ е квадратична мярка, която съответствува на нагледната ни представа за осцилация.

Сега задачата за изглаждане се формулира така:

Нека имаме наблюденията $\{e_i, i = 0,...,n\}$, които предполагаме, че са еквидистантни за простота на записа. Ако с $\{s_i, i = 0,...,n\}$ означим изгладената последователност, то ние искаме тя да се конструира от условието да се намери минимумът на функционала:

$$Q = \varepsilon \int_0^n dx\, \frac{s''^2(x)}{1+s'^2(x)} + \int_0^n dx\, [s(x) - e(x)]^2 . \qquad (8)$$

Смисълът на така конструирания функционал Q е следният: той се състои от два члена, от които първият е (квадратична) мярка за осцилацията на изгладената функция, а вторият е (също квадратична) мярка за близостта на изгладената последователност до оригиналните наблюдения. И тъй, ние

искаме с конструирането на Q да постигнем две цели - хем изгладената последователност да е близо до експерименталните точки, хем тя да има минимална осцилация. Лесно е да се убедим, че тези изисквания са противоречиви. Големината на числото ε, което ние можем свободно да избираме, е мярка за компромиса между тези две условия, който ние правим. Ако искаме изгладената последователност $\{s\}$ да бъде по-гладка, ε трябва да се увеличава; ако пък искаме $\{s\}$ да е по-близо до $\{e\}$, ε трябва да намалява. Граничните случаи са, очевидно:

- $\varepsilon = 0$, когато се получава баналното решение $s \equiv e$ (т.е. функцията s ще минава точно през наблюденията e, но няма да има никакво изглаждане), или
- $\varepsilon \to \infty$, когато s'' трябва да клони към 0, за да има крайно решение. Това е еквивалентно на изискването функцията s да клони към права линия. Така ще получим най-гладката функция, която изобщо може да се построи, но която, разбира се, изобщо не е длъжна да минава близо до нашите наблюдения, респ. да служи за изглаждане на тези данни.

Между тези гранични случаи ние сме свободни в избора на изглаждащия параметър ε в зависимост от нашите предпочитания и вида на поставената задача.

За да се допълни така формулираната задача за изглаждане, трябва да се прибавят гранични условия за краищата:

$$s(0) = e(0), \quad s(n) = e(n), \quad s''(0) = e''(0), \quad s''(n) = e''(n) \tag{9}$$

Този избор на гранични условия всъщност изразява липсата на информация за гладкостта на търсената зависимост в крайните точки. Поради това ние предполагаме линейно поведение на тази функция в краищата на интервала и по посока извън него.

Сега минаваме към дискретния аналог на съотношението (8), като си даваме сметка, че ние познаваме изглажданите данни само в точките, където имаме измервания. Първо опростяваме израза за осцилацията, тъй като в него s влиза силно нелинейно (чрез знаменателя). Да си спомним, че при конструирането на изглаждащи филтри за анализиране на данните с неправдоподобни стойности ние приехме $1 + s'^2(x) \approx const$. Тогава ставаше въпрос за заключения от качествен характер и това приближение беше оправдано. Сега не можем да се задоволим с толкова грубо приближение;

затова ние заместваме израза в знаменателя с някаква апроксимация на дължината на дъгата, *която да не зависи от* s. Най-добре е като приближение на $\{s\}$ да използуваме самите наблюдения $\{e\}$ (тъй като ние искаме в крайна сметка $\{s\}$ да е близо до $\{e\}$). Така знаменателят става независим от $\{s\}$ и задачата за минимизация става линейна.

И така, уравнението за определяне на $\{s\}$ добива вида:

$$Q = \sum_{i=0}^{n} (s_i - e_i)^2 + \sum_{i=1}^{n-1} v_i (s_{i+1} - 2s_i + s_{i-1})^2 = \min_{\{s_i, i=0..n\}}, \qquad (10)$$

където:

$$v_i = \frac{\varepsilon}{1 + (e_i - e_{i-1})^2} = const(s). \qquad (11)$$

Тук трябва още веднъж да изтъкнем, че линейността на задачата се осигурява от замяната $(s_i - s_{i-1})$ с $(e_i - e_{i-1})$. Тя се оправдава от обстоятелството, че при неголям статистически шум в данните e_i е сравнително добра апроксимация на s_i и като се има предвид, че тази замяна се прави само в теглото на единия член на функционала Q, тази апроксимация в много случаи е приемлива. Това приближение е обаче само за наше удобство. Ако това заместване не води до добри резултати, възможно е задачата да се реши по-точно, като се организира итерационен процес, в нулевата итерация на който се използува приближението (11), а в следващите – стойността на изгладената последователност от предишната итерация:

$$v_i^{(k)} = \frac{\varepsilon}{1 + (s_i^{(k-1)} - s_{i-1}^{(k-1)})^2}, \quad k = 1,... \ . \qquad (12)$$

Практически това рядко се налага, като се има предвид, че все пак става дума за първична обработка на данните.

За да намерим минимума на функционала Q, трябва да приравним на нула всички негови първи частни производни спрямо параметрите $\{s\}$, които се явяват в тази задача неизвестни величини, подлежащи на определяне.

Ето тези производни:

$$\begin{aligned}
\frac{\partial Q}{\partial s_0} &= 2(s_0 - e_0) + 2v_1(s_2 - 2s_1 + s_0) = 0 \\
\frac{\partial Q}{\partial s_1} &= 2(s_1 - e_1) + 2v_1(s_2 - 2s_1 + s_0)(-2) + 2v_2(s_3 - 2s_2 + s_1) = 0 \\
\frac{\partial Q}{\partial s_2} &= 2(s_2 - e_2) + 2v_1(s_2 - 2s_1 + s_0) \\
&\quad + 2v_2(s_3 - 2s_2 + s_1)(-2) + 2v_3(s_4 - 2s_3 + s_2) = 0
\end{aligned} \tag{13}$$

Общият член е:

$$\begin{aligned}
\frac{\partial Q}{\partial s_k} &= 2(s_k - e_k) + 2v_{k-1}(s_k - 2s_{k-1} + s_{k-2}) \\
&\quad + 2v_k(s_{k+1} - 2s_k + s_{k-1})(-2) + 2v_{k+1}(s_{k+2} - 2s_{k+1} + s_k) = 0
\end{aligned} \tag{14}$$

Като прегрупираме членовете пред неизвестните $\{s_k\}$, получаваме:

$$\begin{pmatrix}
1+v_1 & -2v_1 & v_1 & 0 & \dots & 0 & 0 \\
-2v_1 & 1+4v_1+v_2 & -2(v_1+v_2) & v_2 & 0 & \dots & 0 \\
v_1 & -2(v_1+v_2) & 1+v_1+4v_2+v_3 & -2(v_2+v_3) & v_3 & 0 & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & 0 & \dots & v_{n-2} & -2(v_{n-2}+v_{n-1}) & 1+4v_{n-1}+v_{n-2} & -2v_{n-1} \\
0 & \dots & \dots & 0 & v_{n-1} & -2v_{n-1} & 1+v_{n-1}
\end{pmatrix}
\begin{pmatrix} s_0 \\ s_1 \\ s_2 \\ \dots \\ \dots \\ s_{n-1} \\ s_n \end{pmatrix} =
\begin{pmatrix} e_0 \\ e_1 \\ e_2 \\ \dots \\ \dots \\ e_{n-1} \\ e_n \end{pmatrix} \tag{15}$$

Това е линейна система алгебрични уравнения и техният брой е твърде голям (колкото е броят на измерванията, често от порядъка на хиляди). Тази система обаче притежава отново особено благоприятните свойства на системата от предишния раздел, а именно:

- матрицата на системата (15) е реална и симетрична;
- тя е с доминиращ главен диагонал (=положително дефинитна).
- тази матрица в допълнение е лентова матрица с ненулеви елементи само по главния диагонал и съседните му два кодиагонала.

Това позволява за решаването на тази система отново да се използува разложението на Холецки, което е много рационално по отношение на числените пресмятания.

Да добавим, че стойността на параметъра ε по принцип е свързана с точността на измерванията $\{e\}$. В ситуацията на първична обработка на данни, когато ние практически нямаме информация за статистическите свойства на

данните, стойността на този регулиращ параметър обикновено се избира въз основа на тестове.

Трябва да споменем още няколко неща във връзка с този метод:

- при него ние получаваме стойностите на изгладената последователност $\{s\}$ само в точките на измерванията $\{e\}$. Ако са ни необходими интерполации за междинни стойности, ние можем да използуваме например сплайнова интерполация, построена върху стойностите на изгладената последователност $\{s\}$;
- Решаването на задачата за изглаждане на данните по описания по-горе метод е непосредствено свързано и много сходно с подхода, използуван при конструирането на сплайнова интерполация. Ние можем да търсим директно решение на задачата за изглаждане чрез сплайни; такива сплайни се наричат ***апроксимационни сплайни.*** Решаването на такава задача обаче е по-сложно и излиза вън от рамките на този курс.

Литература:

[1]. Дж. Форсайт, М.Малкълм, К.Молър: Компютърни методи за математически пресмятания, Наука и изкуство, София (1986).

[2]. В.Б.Злоказов, Метод для автоматического поиска пиков в гамма-спектрах, Препринт Р10-81-204, ОИЯИ-Дубна (1981).

4. Определяне на фона в последователности от данни

Следващата операция при първичната обработка на данните е определянето на фона. Да си спомним, че фон (подложка, англ. background) се нарича тази компонента на данните, която се дължи на процеси, които са, разбира се, също физични, но те са странични за предприетото изследване и не ни интересуват в момента. Това са основно процеси на взаимодействие на разсеяни лъчения, електрически и други шумове и пр. Фонът се счита за безполезна (и всъщност вредна) компонента, тъй като неговите статистически флуктуации се наслагват с тези на полезния сигнал и затрудняват определянето на интересуващата ни физична информация. Възможно е дори (при силен фон и съответно слаб сигнал) полезният сигнал да се "скрие" в шума.

Поради тези причини фонът трябва да се определи колкото е възможно по-добре и да се отдели от полезната компонента на данните.

Като правило фонът се мени плавно от канал в канал (във всеки случай по-плавно от изучавания сигнал). От теоретична гледна точка най-добре е фонът да бъде определен в отделно измерване, което се състои в натрупването на данните при същите условия (апаратура, геометрия на експеримента и др.), но в отсъствие на източника на полезния сигнал. Тогава "чистият" сигнал се получава просто чрез поточково изваждане на фоновия спектър от спектъра на основното измерване.

Това въобще е единственият възможен начин да се определи фона в случай, че полезният сигнал има скорост на изменение от канал към канал, сравнима с тази на фона. Такъв подход обаче не винаги е възможен, понеже част от фона може да е породена от вторични процеси в детектора, които се дължат на самия източник на сигнала, който се изучава. В такъв случай трябва да се използуват методи за оценка на фона непосредствено от измервания спектър. В общия случай не е възможно да се изхожда от някакъв аналитичен израз за разпределението на фона, който да е съгласуван с физическото му формиране поради извънредно сложния характер на процесите, водещи до възникването на фон (за да си дадем представа за тяхната сложност, достатъчно е да помислим, че това са обикновено вторични или многократни взаимодействия). При това положение информацията, на която може да се разчита, че е известна, е:

- че скоростта на изменение на фона е по-малка от тази на полезния сигнал, и
- че фонът е по-малък от сумарния сигнал (ние ще говорим за емисионни спектри, при които е именно така; при абсорбционните спектри, напротив, фонът мажорира полезния сигнал). С други думи, считаме, че фонът представлява обвивка отдолу на полезната компонента на сигнала.

На тези свойства се основава и алгоритъмът, който ще разгледаме тук.

Този алгоритъм се основава на идея, подобна на тази на метода за изглаждане на данните на Злоказов. Подобно на мярката за осцилацията на една функция се въвежда мярка за нейната ***изменчивост,*** която е най-естествено да бъде свързана с първата ѝ производна:

$$dev(y(x)) = y'(x). \quad (1)$$

Интегралната (квадратична) мярка за изменчивост на функцията в краен интервал е:

$$\eta(y(x))_{[x_1,x_2]} = \int_{x_1}^{x_2} dx\ dev^2(y(x)) \quad (2)$$

И тъй, ние искаме да построим апроксимация на фона - гладка обвивка на наблюденията $\{e\}$ отдолу, като решение на една вариационна задача: да се намери фонът $b(x)$ от минимума на функционала:

$$Q = \varepsilon \int_0^n dx\,[b'(x)]^2 + \int_0^n dx\ w(x)[b(x) - e(x)]^2 \quad (3)$$

при граничните условия:

$$b(0) = e(0); b(n) = e(n). \quad (4)$$

Тук се въвежда специална тегловна функция $w(x)$, която има за задача да осигури близост на фона $\{b\}$ към данните $\{e\}$ за малките стойности на данните и да отслабва това ограничение за големите стойности на $\{e\}$. Смисълът на тази тегловна функция е ясен - при голям сигнал фонът е далеч отдолу и не е необходимо да се поставя (силно) ограничение за близост на сигнала към фона, а при малък сигнал фонът очевидно трябва да е близо под него, тъй като тогава целият сигнал се състои почти само от фон.

Подходящ избор на тегловната функция е:

$$w(x) = \frac{1}{e(x) + \alpha}. \quad (5)$$

Малкото число α в (5) служи за ограничаване на теглото до крайни стойности при нулев сигнал.

Дискретизираният вариант на функционала Q има вида:

$$Q = \varepsilon \sum_{i=0}^{n-1} (b_{i+1} - b_i)^2 + \sum_{i=0}^{n} w_i (b_i - e_i)^2. \quad (6)$$

Тук конкретният вид на $w(x)$ не е написан, важно е само w да не зависи от фона $\{b\}$.

За да намерим минимума на Q спрямо $\{b_i, i = 0,...,n\}$, които тук са неизвестните величини, подлежащи на определяне, трябва отново да намерим първите частни производни на Q спрямо тези параметри и да ги приравним на нула.

Ето как изглеждат те:

$$\frac{\partial Q}{\partial b_1} = 2w_1(b_1 - e_1) + 2\varepsilon(b_2 - b_1)(-1) + 2\varepsilon(b_1 - b_0) = 0$$
$$\frac{\partial Q}{\partial b_2} = 2w_2(b_2 - e_2) + 2\varepsilon(b_3 - b_2)(-1) + 2\varepsilon(b_2 - b_1) = 0 \qquad (7)$$

...

В краищата производните се определят като се имат предвид и граничните условия (4) (които водят всъщност до анулиране на първите членове в следващите изрази):

$$\frac{\partial Q}{\partial b_0} = 2w_0(b_0 - e_0) + 2\varepsilon(b_1 - b_0)(-1) = 0$$
$$\frac{\partial Q}{\partial b_n} = 2w_n(b_n - e_n) + 2\varepsilon(b_n - b_{n-1})(-1) = 0 \qquad (8)$$

Системата линейни уравнения за $\{b\}$ в този случай изглежда така:

$$\begin{pmatrix} \varepsilon & -\varepsilon & 0 & 0 & ... & 0 \\ -\varepsilon & 2\varepsilon + w_1 & -\varepsilon & 0 & ... & 0 \\ 0 & -\varepsilon & 2\varepsilon + w_2 & -\varepsilon & 0 & 0 \\ ... & ... & ... & ... & ... & ... \\ 0 & ... & 0 & -\varepsilon & 2\varepsilon + w_{n-1} & -\varepsilon \\ 0 & 0 & 0 & ... & -\varepsilon & \varepsilon \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ ... \\ b_{n-1} \\ b_n \end{pmatrix} = \begin{pmatrix} 0 \\ w_1 e_1 \\ w_2 e_2 \\ \\ w_{n-1} e_{n-1} \\ 0 \end{pmatrix} \qquad (9)$$

Матрицата на тази система е също реална, симетрична, положително дефинитна и лентова, както и онази при метода за изглаждане. Но тя има сега само един ненулев кодиагонал, което опростява още повече решаването на уравненията.

Конкретният вид на тегловната функция $w(x)$ и стойността на параметъра на изглаждане ε се избират обикновено въз основа на тестове.

5. Търсене на пикове в спектрите

Тук ще говорим условно за “спектри”, т.е. за последователности от данни с еквидистантни стойности на аргумента.

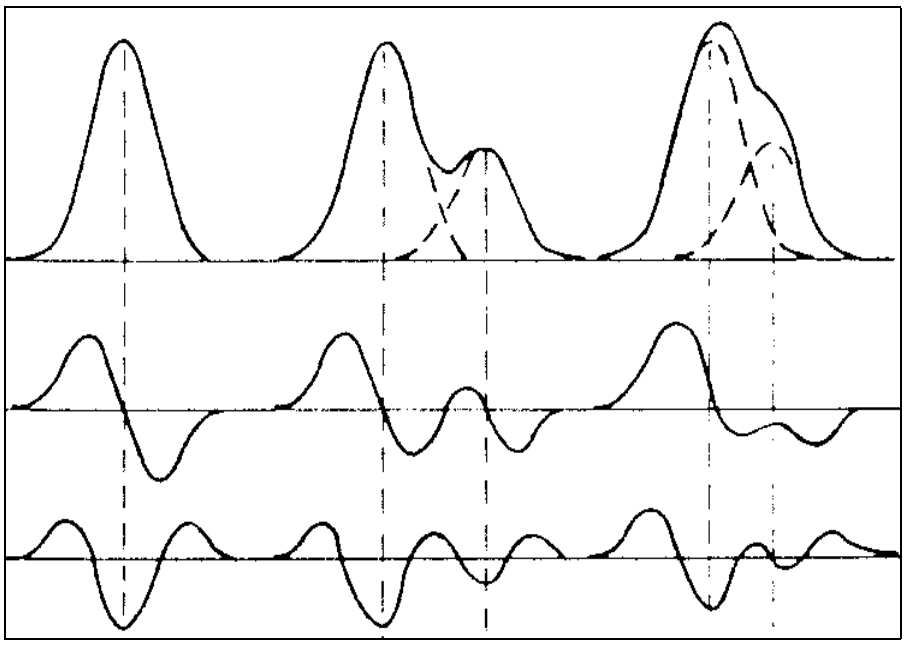
Както вече отбелязахме, пиковете са области от спектрите, които съдържат съществената информация, заради която спектрите са измерени. Поради това откриването на пикове в един спектър е една важна част от предварителната му обработка. Тук ще се спрем накратко на методите за автоматично търсене на пикове.

Основните методи, по които се търсят пикове в данните, са два:

- чрез използуване на екстремалните свойства на пиковете (известни от математическия анализ). При това се построяват апроксимации за първата (а понякога и за втората) производна на наблюденията (след като те са вече изгладени и фонът е изваден);
- чрез използуване на корелациите между последователните стойности на спектралните данни в областта на пиковете.

Методът на производните се основава на обстоятелството, че в областта на пика първата производна на данните спрямо аргумента се променя съществено, като преди пика тя е слабо положителна, после значително се увеличава, след което намалява, минава през нулата и след това симетрично се изменя, минавайки през минимум и асимптотично отива пак към нула, оставайки отрицателна. Пикът се намира при междинното й нулиране. Втората производна пък (тъй като е мярка за кривината) в областта на пика добива съществено отрицателни стойности, които са заобиколени от двете страни симетрично с положителни стойности.

Тези функции имат твърде характерно поведение, което служи за идентификация на пика. Основна пречка за използуването на този метод е статистическият шум в данните, който ограничава възможността за откриване на слаби пикове, тъй като влияе върху стабилното определяне на първата и особено на втората производна, въпреки междинното числено изглаждане на данните, което обикновено се предприема. Картината се усложнява значително в случай, че има два или повече пикове в близост (мултиплетни структури), особено ако те не могат да бъдат разделени от регистриращата апаратура (вж. фигурата). Комбинирането на критериите на първата и втората производна помага за по-сигурното определяне на наличието или отсъствието на пик в спектъра.



единичен пик *частично разделен дублет* *неразделен дублет*

Поведение на y(x), y'(x) и y"(x) около максимум

Корелационният метод за търсене на пикове се основава индиректно на локалната кривина на спектъра и изменението й в областта на пика. За да се пресметне промяната в наклона, необходимо е да се филтрира високочестотният статистически шум, което става обикновено чрез усредняване по няколко канала в рамките на ширината на линията. За целта се използува трансформация от вида:

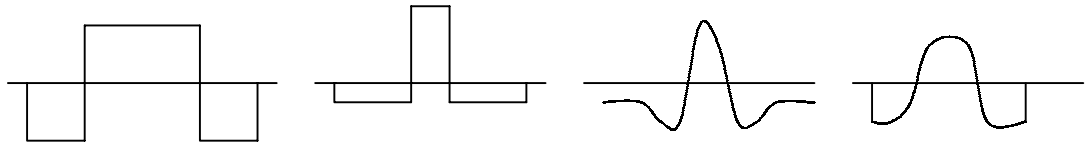
$$T(j) = \sum_{i=j-m}^{j+m} c(i-j)e(i),\ j = m,...,n-m \qquad (1)$$

където c е подходящо избрана функция-филтър. Към нея има две изисквания:

- тя трябва да има нулева площ;
- и да бъде симетрична относно центъра си.

При наличие на пик $T(j)$ приема голяма положителна стойност (пропорционална на площта на пика). В отсъствие на пикове T флуктуира около нулата.

Най-често използуваните корелационни филтри са:



квадратична вълна | *правоъгълна вълна* | *обърната втора производна на Гаусова функция* | *Гаусов корелатор*

Общо взето, и четирите функции дават сходни резултати при търсенето на пикове. Последните две имат известни предимства по отношение на точност, но с първите две пресмятанията се извършват по-бързо.

Основно предимство на корелационния метод пред метода на производните е, че тук за идентифициране на пика се използува информацията едновременно от много точки около него. Това придава на този метод стабилност, присъща на всички изглаждащи методи. Известен недостатък на корелационния метод обаче е необходимостта от предварителна оценка на ширината на търсените пикове.

При излагането на тези методи ние съзнателно оставяме открит въпроса за пиковите *мултиплети*, т.е. за недобре разграничените един от друг близки пикове. Такива конфигурации внасят допълнително усложняване на картината и затрудняват откриването на пикове. В известна степен обаче изложените тук методи могат с някои модификации да бъдат прилагани и в такива случаи.

Накрая, след като един пик е идентифициран като възможен според горните критерии, е необходимо да се провери неговата *статистическа значимост.* Това става най-общо чрез сравняване на амплитудата (или по-добре – на площта) на пика с флуктуациите на фона в тази област. Решаването на тази задача обаче (както и при останалите методи за обработка на данни) не е възможно без използуването на информация за статистическия характер на обработваните данни.

Литература:

[1]. G.W.Philips, K.W.Marlow, Nucl.Instr.Meth. 137 (1976) 525-536.

III. Статистически свойства на наблюденията

6. Случайни величини

6.1. Понятие за вероятност

Експериментът се дефинира като строга последователност от предварително определени действия, водещи до определяне на стойностите на една (или повече) величини, които се наричат *резултат(и) от експеримента*. Те могат да се изменят дискретно или непрекъснато.

Независимо от точността на изпълнение на условията за провеждане на експеримента обаче, резултатите от повторния експеримент са, изобщо казано, различни. Това се дължи или на ограничената точност на измерванията, или на вътрешната природа на изследваното явление.

Поради това, възможните стойности на всяка величина, която е резултат от един експеримент, не са някакви фиксирани числа, а някакво (най-често ограничено) множество. Съвкупността от тези области за всички величини, представляващи резултати от даден експеримент, се нарича *пространство на извадките (основно пространство)* на този експеримент. Обикновено е трудно да се определят точните граници на допустимите области за всички величини, които са измеряеми в даден експеримент. Затова обикновено пространството на извадките се взема по-голямо и то съдържа фактическото пространство като подмножество.

Елементарно събитие се нарича всеки елемент от пространството на извадките, т.е. всяка реализация на експеримента, при която се получават определени конкретни резултати за всички величини в този експеримент.

Събитие се нарича пък фактът на принадлежност на резултата към някакво подмножество на пространството на извадките. Ако едно събитие A не е настъпило, ние говорим за настъпването на допълнителното събитие $\overline{A}$. Цялото пространство на извадките пък съответствува на събитие, което се случва винаги в резултат на този експеримент (*сигурно* или *достоверно събитие* - E).

Понятието *вероятност на дадено събитие* има различни определения:

- *честотно* - като граница на отношението на броя на благоприятните изходи за събитието към общия брой реализации на експеримента: $P(A) = \lim_{N\to\infty} \frac{n}{N}$. Трудностите с употребата на това определение са

свързани с невъзможността за провеждането на безкраен брой експерименти;

- вероятността може да се въведе и ***аксиоматично*** (Колмогоров) чрез един минимален брой аксиоми. Тук ние ще ги скицираме:
    1. (неотрицателност): На всяко събитие A съответствува неотрицателно число, което се нарича негова ***вероятност***: $P(A) \geq 0$;
    2. (нормировка): Достоверното събитие E има вероятност единица: $P(E) = 1$;
    3. (адитивност): Ако събитията A и B са ***несъвместни*** (не могат да се случат едновременно), то вероятността за настъпване на A или B е: $P(A+B) = P(A) + P(B)$;
    4. (непрекъснатост): За всяка редица от събития, клоняща към празното множество, редицата от техните вероятности е сходяща с граница 0: $A_n \to \varnothing \Rightarrow \lim_{n\to\infty} P(A_n) = 0$.

От тези аксиоми лесно се получава като следствие например, че $P(A) + P(\overline{A}) = 1$.

Условната вероятност да се случи събитието B при условие, че вече е настъпило събитието A, се определя така:

$$P(B/A) = \frac{P(AB)}{P(A)}$$

Оттук:

$$P(AB) = P(A)P(B/A)$$

В тези изрази AB е сечението на събитието A със събитието B.

Оттук следва формулата за ***пълната вероятност*** за дадено събитие:

Нека даден експеримент води към някое от следните несъвместни събития:

$$E = A_1 + A_2 + \ldots + A_n$$

Тогава вероятността да се осъществи кое да е събитие B е:

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B/A_i).$$

Верността на това равенство следва от предишните разглеждания.

6.2. Случайни величини. Разпределения на случайните величини

Всяко събитие може да се характеризира чрез получаването в резултат на настъпването му на една от възможните стойности на някаква величина. Тъй като при всяка реализация на един експеримент тази величина приема различни стойности (които са неизвестни отнапред), тя се нарича *случайна*. Случайните величини се делят на *дискретни* или *непрекъснати* според множествата от стойности, които могат да заемат.

Основната характеристика на всяка случайна величина е нейната *функция на разпределение*. Тя се определя така:

ако $\bar{x}$ е случайна величина и $x \in (-\infty, \infty)$ е реално число, то вероятността за събитието ($\bar{x} < x$) е функция на x и се нарича *функция на вероятностното разпределение на случайната величина* $\bar{x}$:

$$F(x) = P(\bar{x} < x). \qquad (1)$$

Ако $\bar{x}$ е дискретна случайна величина, $F(x)$ е стъпаловидна функция. Очевидно $F(x)$ е монотонно ненамаляваща функция.

От определението следва:

$$P(\bar{x} \geq x) = 1 - P(\bar{x} < x) = 1 - F(x) \qquad (2)$$

Специален интерес представлява не само (и не толкова) функцията $F(x)$, а нейната производна по аргумента й (разбира се, ако такава съществува):

$$f(x) = \frac{dF(x)}{dx}, \qquad (3)$$

която се нарича *плътност на вероятностното разпределение (вероятностна плътност)* на случайната величина $\bar{x}$**.** Тя е пропорционална на вероятността за събитието $(x \leq \bar{x} \leq x + dx))$- т.е. след експеримента случайната величина $\bar{x}$ да се окаже в един безкрайно малък интервал около числото x. Очевидно:

$$P(\bar{x} < a) = \int_{-\infty}^{a} dx\ f(x) = F(a) \qquad (4)$$

$$P(a \leq \bar{x} \leq b) = \int_{a}^{b} dx\ f(x) = F(b) - F(a) \qquad (5)$$

В частност:

$$\int_{-\infty}^{\infty} dx\ f(x) = 1 = P(E) \qquad (6)$$

6.3. Функции на случайни величини. Математическо очакване. Дисперсия. Моменти

Често ние се интересуваме освен от дадена случайна величина още и от някакви други величини, които са нейни функции. Всяка функция на една случайна величина е също случайна величина, доколкото също заема различни стойности при различни реализации на един и същ експеримент. Следователно тя има също вероятностна функция на разпределение, както и вероятностна плътност.

Средна стойност (математическо очакване) на една дискретна случайна величина $\bar{x}$ се нарича сумата от всевъзможните стойности на тази величина, умножени по съответните им вероятности:

$$E(\bar{x}) \stackrel{def}{=} \sum_{i=1}^{n} x_i P(\bar{x} = x_i) \qquad (1)$$

Средната стойност на случайната величина обаче *не е* случайна величина. Тя всъщност не зависи от $\bar{x}$, тъй като в (1) се сумира по *всички възможни* стойности на случайната величина. Средната стойност не е функция, а функционал на $\bar{x}$, т.е. правило, по което на една функция – $P(\bar{x})$, се съпоставя едно число – математическото очакване.

Средна стойност на функцията $H(\bar{x})$ на случайната величина $\bar{x}$ се определя така:

$$E\{H(\bar{x})\} \stackrel{def}{=} \sum_{i=1}^{n} H(x_i) P(\bar{x} = x_i) \qquad (2)$$

При непрекъснати случайни величини тези формули се обобщават непосредствено:

$$E(\bar{x}) = \int_{-\infty}^{\infty} dx\, x f(x) \qquad (3)$$

$$E\{H(\bar{x})\} = \int_{-\infty}^{\infty} dx\, H(x) f(x) \qquad (4)$$

Сега ще разгледаме една функция от специален вид:

$$H(\bar{x}) = (\bar{x} - c)^k \qquad (5)$$

Нейното математическо очакване:

$$\alpha_k = E\{(\bar{x} - c)^k\} \qquad (6)$$

се нарича *момент от ранг к* на случайната величина $\bar{x}$ спрямо точката c.

Особен интерес представляват моментите спрямо средната стойност на случайната величина:

$$\mu_k = E\{[\bar{x} - E(\bar{x})]^k\} = \int_{-\infty}^{\infty} dx\,[x - E(\bar{x})]^k f(x) \qquad (7)$$

Те се наричат ***централни моменти*** на случайната величина $\bar{x}$. Причината за специалното предпочитание към централните моменти се основава на обстоятелството, че функцията $\alpha_k(c)$ има минимум при $c = E(\bar{x})$ поне при $k = 2$.

Очевидно, $\mu_0 = 1, \mu_1 = 0.\ \mu_2$ е низшият по порядък централен момент, който съдържа информация за отклонението на случайната величина от средната й стойност. Централният момент от втори порядък μ_2 се нарича ***дисперсия*** (variance) на случайната величина $\bar{x}$. Ето няколко синонимни означения, използувани в различни книги:

$$\mu_2(\bar{x}) = \sigma^2(\bar{x}) = \mathrm{var}(\bar{x}) = E\{[\bar{x} - E(\bar{x})]^2\}$$

Моментите на случайните величини имат и физична интерпретация:

При множество измервания на една и съща случайна величина резултатите се групират предимно около една "истинска" стойност на тази величина. Или, казано по-точно, по-вероятно е да се получат (и по-често се получават) стойности, близо до някакво число (напр. x_0), отколкото далеч от него. Вероятностната плътност на случайната величина тогава ще има типичната и често срещана камбановидна форма. Математическото очакване може да се интерпретира като оценка за действителната стойност на измеряемата величина. Изразът за него прилича на този за координатите на центъра на масите на едно (едномерно) тяло. Дисперсията пък в такъв случай може да се интерпретира като израз за инерчния момент на такова тяло (относно центъра на масите му), и тя е мярка за *разсейването* на резултатите от измерванията около средната им стойност. Ако дисперсията е малка, измерванията могат да се считат за по-точни (по-голяма част от тях са близо до средната стойност).

Дисперсията е квадратична мярка. Затова се използува величината $\sigma(\bar{x}) = +\sqrt{\mu_2}(\bar{x})$, която се нарича ***средноквадратично (стандартно) отклонение*** на случайната величина $\bar{x}$. Обикновено именно тази величина се посочва като мярка за ***грешката*** (правилният термин е ***статистическата неопределеност***) на $\bar{x}$ и се пише: $x = E(\bar{x}) \pm \sigma(\bar{x})$.

В много случаи математическото очакване и дисперсията не дават достатъчна информация за изследваните случайни величини; затова за по-пълното им характеризиране се използуват и моменти от по-висок ред.

Третият централен момент μ_3 характеризира несиметричността на вероятностното разпределение около математическото очакване. Прието е безразмерната величина

$$\gamma_1 = \frac{\mu_3}{\sigma^3} \qquad (8)$$

да се използува за количествена характеристика на различията между относителния дял на положителните и отрицателните отклонения от средната стойност. Тази величина се нарича ***асиметрия*** (skewness) на разпределението на случайната величина. За симетрични разпределения (които се представят с четни функции на вероятностната плътност) $\gamma_1 = 0$.

Четвъртият централен момент $\mu_4 = E(x - E(\bar{x}))^4$ определя характера на максимума в точката на математическото очакване при симетрични разпределения. За количествена характеристика се приема безразмерната величина

$$\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3 \qquad (9)$$

която се нарича ***ексцес*** (kurtosis).

Величините асиметрия и ексцес са въведени чрез равенствата (8) и (9) така, че за най-често срещаното разпределение - нормалното - те да са нули: $\gamma_1 = 0$, $\gamma_2 = 0$. При $\gamma_1 > 0$ по-голямата част от вероятностното разпределение се намира вдясно от математическото очакване (респ. обратното). При $\gamma_2 > 0$ пък максимумът на вероятностната плътност е по-остър от този при нормалното разпределение с параметри като в равенството (9), съответно при $\gamma_2 < 0$ максимумът на разпределението е по-тъп от този на нормалното разпределение.

По-висшите моменти на разпределенията рядко се използуват. Трябва да се отбележи, че колкото по-голям е рангът k на даден момент, толкова по-голям е приносът на отдалечените от средната стойност участъци от функцията на вероятностната плътност (тъй наречените "опашки" на разпределението) при определянето на к-ия момент.

Освен важността на отделните моменти, съществено теоретично значение има и пълният набор от *всички* моменти (централни или нецентрални) на даденото разпределение. ***Нецентралните моменти*** се дефинират така:

$$\mu'_k = \int_{-\infty}^{\infty} dx\, x^k f(x) \qquad (10)$$

Доказва се (това ще видим малко по-нататък), че при много общи условия (които са изпълнени за всички разпределения, представляващи практически интерес) съвкупността от всички моменти напълно определя вероятностното разпределение. Има случаи, когато е по-лесно да се построят моментите на едно неизвестно разпределение, отколкото самото разпределение (т.е. вероятностната му плътност). Това позволява тогава да се построи и самото разпределение чрез неговите моменти.

В сила е по-конкретно следното твърдение (теорема):

Ако $f_1(x)$ и $f_2(x)$ са две вероятностни плътности, всичките съответни нецентрални моменти на които са равни, и ако разликата $f_1(x) - f_2(x)$ е разложима в степенен ред, то $f_1(x) \equiv f_2(x)$.

Ние ще установим верността на това твърдение. Нека означим:

$f_1(x) - f_2(x) = c_0 + c_1 x + c_2 x^2 + \ldots$ и нека моментите на двете функции са съответно $\mu'_1 = \mu''_1, \mu'_2 = \mu''_2, \ldots$

Тогава:

$$\int_{-\infty}^{\infty} dx\,[f_1(x) - f_2(x)]^2 = \int_{-\infty}^{\infty} dx\,[f_1(x) - f_2(x)](c_0 + c_1 x + c_2 x^2 + \ldots)$$

$$= c_0 \int_{-\infty}^{\infty} dx\,[f_1(x) - f_2(x)] + c_1 \int_{-\infty}^{\infty} dx\, x[f_1(x) - f_2(x)] + c_2 \int_{-\infty}^{\infty} dx\, x^2[f_1(x) - f_2(x)] + \ldots$$

$$= c_0(1-1) + c_1(\mu'_1 - \mu''_1) + c_2(\mu'_2 - \mu''_2) + \ldots = 0,$$

откъдето следва, че $f_1(x) \equiv f_2(x)$ (тъй като подинтегралната функция е неотрицателна и тогава горният интеграл е 0 тогава и само тогава, когато самата подинтегрална функция е тъждествено равна на нула).

При зададена вероятностна плътност $f(x)$ пресмятането на моментите $\{\mu(\bar{x})\}$ не представлява проблем. Проблем представлява обаче често самото намиране на вероятностната плътност, която да описва адекватно процеса на измерването на изучаваната случайна величина.

6.4. Свойства на математическото очакване и дисперсията

1. Ако $c = const$, то:

$$E(c) = c$$

$$E(c\bar{x}) = cE(\bar{x}) \qquad (1)$$

$$\sigma^2(c) = 0$$

$$\sigma^2(c\bar{x}) = c^2\sigma^2(x) \qquad (2)$$

От последното равенство следва:

$$\sigma^2(\bar{x}) = E\{[\bar{x} - E(\bar{x})]^2\} = E\{\bar{x}^2 - 2\bar{x}E(\bar{x}) + E^2(\bar{x})\} = E(\bar{x}^2) - E^2(\bar{x}) \qquad (3)$$

2. Разглеждаме функцията

$$\bar{u} = \frac{\bar{x} - E(\bar{x})}{\sigma(\bar{x})} \qquad (4)$$

Тази функция е също случайна величина. Нейното математическо очакване е:

$$E(\bar{u}) = \frac{1}{\sigma(\bar{x})}\{E[\bar{x} - E(\bar{x})]\} = \frac{1}{\sigma(\bar{x})}[E(\bar{x}) - E(\bar{x})] = 0 \qquad (5)$$

Дисперсията на случайната величина (4) е:

$$\sigma^2(\bar{u}) = \frac{1}{\sigma^2(\bar{x})}E\{[\bar{x} - E(\bar{x}) - 0]^2\} = \frac{\sigma^2(\bar{x})}{\sigma^2(\bar{x})} = 1 \qquad (6)$$

Така определената функция $\bar{u}$ се нарича ***приведена (стандартизирана, нормирана) случайна величина***. Такива величини са удобни с това, че математическото им очакване е нула, а дисперсията - единица. Таблиците и процедурите за компютърни пресмятания на статистическите разпределения като правило се дават именно за нормираните случайни величини. Самите случайни величини (ненормираните) се получават от нормираните чрез прости линейни трансформации.

Ето някои други мерки за характеристика на местоположението в статистическите разпределения:

мод, мода (mode) се нарича стойността x_m на аргумента, която съответствува на максималната вероятност: $P(\bar{x} = x_m) = \max$. Тя може да се определи чрез диференциране на вероятностната плътност по аргумента. Разпределенията с един максимум се наричат ***унимодални***. Такива са почти всички разпределения, които представляват интерес във физиката;

медиана (median) се нарича стойността на аргумента $x(0.5)$, която съответствува на $F(x(0.5) = P(\bar{x} < x(0.5)) = 0.5$. Медианата дели дефиниционната област на вероятностната плътност на две части, във всяка от които случайната величина попада с равна вероятност (т.е. това са две равноплощни

части на разпределението). За симетричните разпределения математическото очакване, модата и медианата съвпадат;

по-общо ***квантил (quantile)*** на ниво p, $0 < p < 1$ се нарича стойността на аргумента $x(p)$, за която: $F[x(p)] = P(\bar{x} < p) = p$. Така че медианата е квантилът на ниво 1/2. Квантилите на нивата 1/4, 2/4, 3/4 се наричат ***квартили*** , на нивата 0.1, 0.2,...0.9 се наричат ***децили***, а на нивата 0.01, 0.02,..., 0.99 - ***процентили***. За често използваните разпределения най-важните квантили са дадени в таблици и има готови изчислителни процедури за тяхното пресмятане.

7. Разпределения на няколко случайни величини

7.1. Разпределения на две случайни величини

Рядко при обработката на данни се оказва възможно да се ограничим с разглеждането само на една случайна величина. Дори и да се опитваме да измерваме само една физична величина (параметър), ние се натъкваме на наличието и на други фактори, които влияят върху резултатите от измерването и които също се описват със случайни величини (типичен пример са измерванията на някакъв сигнал при наличието на фон). Адекватното описание на такива експерименти предполага едновременното изучаване на статистическите свойства на всички величини, които реално участвуват във формирането на резултатите. Затова е нужно да се изучават *съвместните им вероятностни разпределения*. При това възникват и някои нови свойства, които не съществуват при едномерните разпределения.

Най-простият вариант е на две случайни величини, $\bar{x}, \bar{y}$. Ще ни интересува вероятността едновременно $\bar{x} < x, \bar{y} < y.$ Както и при една случайна величина, предполагаме, че съществува функция на разпределение:

$$F(x,y) = P(\bar{x} < x, \bar{y} < y) \qquad (1)$$

Във всички интересни случаи $F(x,y)$ съществува. Ако тя е диференцируема по двата си аргумента, то функцията:

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y} \qquad (2)$$

се нарича ***съвместна вероятностна плътност*** на случайните величини $\bar{x}, \bar{y}$. Тогава, очевидно:

$$P(a \le x < b, c \le y < d) = \int_a^b dx \int_c^d dy\ f(x,y) \qquad (3)$$

Понякога възниква следният проблем: в резултат на измервания се получават множество стойности на случайните величини $\bar{x}, \bar{y}$. По тях може приближено да се построи съвместната функция на разпределение $F(x,y)$, респективно съвместната вероятностна плътност $f(x,y)$. Ние обаче можем да се интересуваме само от поведението на величината $\bar{x}$, независимо от $\bar{y}$. Плътността на вероятностното разпределение по $\bar{x}$ може да се получи чрез интегриране на $f(x,y)$ по y в цялото пространство на изменение на y:

$$P(a \le x < b, -\infty < y < \infty) = \int_a^b dx \left[\int_{-\infty}^{\infty} dy\ f(x,y) \right] = \int_a^b dx\ g(x) \qquad (4)$$

Тук:

$$g(x) \overset{def}{=} \int_{-\infty}^{\infty} dy\ f(x,y) \qquad (5)$$

е вероятностната плътност на случайната величина $\bar{x}$. $g(x)$ се нарича *безусловна (маргинална) плътност* на разпределението на $\bar{x}$. По подобен начин за $\bar{y}$ се дефинира функцията:

$$h(y) = \int_{-\infty}^{\infty} dx\ f(x,y) \qquad (6)$$

Аналогично на понятието *независими събития* ние можем да определим *независимостта на случайните величини* $\bar{x}$ и $\bar{y}$. Именно, двете случайни величини $\bar{x}$ и $\bar{y}$ се наричат независими, ако е изпълнено условието:

$$f(x,y) = g(x)h(y) \qquad (7)$$

Условната вероятностна плътност също се определя с помощта на маргиналните разпределения:

$$f(y \mid x) = \frac{f(x,y)}{g(x)} \qquad (8)$$

се нарича условна плътност на вероятността за случайната величина $\bar{y}$ при зададено $\bar{x}$.

За *пълната вероятностна плътност* е също в сила съотношението:

$$h(y) \overset{def}{=} \int_{-\infty}^{\infty} dx\ f(x,y) = \int_{-\infty}^{\infty} dx\ f(y \mid x) g(x) \qquad (9)$$

За независими случайни величини:

$$f(y \mid x) = \frac{g(x)h(y)}{g(x)} = h(y) \qquad (10)$$

т.е. знанията ни за статистическото разпределение на случайната величина $\bar{x}$ не увеличават информацията за разпределението на $\bar{y}$ (и обратно).

7.2. Моменти на съвместните разпределения

Аналогично на едномерния случай, определяме математическото очакване на една дадена функция $H(\bar{x},\bar{y})$ на случайните величини $\bar{x}$ и $\bar{y}$ като:

$$E\{H(\bar{x},\bar{y})\} = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\, H(x,y) f(x,y) \qquad (1)$$

а дисперсията:

$$\sigma^2\{H(\bar{x},\bar{y})\} = E\{[H(\bar{x},\bar{y}) - E\{H(\bar{x},\bar{y})\}]^2\} \qquad (2)$$

Нека разгледаме функцията:

$$H(\bar{x},\bar{y}) = \bar{x}^l \bar{y}^m \qquad (3)$$

Математическото очакване на тази функция се нарича (lm) *-ти момент на вероятностното разпределение на случайните величини* $\bar{x}$ *и* $\bar{y}$ *спрямо точката (0,0):*

$$\lambda_{lm} \overset{def}{=} E\{\bar{x}^l \bar{y}^m\} \qquad (4)$$

Това е нецентрален момент. По-общо се разглежда функцията:

$$H(\bar{x},\bar{y}) = (\bar{x}-a)^l (\bar{y}-b)^m \qquad (5)$$

Нейното математическо очакване определя централните моменти на разпределението на $\bar{x}$ и $\bar{y}$ относно точката (a,b):

$$\alpha_{lm} \overset{def}{=} E\{(\bar{x}-a)^l (\bar{y}-b)^m\} \qquad (6)$$

Особен интерес представляват моментите μ_{lm} спрямо точката $(\lambda_{10},\lambda_{01})$, която е точно математическото очакване на $f(x,y)$. Младшите по ранг моменти са:

$$\begin{aligned} &\lambda_{00} = 1 \\ &\lambda_{10} = E\{\bar{x}\} \overset{def}{=} \hat{x} \\ &\lambda_{01} = E\{\bar{y}\} \overset{def}{=} \hat{y} \end{aligned} \qquad (7)$$

$$\mu_{00} = 1$$
$$\mu_{10} = 0$$
$$\mu_{01} = 0$$
$$\mu_{11} = E\{(\bar{x} - \hat{x})(\bar{y} - \hat{y})\} \overset{def}{=} \operatorname{cov}\{\bar{x}, \bar{y}\} \quad (8)$$
$$\mu_{20} = E\{(\bar{x} - \hat{x})^2\} = \sigma^2\{\bar{x}\}$$
$$\mu_{02} = E\{(\bar{y} - \hat{y})^2\} = \sigma^2\{\bar{y}\}$$

7.2.1.Свойства на моментите на съвместните разпределения

При повече от една случайна величина може да се говори и за по-сложни свойства на моментите, свързани с операциите с няколко променливи. Ето някои от тях:

$$E\{a\bar{x} + b\bar{y}\} = aE\{\bar{x}\} + bE\{\bar{y}\} \quad (9)$$

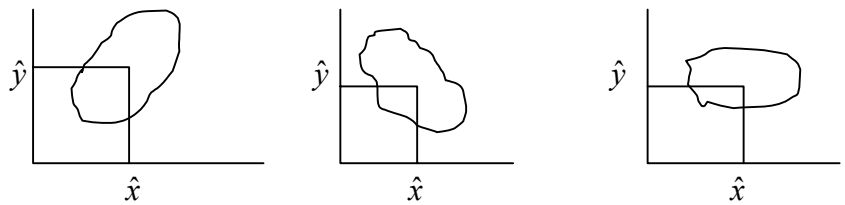
(следва непосредствено от дефиницията);

$$\begin{aligned}\sigma^2\{a\bar{x} + b\bar{y}\} &= E\{[(a\bar{x} + b\bar{y}) - E\{a\bar{x} + b\bar{y}\}]^2\} \\ &= E\{[a\bar{x} + b\bar{y} - a\hat{x} - b\hat{y}]^2\} \\ &= E\{[a^2(\bar{x} - \hat{x})^2 + b^2(\bar{y} - \hat{y})^2 + 2ab(\bar{x} - \hat{x})(\bar{y} - \hat{y})]\}\end{aligned} \quad (10)$$

Или:

$$\sigma^2\{a\bar{x} + b\bar{y}\} = a^2\sigma^2(\bar{x}) + b^2\sigma^2(\bar{y}) + 2ab\operatorname{cov}\{\bar{x}, \bar{y}\} \quad (11)$$

Величините $E\{\bar{x}\}, E\{\bar{y}\}, \sigma^2\{\bar{x}\}, \sigma^2\{\bar{y}\}$, които участвуват тук, са много сходни с математическото очакване и дисперсията за едномерния случай. Що се отнася до $\operatorname{cov}\{\bar{x}, \bar{y}\}$, която се нарича ***ковариация*** на $\bar{x}$ и $\bar{y}$, тя изисква специално внимание предвид голямата й важност. От дефиницията й следва, че ковариацията е положителна, ако събитията $\bar{x} > \hat{x}$ & $\bar{y} > \hat{y}$ (респективно $\bar{x} < \hat{x}$ & $\bar{y} < \hat{y}$) се появяват по-често едновременно, отколкото комбинациите на малки $\bar{x}$ с големи $\bar{y}$. Ковариацията е отрицателна в обратния случай, т.е. когато по-често малки (спрямо средното) стойности на $\bar{x}$ се наблюдават едновременно с големи стойности на $\bar{y}$. Накрая, $\operatorname{cov}\{\bar{x}, \bar{y}\} \approx 0$, ако знанието на $\bar{x}$ не ни дава информация за стойностите на $\bar{y}$ (и обратно). Това се пояснява на следните картинки:



$\operatorname{cov}\{\bar{x}, \bar{y}\} > 0$ $\operatorname{cov}\{\bar{x}, \bar{y}\} < 0$ $\operatorname{cov}\{\bar{x}, \bar{y}\} \approx 0$

В много случаи е удобно вместо ковариацията да се използува величината:

$$\rho(\bar{x},\bar{y}) \overset{def}{=} \frac{\mathrm{cov}(\bar{x},\bar{y})}{\sigma(\bar{x})\sigma(\bar{y})} \tag{12}$$

която се нарича ***коефициент на корелация***.

Ковариацията и коефициентът на корелация дават обобщена оценка за взаимната зависимост между случайните величини $\bar{x}$ и $\bar{y}$.

За да изясним смисъла на коефициента на корелация, разглеждаме нормираните случайни величини:

$$\bar{u} = \frac{\bar{x}-\hat{x}}{\sigma(\bar{x})}; \quad \bar{v} = \frac{\bar{y}-\hat{y}}{\sigma(\bar{y})} \tag{13}$$

Дисперсията на сумата $\bar{u}+\bar{v}$ е:

$$\sigma^2(\bar{u}+\bar{v}) = \sigma^2(\bar{u}) + \sigma^2(\bar{v}) + 2\rho(\bar{u},\bar{v})\sigma(\bar{u})\sigma(\bar{v}) \tag{14}$$

Тъй като за нормираните случайни величини дисперсията е винаги 1, то:

$$\sigma^2(\bar{u}+\bar{v}) = 2[1+\rho(\bar{u},\bar{v})] \tag{15}$$

По аналогичен начин за разликата $\bar{u}-\bar{v}$ се получава:

$$\sigma^2(\bar{u}-\bar{v}) = 2[1-\rho(\bar{u},\bar{v})] \tag{16}$$

Но дисперсията е винаги неотрицателна, $\sigma^2 \geq 0$. Тогава от (15) и (16) следва:

$$-1 \leq \rho(\bar{u},\bar{v}) \leq 1 \tag{17}$$

В сила са следните равенства:

$$\begin{aligned} \rho(\bar{u},\bar{v}) &\overset{def}{=} \frac{\mathrm{cov}(\bar{u},\bar{v})}{1.1} \overset{def}{=} E\{(\bar{u}-0)(\bar{v}-0)\} = E\left\{\frac{(\bar{x}-\hat{x})}{\sigma(\bar{x})}.\frac{(\bar{y}-\hat{y})}{\sigma(\bar{y})}\right\} \\ &= \frac{\mathrm{cov}(\bar{x},\bar{y})}{\sigma(\bar{x})\sigma(\bar{y})} = \rho(\bar{x},\bar{y}) \end{aligned} \tag{18}$$

Следователно, за всеки две случайни величини (не непременно нормирани):

$$-1 \leq \rho(\bar{x},\bar{y}) \leq 1 \tag{19}$$

Оттук виждаме още, че коефициентът на корелация ρ е и нормиран; той не зависи от мащаба по x и по y (за разлика от ковариацията).

Граничните случаи са:

$$\text{а)}\ \rho = +1, \ \Rightarrow \sigma^2(\bar{u}-\bar{v}) = 0, \ \Rightarrow \bar{u}-\bar{v} = \frac{\bar{x}-\hat{x}}{\sigma(\bar{x})} - \frac{\bar{y}-\hat{y}}{\sigma(\bar{y})} = const \tag{20}$$

Това е *функционално* равенство. То може да бъде изпълнено за всяко $\bar{x}$ и всяко $\bar{y}$ само ако е в сила:

$$\bar{y} = a + b\bar{x} \tag{21}$$

т.е. случайните величини $\bar{x}$ и $\bar{y}$ са линейно свързани и $b>0$. Равенството (21) съответствува на случая на ***пълна линейна корелация между случайните величини*** $\bar{x}, \bar{y}$;

б) напълно аналогично при $\rho=-1, \Rightarrow \sigma^2(\bar{u}+\bar{v})=0,$ откъдето следва:

$$\bar{y}=a-b\bar{x} \qquad (22)$$

който случай се нарича ***пълна линейна антикорелация***.

Ако случайните величини $\bar{x}$ и $\bar{y}$ са независими, то $\operatorname{cov}(\bar{x},\bar{y})=\rho(\bar{x},\bar{y})=0.$

Наистина:

$$\operatorname{cov}(\bar{x},\bar{y})=\int_{-\infty}^{\infty}dx\int_{-\infty}^{\infty}dy\,(x-\hat{x})(y-\hat{y})g(x)h(y)=\left\{\int_{-\infty}^{\infty}dx(x-\hat{x})g(x)\right\}\left\{\int_{-\infty}^{\infty}dy(y-\hat{y})h(y)\right\}=0.$$

Обратното обаче не е в сила, т.е. ако ковариацията на две случайни величини е нула, те *не са* непременно независими.

7.3. Съвместни разпределения на повече от две случайни величини

Функцията на разпределение на n случайни величини $\bar{x}_1,\bar{x}_2,...\bar{x}_n$ се определя така:

$$F(x_1,x_2,...,x_n)=P(\bar{x}_1<x_1,\bar{x}_2<x_2,...\bar{x}_n<x_n) \qquad (1)$$

Вероятностната плътност се дефинира по следния начин:

$$f(x_1,x_2,...,x_n)=\frac{\partial^n}{\partial x_1\partial x_2...\partial x_n} \qquad (2)$$

По аналогичен начин се въвеждат всички останали характеристики: маргинално разпределение, математическо очакване, дисперсия, моменти и т.н. Този случай не ни дава нищо ново, но позволява да се направи обобщено разглеждане. Именно, случайните величини $\bar{x}_1,\bar{x}_2,...\bar{x}_n$ могат да се разглеждат като компоненти на вектор в n-мерното пространство: $x=\begin{pmatrix}x_1\\ ...\\ x_n\end{pmatrix}$. Това е вектор-стълб. Транспонираният му вектор е: $x^T=(x_1,x_2,...,x_n)$. Тогава функцията на разпределение се записва така: $F=F(x)$, съответната вероятностна плътност: $f(x)=\dfrac{\partial F(x)}{\partial x}$. Математическото очакване на една функция $H(x)$ добива вида:

$$E\{H(x)\}=\int dxH(x)f(x). \qquad (3)$$

Тук навсякъде x е вектор с n компоненти, интегрирането е също в n-мерното пространство.

Вторите централни моменти - дисперсиите и ковариациите - могат да се обединят в една матрица с размерност $n \times n$, която се нарича ***ковариационна матрица:***

$$c = \begin{pmatrix} c_{11} & \dots & c_{1n} \\ \dots & \dots & \dots \\ c_{n1} & \dots & c_{nn} \end{pmatrix} \qquad (4)$$

Нейните елементи са:

$$c_{ij} = \mathrm{cov}(\bar{x}_i, \bar{x}_j) = E\{(\bar{x}_i - \hat{x}_i)(\bar{x}_j - \hat{x}_j)\} \qquad (5)$$

Диагоналните елементи $c_{ii} = \sigma^2(x_i)$ са дисперсиите (които се наричат още ***вариации***). Матрицата c е симетрична по определение (5).

Като си дадем сметка, че множителите в (5) са компоненти на n-мерен вектор, от правилото за умножение на матрици се получава:

$$c = E\{x - \hat{x})(x - \hat{x})^T\} \qquad (6)$$

Ковариационната матрица (наричана понякога вариационно-ковариационна матрица, а също и матрица на грешките) съдържа основната информация за разсейването на компонентите на един n-мерен случаен вектор. По тази причина нейното пресмятане е една от главните задачи при обработката на експерименталните данни.

8. Трансформации на случайни величини. Разпространение на грешките

Стана ясно, че функцията на една случайна величина е също случайна величина: $\bar{y} = y(\bar{x})$. Въпросът, който възниква често, е: каква е вероятностната плътност $g(y)$, ако е известна плътността $f(x)$?

Общият отговор на този въпрос е свързан с изискването за взаимна еднозначност и непрекъснатост на изображенията $y(x)$ и $x(y)$ и непрекъснатост на частните производни $\frac{\partial x}{\partial y}$. Тогава търсената връзка е от вида:

$$g(y) = \left|\frac{dx}{dy}\right| f(x) \qquad (1)$$

Съотношението (1) има и многомерно обобщение. Ние няма да се занимаваме с общия случай, а ще разглеждаме само най-важния специален случай на ***линейни трансформации***.

Линейните трансформации са най-често използуваните на практика. Причината е в простотата на действията с тях. Те например водят до линейни уравнения и системи, които много по-лесно се решават. Затова при възможност ние се стараем да преминем от нелинейни към линейни трансформации. Обикновено това се постига чрез ограничаване на Тейлъровото (Taylor) разложение на функциите до линейни членове.

Нека x и y са n-мерни вектори от случайните величини $(\bar{x}_1,...,\bar{x}_n)$ и $(\bar{y}_1,...,\bar{y}_n)$. Нека y е линейна функция на x. Това означава:

$$y = a + Tx \tag{2}$$

където a е n-мерен вектор, а $T - n \times n$ матрица.

Математическото очакване на случайната величина y е:

$$E(\bar{y}) \stackrel{def}{=} \hat{y} = a + T\hat{x} \tag{3}$$

Нека да видим как ще изглежда ковариационната матрица на линейната трансформация $y(x)$:

$$\begin{aligned} C_y &= E\{(y-\hat{y})(y-\hat{y})^T\} \\ &= E\{(a+Tx-a-T\hat{x})(a+Tx-a-T\hat{x})^T\} \\ &= E\{T(x-\hat{x})(x-\hat{x})^T T^T\} = TC_xT^T \end{aligned} \tag{4}$$

(Тук помним, че матричното умножение не е комутативно!).

И така:

$$C_y = TC_xT^T \tag{5}$$

С помощта на съотношението (5) ще формулираме едно от най-известните правила в статистиката, наречено ***закон за разпространение на грешките*** (англ. error propagation).

Нека $x = (\bar{x}_1,...,\bar{x}_n)$ е n-вектор от известни измервания. Ние предполагаме, че:

- средните стойности на измерванията са известни; нека те образуват вектора $\hat{x}$;
- средноквадратичните отклонения ("грешките") на x са също известни; те образуват ковариационната матрица C_x.

Ние искаме да разберем каква е "грешката" на някаква функция на измерването $y = y(x)$.

Ако грешките на x са сравнително малки, то вероятностната му плътност $f(x)$ ще е съществено различна от нула само в неголяма околност на $\hat{x}$ (тази околност е от порядъка на $\sigma(x)$). Тогава ще е в сила разложението:

$$y_i = y_i(\hat{x}) + \sum_{j=1}^{n} \left.\frac{\partial y_i}{\partial x_j}\right|_{x=\hat{x}} (x_j - \hat{x}_j) + O(x-\hat{x})^2, \qquad i = 1,...,n \tag{6}$$

Или, в матрична форма:

$$y = y(\hat{x}) + T(x-\hat{x}) + O(x-\hat{x})^2, \quad T_{ij} = \left.\frac{\partial y_i}{\partial x_j}\right|_{x=\hat{x}} \tag{7}$$

Тогава, използувайки (5), получаваме:

$$C_y = TC_xT^T \tag{8}$$

където матрицата на трансформацията T се определя от (7b). Равенството (8) изразява закона за разпространение на грешките при измерванията.

Казаното дотук може да се обобщи така:

- законът за разпространение на грешките (5) е точен за линейни трансформации на случайните величини;
- този закон може да се използува приближено и за нелинейни трансформации на случайни величини във вида (8). Това приближение е добро, когато дисперсиите на базисните случайни величини $\{x\}$ са малки.

Тук трябва да се направи следната важна забележка:

Дисперсиите са, както знаем, диагоналните елементи на съответната ковариационна матрица. Тъй като обикновено ние се интересуваме главно от дисперсиите като мярка за разсейването на резултатите от измерванията, ние сме склонни да оценяваме само $\sigma^2(x)$. Но, забележете, че дисперсиите на трансформациите y зависят не само от дисперсиите на x, но и от ковариациите $\mathrm{cov}(x_i, x_j)$. Без отчитането на ковариациите между измерванията (т.е. техните взаимни зависимости), ние можем да получим съвършено нереалистични резултати за $\sigma^2(y)$. Само ако наблюденията $\{x\}$ са взаимно независими помежду си, тогава $\mathrm{cov}(x_i, x_j) = 0 \ (i \neq j)$ и тогава те могат да се пренебрегнат. Тогава матрицата C_x става диагонална и е в сила:

$$(C_y)_{ii} = \sigma^2(y_i) = \sum_{j=1}^{n}\left(\frac{\partial y_i}{\partial x_j}\right)^2 \sigma^2(x_i) \qquad (9)$$

Последното съотношение е известно също като закон за разпространение на грешките и фигурира в много учебни пособия именно в този вид. Напомняме още веднъж, че (9) е валидно само при независими измервания $\{x\}$.

Ето формули за дисперсията на някои прости функции от случайни величини, пресметнати според закона за разпространение на грешките:

Нека x_1 и x_2 са две ***независими*** случайни величини. Тогава:

$y = x_1 \pm x_2$	$\sigma^2(y) = \sigma^2(x_1) + \sigma^2(x_2)$
$y = x_1.x_2$, $y = \frac{x_1}{x_2}$	$\frac{\sigma^2(y)}{y^2} = \frac{\sigma^2(x_1)}{x_1^2} + \frac{\sigma^2(x_2)}{x_2^2}$
$y = x^n$	$\frac{\sigma^2(y)}{y^2} = \frac{n^2\sigma^2(x)}{x^2}$
$y = e^x$	$\frac{\sigma^2(y)}{y^2} = \sigma^2(x)$

Накратко тези правила могат да се изкажат така:

- при събиране и изваждане се сумират дисперсиите на събираемите;
- при умножение и деление се сумират относителните дисперсии на множителите/делителите.

Ако x_1 и x_2 ***не са независими***, то съответните формули за дисперсията на y в случая на произведение и частно са:

$y = x_1.x_2$	$\frac{\sigma^2(y)}{y^2} = \frac{\sigma^2(x_1)}{x_1^2} + \frac{\sigma^2(x_2)}{x_2^2} + 2\rho(x_1x_2)\frac{\sigma(x_1)}{x_1}\frac{\sigma(x_2)}{x_2}$
$y = \frac{x_1}{x_2}$	$\frac{\sigma^2(y)}{y^2} = \frac{\sigma^2(x_1)}{x_1^2} + \frac{\sigma^2(x_2)}{x_2^2} - 2\rho(x_1x_2)\frac{\sigma(x_1)}{x_1}\frac{\sigma(x_2)}{x_2}$

С други думи, дисперсията на произведението се увеличава при положителна корелация между множителите, а дисперсията на частното намалява при положителна корелация между делителите (в сравнение с

дисперсиите за независими множители/делители). При антикорелация между множителите/делителите е обратното.

IV. Някои важни статистически разпределения

Тук ще разгледаме свойствата на някои конкретни статистически разпределения, които се срещат най-често в практиката на обработката на данни.

9. Дискретни разпределения

9.1. Експеримент с два изхода (разпределение на Бернули)

Най-простият възможен експеримент е този, чиято реализация води до едно от двете несъвместни събития A и $\overline{A}$ (например хвърляне на монета). Той е известен и като *схема на Бернули* (Bernouli). В такъв случай е в сила разложението:

$$E = A + \overline{A} \qquad (1)$$

Означаваме вероятността за "благоприятен изход" (нека за такъв приемем настъпването на събитието A) с:

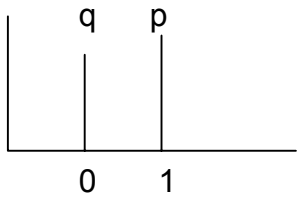
$$P(A) = p, \quad P(\overline{A}) = 1 - p = q \qquad (2)$$

Резултатът от един такъв експеримент може да се изрази чрез една случайна величина $\bar{x}$, която приема стойности 1 (0) в зависимост от това, дали е настъпило събитието $A\,(\overline{A})$. Статистическите характеристики на случайната величина $\bar{x}$ са:

$$E\{\bar{x}\} = \sum xp(x) = 1.p + 0.q = p \qquad (3)$$

$$\sigma^2\{\bar{x}\} = \sum [x - p]^2 p(x) = (1-p)^2.p + (0-p)^2.q = pq^2 + p^2q = pq \qquad (4)$$

Разпределението на тази случайна величина изглежда така:



Това е почти всичко, което може да се каже за тази най-проста случайна величина. Тя е интересна не толкова сама по себе си, колкото с това, че тя участвува в изграждането на всички важни дискретни разпределения.

9.2. Биномно разпределение

Това е вероятностното разпределение на случайна величина, описваща всевъзможните резултати от n-кратното повторение на експеримента с два изхода:

$$\bar{x} = \sum_{i=1}^{n} \bar{x}_i \qquad (5)$$

където $\bar{x}_i$ е случайна величина, описваща единичен експеримент с два изхода. Нейното разпределение ни е вече известно.

Важно допълнение тук е, че отделните единични експерименти се предполагат *независими* помежду си. Тогава вероятността за k на брой благоприятни изходи (т.е. събитието A в предишната схема) при n-кратно повторение ще бъде:

$$W_k^n = \binom{n}{k} p^k q^{n-k} \qquad (6)$$

Отделните членове имат такъв смисъл: $p^k q^{n-k}$ е вероятността точно k на брой реализации на единичния експеримент да дадат благоприятен изход, а $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ е броят на всички възможни начини, по които може да се получат k благоприятни изхода при n на брой експерименти от този тип.

W_k^n е вероятностното разпределение на случайната величина (5). То се нарича ***биномно разпределение***. При него k е *аргумент* (това е броят на благоприятните изходи, който е независимата променлива), а n и p са *параметри* (това са съответно броят реализации на единичния експеримент с два изхода и вероятността за благоприятен изход при единична реализация).

Очевидно:

$$\sum_{k=0}^{n} W_k^n = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = (p+q)^n = 1^n = 1 \qquad (7)$$

така че вероятността (6) е правилно нормирана.

Математическото очакване на биномното разпределение е:

$$\begin{aligned} E\{\bar{x}\} &= \sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k} = \sum_{k=0}^{n} k \frac{n!}{k!(n-k)!} p^k q^{n-k} \\ &= np \sum_{k=0}^{n} \frac{(n-1)!}{(k-1)![n-1-(k-1)]!} p^{k-1} q^{n-1-(k-1)} = np \sum_{l=0}^{n-1} \binom{n-1}{l} p^l q^{n-1-l} = np \end{aligned} \qquad (8)$$

Този резултат може да се пресметне и по-лесно - като използуваме вече известния ни израз за математическото очакване на единичния експеримент с два изхода (3) и правилото за математическото очакване на сума от независими случайни величини:

$$E\{\bar{x}\} = \sum_{i=1}^{n} p = np \qquad (9)$$

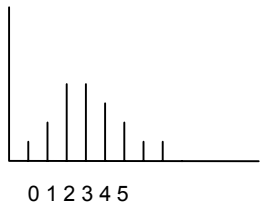
Съответно, за дисперсията на биномното разпределение имаме:

$$\sigma^2\{\bar{x}\} = \sum_{k=0}^{n} (k - np)^2 \; W_k^n \qquad (10)$$

Като използуваме отново формулата за дисперсията на сума от независими случайни величини (ковариациите са нули), получаваме:

$$\sigma^2\{\bar{x}\} = \sum_{k=0}^{n} \sigma^2(x_k) = npq \qquad (11)$$

Ето как изглежда биномното разпределение:



То е дефинирано само за цели неотрицателни стойности на аргумента k, има максимум при $k \approx np$ и е несиметрично (с коефициент на асиметрия $\gamma_1 > 0$).

От свойствата на биномното разпределение може да се направи едно важно наблюдение за *връзката между честота и вероятност.* Именно, да дефинираме честотата на събитието A (благоприятният изход) в n Бернулиеви опита като:

$$\bar{v} = \frac{\bar{x}}{n}. \qquad (12)$$

Величината $\bar{v}$ е случайна величина. Нейното математическо очакване е:

$$E\{\bar{v}\} = \frac{np}{n} = p \qquad (13)$$

а дисперсията ѝ е:

$$\sigma^2(\bar{v}) = \sigma^2\left(\frac{\bar{x}}{n}\right) = \frac{1}{n^2}\sigma^2(\bar{x}) = \frac{p(1-p)}{n} \qquad (14)$$

Какво се вижда оттук?

- Първо, че <u>математическото очакване на честотата съвпада с вероятността</u> (която можем да предполагаме, че е неизвестна в един реален експеримент);
- Второ - <u>с нарастването на n дисперсията на честотата намалява</u>: $\lim_{n\to\infty} \sigma^2(\bar{v}) = 0$. Тъй като в (14) $p(1-p) \le \frac{1}{4}$, то при фиксирано n,

$\sigma(\bar{\nu}) \le \frac{1}{2\sqrt{n}}$. Следователно, с нарастването на n честотата $\bar{\nu}$ на дадено събитие все повече се доближава до вероятността за това събитие. Това свойство на честотата популярно се нарича *закон за големите числа* (ние обаче знаем точната формулировка на този закон: за всяко число $\varepsilon > 0$ и за всяко $\delta > 0$ съществува такова n, че $P(|\nu - p| > \delta) < \varepsilon$).

Ще споменем някои обобщения на биномното разпределение:

Полиномното разпределение е непосредствено обобщение на биномното за експерименти с краен брой изходи (повече от два). Предполагайки, че е в сила разложението:

$$E = A_1 + A_2 + ... + A_l \tag{15}$$

където събитията A_j са несъвместни, и ако

$$P(A_j) = p_j, \ \sum_{j=1}^{l} p_j = 1 \tag{16}$$

то вероятността при n опита всяко от събитията A_j да се случи k_j пъти $(j = 1,...,l)$ се дава с:

$$W^n_{(k_1,k_2,...,k_l)} = \frac{n!}{\prod_{j=1}^{l} k_j!} \prod_{j=1}^{l} p_j^{k_j} \tag{17}$$

(На това разпределение се подчиняват например игрите със зар).

По-сложно е т.нар. *хипергеометрично разпределение*, което описва вероятността от n извадени топки (бели и черни) k да са бели и $l = n - k$ - черни, ако K и L са всичките бели и черни топки и $N = K + L$ е общият брой топки в урната. Това е изваждане без повторение. В този случай последователните опити не са независими, тъй като те зависят от предисторията (така например цветът на последната топка е еднозначно известен от предишните резултати).

Тази зависимост може да се усили, ако всеки път, когато извадим топка от определен цвят, добавим в урната фиксиран брой топки от същия вид. Това е *разпределението на Пойа*, което има отношение към възникването на епидемии - тогава появяването на един случай увеличава вероятността за подобни заболявания.

Нека да се върнем към биномното разпределение и да разгледаме граничните случаи. Те са два:

- при фиксирана вероятност p за благоприятен изход при единичния експеримент, ако броят на неговите повторения n нараства неограничено, биномното разпределение клони към *нормално разпределение (разпределение на Гаус (Gauss))*;
- ако с нарастването на n вероятността за благоприятен изход p намалява така, че $np = const$, то биномното разпределение при $n \to \infty$ клони към *разпределението на Поасон (Poisson)*.

9.3. Разпределение на Поасон

9.3.1. Разпределението на Поасон като граничен случай на биномното разпределение

Нека означим:

$$\lambda = np \tag{1}$$

Предполагаме, че λ е фиксирано. Тогава:

$$W_k^n = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k \frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^k} = \frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n \frac{n(n-1)...(n-k+1)}{n^k}\frac{1}{\left(1-\frac{\lambda}{n}\right)^k} \tag{2}$$

При $n \to \infty$ първият множител в (2) е $const$, третият и четвъртият клонят всеки към единица, а вторият клони към експоненциална функция. Следователно:

$$W_k^n \xrightarrow[n\to\infty]{} \frac{\lambda^k}{k!} e^{-\lambda} \tag{3}$$

9.3.2. Директен извод на разпределението на Поасон

Този подход се основава на конкретен физически модел и по тази причина изяснява по-добре свойствата на Поасоновото разпределение от гледна точка на физическия му смисъл.

Нека имаме радиоактивен източник с множество радиоактивни ядра, които се разпадат спонтанно. При това са в сила следните предположения:

1. Отделните събития са независими във времето (т.е. наличието на разпадане в момента t не зависи от предисторията);

2. Вероятността за отделно събитие (разпадане) за *малък интервал* от време δt е пропорционална на дължината на интервала: $p(t, t+\delta t) = \mu \delta t + O(\delta t)^2$ (4)
3. Вероятността за повече от едно събитие (разпадане) *в малък интервал* от време δt е 0.

Ние търсим вероятността $P_k(t)$ в интервала време $(0,t)$ да се случат k на брой разпадания. За целта ще образуваме $P(t)$ и $P(t+\delta t)$ и ще получим диференциално уравнение за $P(t)$ при $\delta t \to 0$.

Първо търсим P_0. За да има нула разпадания в интервала $(0, t+\delta t)$ трябва да няма разпадания както в интервала $(0,t)$, така и в интервала $(t, t+\delta t)$. Съгласно първото условие тези събития са независими, следователно:

$$P_0(t+\delta t) = P_0(t)(1-\mu\delta t) \quad (5)$$

(напомняме, че $1-\mu\delta t$ е вероятността за неразпадане в интервал с дължина δt). Следователно:

$$\frac{P_0(t+\delta t) - P_0(t)}{\delta t} = -\mu P_0(t) \quad (6)$$

При $\delta t \to 0$:

$$\frac{dP_0}{dt} = -\mu P_0(t) \quad (7)$$

Общото решение на (7) е:

$$P_0(t) = Ae^{-\mu t} \quad (8)$$

където A е произволна константа.

За да се определи стойността на A, е необходимо едно начално условие. Логично е да предпоставим:

$$P_0(0) = 1 \quad (9)$$

т.е. вероятността за неразпадане в интервал с дължина нула е равна на тази на достоверното събитие. Тогава, окончателно:

$$P_0(t) = e^{-\mu t} \quad (10)$$

Това всъщност е известният закон за радиоактивното разпадане.

Нека сега $k=1$, т.е. търсим вероятността за *точно едно* разпадане в интервала $(0, t+\delta t)$. Има две възможности: едно разпадане в $(0,t)$ и нула разпадания в $(t, t+\delta t)$ или обратното. Тези сложни събития са несъвместни, така че общата вероятност е сума от техните вероятности:

$$P_1(t+\delta t) = P_1(t)(1-\mu\delta t) + P_0(t)\mu\delta t \qquad (11)$$

Следователно:

$$\frac{P_1(t+\delta t) - P_1(t)}{\delta t} = -\mu P_1(t) + \mu e^{-\mu t} \qquad (12)$$

Това уравнение има решение:

$$P_1(t) = \mu t e^{-\mu t} \qquad (13)$$

Това е вероятността за точно едно разпадане в интервала $(0,t)$.

Горното разглеждане е напълно приложимо и за общия случай ($k \geq 1$):

$$P_k(t+\delta t) = P_k(t)(1-\mu\delta t) + P_{k-1}(t)\mu\delta t \qquad (14)$$

тъй като според третото предположение на модела повече от едно разпадане в интервала $(t, t+\delta t)$ не е възможно. Следователно:

$$\frac{P_k(t+\delta t) - P_k(t)}{\delta t} = -\mu P_k(t) + \mu P_{k-1}(t) \qquad (15)$$

При $\delta t \to 0$, това диференциално уравнение има решение:

$$P_k(t) = \frac{(\mu t)^k}{k!} e^{-\mu t} \qquad (16)$$

което е всъщност Поасоновото разпределение.

Разглеждайки тази картина статично, (т.е. като анализираме броя на наблюдаваните разпадания за даден фиксиран интервал от време), ние виждаме, че $\lambda = \mu t$ има смисъл на среден брой разпадания за дадения интервал, а P_k е вероятността да получим k на брой разпадания, ако средният им брой е λ. Оттук се вижда и физическият смисъл на параметъра μ; това е *средната скорост на разпадане* на радиоактивните ядра.

9.3.3. Свойства на Поасоновото разпределение

Разпределението на Поасон (3) е дефинирано за всички цели неотрицателни стойности на своя аргумент k. То зависи от един параметър λ, който е реално неотрицателно число.

Нормировка:

$$\sum_{k=0}^{\infty} P_k(\lambda) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} \equiv 1 \qquad (17)$$

откъдето се вижда, че разпределението, дефинирано чрез (3), е нормирано коректно.

Средна стойност:

$$E\{P(\lambda)\} = e^{-\lambda}\sum_{k=0}^{\infty} k\frac{\lambda^k}{k!} = \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = \lambda \tag{18}$$

Дисперсия:

$$\sigma^2\{P\} = e^{-\lambda}\sum_{k=0}^{\infty}(k-\lambda)^2\frac{\lambda^k}{k!} \tag{19}$$

Развиваме израза за квадрата в скобите. Вторият и третият членове са от познат вид. Първият е:

$$\begin{aligned} e^{-\lambda}\sum_{k=0}^{\infty}k^2\frac{\lambda^k}{k!} &= \lambda e^{-\lambda}\sum_{k=1}^{\infty}k\frac{\lambda^{k-1}}{(k-1)!} \\ &= \lambda e^{-\lambda}\sum_{j=0}^{\infty}(j+1)\frac{\lambda^j}{j!} = \lambda(\lambda+1) \end{aligned} \tag{20}$$

Следователно:

$$\sigma^2\{P(\lambda)\} = \lambda^2 + \lambda - 2\lambda^2 + \lambda^2 = \lambda \tag{21}$$

Забележка: Дисперсията може да се пресметне и като се използува резултата (9.11), за дисперсията на биномното разпределение, тъй като Поасоновото разпределение е всъщност един негов частен случай. Именно:

$$\sigma^2\{P(\lambda)\} = npq = \lambda(1-\lambda/n) \xrightarrow[n\to\infty]{} \lambda \tag{22}$$

Асиметрия:

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{\sum_{\kappa=0}^{\infty}(k-\lambda)^3\frac{\lambda^k}{k!}e^{-\lambda}}{\lambda^{3/2}} = \frac{1}{\sqrt{\lambda}} \tag{23}$$

От написаното се вижда, че:

- параметърът λ действително има смисъл на средна стойност, което оправдава названието му при втория подход (среден брой разпадания);
- дисперсията на Поасоновото разпределение е равна на средната му стойност;
- асиметрията на Поасоновото разпределение е винаги положителна; с увеличаването на λ тя намалява, т.е. разпределението става по-симетрично.

Поасоновото разпределение се получава от биномното при голям брой повторения на експеримент с малка вероятност за благоприятен изход; поради това то е известно още като *разпределение на редките събития.*

Радиоактивното разпадане е вероятно най-известният пример за явление, подчиняващо се на Поасоновото разпределение. Има обаче още много други такива явления:

- броят на научните открития, направени неколкократно при независими изследвания;
- броят на звездите в определен ъглов диапазон на небесния свод,
- и др.

10. Непрекъснати разпределения

10.1. Равномерно разпределение

Това е най-простият пример за непрекъснато разпределение. Едно статистическо разпределение се нарича *равномерно разпределение*, ако вероятностната му плътност е константа в някакъв краен интервал и нула извън него:

$$f(x)=\begin{cases} c & a\le x\le b \\ 0 & x<a,\quad x>b \end{cases} \tag{1}$$

Нормировка:

$$\int_{-\infty}^{\infty} dx f(x)=1 \;\Rightarrow c(b-a)=1 \;\Rightarrow c=\frac{1}{b-a} \tag{2}$$

Функция на разпределение:

$$F(x)=\int_{-\infty}^{x} dx\, f(x)=\begin{cases} 0 & x<a \\ \dfrac{x-a}{b-a} & a\le x\le b \\ 1 & x>b \end{cases} \tag{3}$$

Математическо очакване:

$$E(\bar{x})=\hat{x}=\int_{-\infty}^{\infty} dx\, xf(x)=\frac{a+b}{2} \tag{4}$$

Дисперсия:

$$\sigma^2(\bar{x})=\frac{1}{12}(b-a)^2 \tag{5}$$

Равномерното разпределение само по себе си е от неголямо значение за практически приложения. То обаче е много удобно за пресмятания и играе

важна роля при трансформациите на статистическите разпределения (например при моделирането със случайни числа).

10.2. Нормално разпределение

10.2.1. Лапласов модел на грешките

В произведението си "Философско есе върху теорията на вероятностите" през 1783г. Лаплас (P.Laplace) е разгледал въпроса за произхода на грешките в наблюденията. Той е предполагал, че:

- истинската стойност на измеряемата величина съществува обективно и има някаква точна и постоянна стойност, равна на някаква константа m;
- m не може да се измери точно, тъй като резултатите от измерването се изкривяват от голям брой (n) независими помежду си смущения, всяко от които е с големина ε;
- всяко едно от смущенията с равна вероятност води до отклоняване на стойността на наблюдаемата величина на $\pm\varepsilon$;
- грешката на измерването се образува от сумирането на влиянието на отделните смущения.

Ясно е, че в този модел вероятностното разпределение на резултатите от измерванията се описва с биномно разпределение. Развитието на влиянието на грешките върху стойността на резултатите от наблюденията изглежда по следния начин:

бр.смущения	-3ε	-2ε	-1ε	m	+1ε	+2ε	+3ε
0				1			
1			1/2		1/2		
2		1/4		1/4+1/4		1/4	
3	1/8		1/8+1/4		1/4+1/8		1/8

В таблицата е дадена вероятността да се получи даден резултат от измерването след въздействието на определен брой смущения. В началната точка отсъствуват смущения и затова с вероятност 1 ние получаваме истинската стойност на измеряемата величина m. С появата на едно смущение с равна вероятност (1/2) се получава резултат $m-\varepsilon$ или $m+\varepsilon$. Същото става и при всяко следващо смущение, т.е. всеки ред се получава от предишния чрез

сумиране на ефекта от предишните смущения и последното по схемата на Лаплас. Това е впрочем вероятността W_k^n за $p = q = 1/2$.

Сега, по-интересно е да видим какво става при неограничено нарастване на броя на смущенията: $n \to \infty$. Макар че в оригиналния модел на Лаплас $p = q = 1/2$, то резултатът, който следва, е валиден и за $p \neq q$ (т.е. за асиметрия в посоките на отклонение на резултата под действието на смущението).

Този резултат (известен като локална теорема на Моавр-Лаплас (Moivre-Laplace), която няма да доказваме) гласи, че за всички нормирани бернулиеви случайни величини $\bar{x}$ от вида:

$$\bar{x} = \frac{k - np}{\sqrt{npq}} \tag{1}$$

където k е някое цяло число, при $n \to \infty$ е в сила:

$$P(x) \equiv \Phi_0(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{2}$$

По-общо, разпределението (2) има вероятностна плътност от вида:

$$\Phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-x_0)^2}{2\sigma^2}} \tag{3}$$

Математическото очакване на случайната величина с вероятностна плътност (3) е:

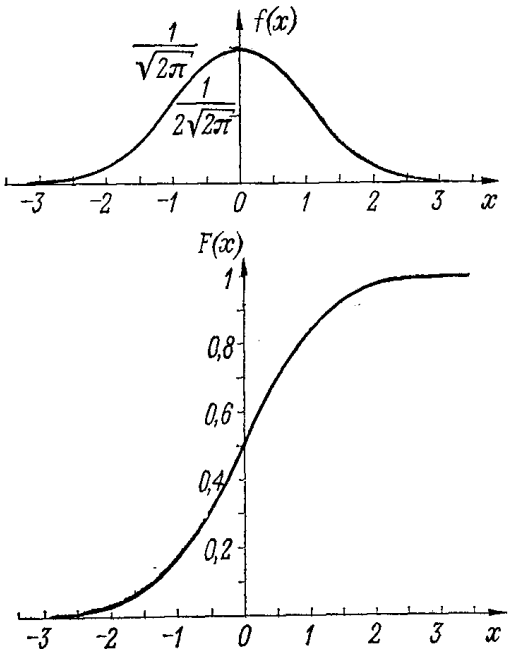
$$E(\bar{x}) = x_0 \tag{4}$$

а дисперсията ѝ е:

$$\sigma^2(\bar{x}) = \sigma^2 \tag{5}$$

Това разпределение се нарича ***нормално, или Гаусово***. То заема централно място в математическата статистика и по-специално в теорията на грешките. Разпределението, описвано с вероятностната плътност (2), се нарича ***стандартно нормално разпределение***. То има средна стойност нула и дисперсия 1.

10.2.2. Свойства на нормалното разпределение



Вероятностната плътност на нормалното разпределение (3) представлява камбановидна крива, симетрична спрямо точката x_0. Тя има инфлексни точки при стойности на аргумента $x = \pm\sigma$, където втората производна става нула и наклонът на кривата се сменя. Това разпределение зависи от два параметъра - средната стойност x_0 и дисперсията σ^2. Използува се и означението:

$$N(x_0, \sigma^2) \equiv \Phi(x) \qquad (6)$$

Функцията на разпределение на стандартното нормално разпределение е:

$$\Psi_0(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} dt e^{-\frac{t^2}{2}} \qquad (7)$$

Тази функция е свързана с известната специална функция $erf(x)$ (функция на грешките, англ. error function):

$$erf(x) \overset{def}{=} \frac{2}{\sqrt{\pi}} \int_{0}^{x} dt e^{-t^2} \qquad (8)$$

Именно:

$$\Psi_0(x) = \frac{1}{2} + erf\left(\frac{x}{\sqrt{2}}\right)$$

Функцията (7) се нарича *интеграл на вероятностите (probabilities integral)*.

Вероятността една нормално разпределена случайна величина $\bar{x}$ да се отличава от средната си стойност с по-малко от $n\sigma$ е:

$$P(|\bar{x} - x_0| < 1\sigma) = 0.682$$
$$P(|\bar{x} - x_0| < 2\sigma) = 0.954 \qquad (9)$$
$$P(|\bar{x} - x_0| < 3\sigma) = 0.998$$

От тези съотношения се вижда, че нормално разпределената случайна величина е групирана в близка околност на средната си стойност (с радиус от порядъка на σ). По-нататък ще видим, че в много случаи в добро приближение може да се счита, че *неопределеностите (грешките) при измерванията са разпределени нормално около средната стойност* x_0. В такъв случай стандартното отклонение σ може да бъде разглеждано като мярка за големината на случайните грешки при измерванията.

Едно друго важно свойство на нормалното разпределение е, че сумата на две независими случайни величини с Гаусови разпределения е също нормално разпределена със средна стойност $\hat{x} = x_1 + x_2$ и с дисперсия $\sigma^2 = \sigma_1^2 + \sigma_2^2$. Това свойство обаче е само частен случай от един много по-общ резултат, известен като *централна гранична теорема*. Самата централна гранична теорема (ЦГТ) има различни формулировки с различна степен на общност. Може да се установи например, че ако $\{\bar{x}_i, i = 1..n\}$ е множество от независими случайни величини, еднакво разпределени със средни стойности x_0 и дисперсии σ^2, то сумата $\frac{1}{n}\sum_{i=1}^{n}\bar{x}_i$ при $n \to \infty$ се подчинява на Гаусово разпределение с параметри $N(x_0, \frac{\sigma^2}{n})$. Забележете, че в тази формулировка никъде не става дума за конкретния вид на вероятностното разпределение на всяка от величините $\{\bar{x}_i\}$ (т.е. тяхното разпределение може да е произволно; достатъчно е само средните стойности и дисперсиите да са еднакви, и теоремата пак ще е вярна!). Може да се покаже, че сумите от случайни величини се подчиняват на нормално разпределение дори и при по-слаби предположения от тези, в частност когато не всичките $\{\bar{x}_i\}$ имат еднакви разпределения. Ние няма да се занимаваме с най-общата формулировка на ЦГТ, но е важно да знаем, че нормалното разпределение е добро приближение за статистическото разпределение на измерваните величини в много широк клас случаи.

10.2.3. Нормалното разпределение и грешките при експеримента: модел на Хершел (J.Hershel)

С помощта на ЦГТ Лапласовият модел за грешките може да бъде изтълкуван като модел за сумарното въздействие на малки индивидуални ефекти върху резултата от измерването. Първоначалното предположение за еднакви амплитуди на всичките смущения сега може да се отхвърли и въпреки това се получава нормално разпределение на грешките в резултата (вж. предишния раздел). В стремежа да се изяснят границите на приложимост на нормалното разпределение като модел за описание на грешките при измерванията е интересно да се разгледа модела на астронома Хершел, който е извел нормалното разпределение при много слаби предположения.

Хершел разглежда падането от точка O върху една равнина на малки топчета. Поради случайни смущения не всички топчета падат върху една и съща точка от равнината. Ние се интересуваме при това от вероятностната плътност на разпределение на координатите на топчетата, т.е. от вероятността за попадане на едно топче върху елементарната площ dA. Този елемент може да се разглежда както в декартови, така и в полярни координати. Предполагаме, че са взети мерки да няма систематични отклонения от центъра; тогава всички ъгли на отклонение са равновероятни (т.е. имаме ***ъглова симетрия***). В такъв случай вероятностната плътност в полярни координати ще зависи само от радиус-вектора. Ние я означаваме с $h(r)$.

Предполагаме още, че попадането на топчето в интервала $(x, x+dx)$ е ***независимо*** от попадането му в интервала $(y, y+dy)$. Поради вече предположената ъглова симетрия, вероятностната плътност по x и по y ще има еднакъв вид, т.е. ще е в сила:

$$g(x,y) = f(x)f(y) \qquad (1)$$

Или, вероятността за попадане в елементарната площ dA ще бъде:

$$h(r)dA = f(x)f(y)dA \qquad (2)$$

Логаритмуваме двете страни:

$$\ln h(r) = \ln f(x) + \ln f(y) \qquad (3)$$

Преминавайки и в дясната страна на (3) към полярни координати, имаме:

$$\ln h(r) = \ln f(r\cos\theta) + \ln f(r\sin\theta) \qquad (4)$$

Сега диференцираме двете страни спрямо полярния ъгъл θ:

$$0 = -r\sin\theta \frac{f'(r\cos\theta)}{f(r\cos\theta)} + r\cos\theta \frac{f'(r\sin\theta)}{f(r\sin\theta)} \qquad (5)$$

Връщайки се отново в декартови координати, получаваме:

$$0 = -y\frac{f'(x)}{f(x)} + x\frac{f'(y)}{f(y)} \tag{6}$$

Прегрупираме и стигаме до израза:

$$\frac{f'(x)}{xf(x)} = \frac{f'(y)}{yf(y)} \tag{7}$$

Поради независимостта на x и y това функционално равенство може да бъде изпълнено *за всяко* x, y *само* ако всяка от двете му страни поотделно е равна на константа:

$$\frac{f'(x)}{xf(x)} = const = c \tag{8}$$

Следователно (интегрирайки (8) по x):

$$\begin{aligned} \ln[f(x)] &= \frac{1}{2}cx^2 + const \\ f(x) &= k\exp\left(\frac{1}{2}cx^2\right) \end{aligned} \tag{9}$$

Тъй като вероятността трябва да е ограничена при $x \to \infty$, то очевидно трябва да е изпълнено: $c < 0$. Означаваме: $c = -\sigma^2$. Тогава получаваме последователно:

$$\begin{aligned} f(x) &= k\exp\left(-\frac{1}{2}\frac{x^2}{\sigma^2}\right) \\ h(r) &= k^2\exp\left(-\frac{1}{2}\frac{(x^2+y^2)}{\sigma^2}\right) \\ h(r) &= k^2\exp\left(-\frac{1}{2}\frac{r^2}{\sigma^2}\right) \end{aligned} \tag{10}$$

Какво се оказа? При много общи и много естествени предположения за ***<u>независимост и симетрия</u>*** ние получихме нормално разпределение за отклоненията. Това е демонстрация на правдоподобността при използуване на Гаусовото разпределение като модел на грешките при измерванията. Това заключение е от решаваща важност в много случаи, особено в теорията на метода на най-малките квадрати.

Въпреки това ние трябва да помним, че *нормалното разпределение не е природен закон*. Причините за появата на неопределености при измерванията са твърде многообразни, за да могат да се обхванат с една формула всевъзможните им проявления. Симетрията и независимостта, макар и да са

много общи качества, не винаги са изпълнени в конкретните експерименти. Затова трябва да се проверява нормалността на разпределението на отклоненията винаги, когато това е възможно, още *преди* да са извършени (понякога твърде обширните) пресмятания, които са валидни обаче само за случая на Гаусово разпределение.

10.3. Разпределение χ^2

Тук ще дадем кратки сведения (без доказателства) за още едно важно непрекъснато разпределение.

Нека $\{\bar{x}_i, i = 1,..., n\}$ са независими случайни величини, които имат стандартно нормално разпределение $N(0,1)$. Тогава сумата:

$$\bar{u} = \sum_{i=1}^{n} \bar{x}_i^2 \qquad (1)$$

е случайна величина, чиято вероятностна плътност има вида:

$$f(u) = \frac{1}{\Gamma(\lambda)2^{\lambda}} u^{\lambda-1} e^{-\frac{u}{2}} \qquad (2)$$

където:

$$\lambda = \frac{n}{2}$$
$$\Gamma(x+1) = \int_0^{\infty} dt t^x e^{-t} \qquad (3)$$

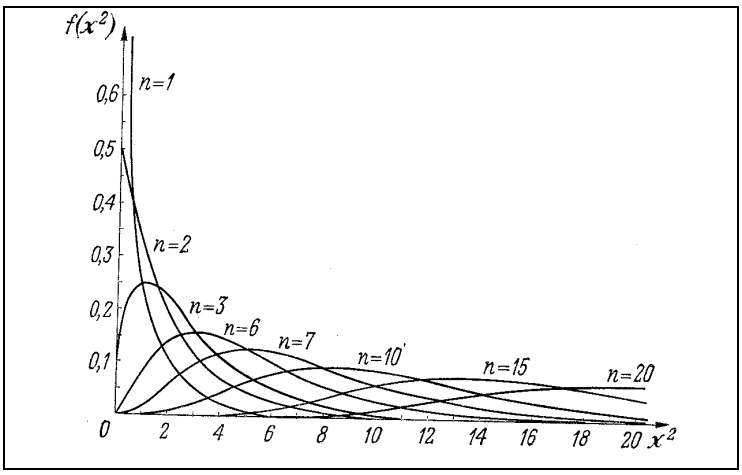
$\Gamma(x)$ е гама-функцията на Ойлер (L.Euler). Числото n, което е единственият параметър на разпределението (2), се нарича *брой на степените на свобода*. Разпределението (2) се нарича *разпределение хи-квадрат (*χ^2*)*.

Едно важно свойство на разпределението χ^2 е следното:

сумата на две независими случайни величини с разпределения χ^2 *съответно с* n_1 *и* n_2 *степени на свобода се подчинява на разпределение* χ^2 *с* n_1+n_2 *степени на свобода.*

Математическото очакване и дисперсията на χ^2 разпределението с плътност (2) са съответно:

$$E\{\bar{\chi}^2\} = 2\lambda = n$$
$$\sigma^2\{\bar{\chi}^2\} = 4\lambda = 2n \qquad (4)$$



На фигурата е показана вероятностната плътност на разпределението χ^2 за различен брой степени на свобода. Вижда се, че с увеличаването на n разпределението става все по-симетрично (и все по-широко).

Разпределението χ^2 намира приложение поне в два много важни случая:

- при проверката на хипотезата за вероятностната плътност на случайна величина, измерена в някакъв експеримент (χ^2 критерий за съгласие);
- при проверката на резултатите от оценките по метода на най-малките квадрати.

Ако величините $\{\bar{x}_i, i = 1,...,n\}$ имат разпределение не $N(0,1)$, а $N(x_0, \sigma^2)$, то величината:

$$\bar{\chi}^2 = \sum_{i=1}^{n} \frac{(\bar{x}_i - x_0)^2}{\sigma^2} \qquad (5)$$

се подчинява на χ^2 разпределение с n степени на свобода.

V. Оценяване на параметри чрез извадки

Досега ние разглеждахме общия вид на статистическите разпределения, а също така и редица техни специфични свойства. Ние обаче не се интересувахме от това, как тези свойства се проявяват при отделните експерименти. Фактически, вероятностната плътност в реалните ситуации зависи от някои параметри (напр. средната стойност при Поасоновото разпределение, средната стойност и дисперсията при Гаусовото разпределение и пр.). Числените стойности на тези параметри като правило са неизвестни в един конкретен експеримент. Следователно, ние не знаем *точното* вероятностно разпределение и сме принудени да го апроксимираме с някакво честотно разпределение, каквото може да се получи експериментално (трябва да си спомним, че конкретните експерименти са свързани винаги с краен брой измервания). Съвкупността от експерименти, провеждани с тази цел, се нарича ***извадка***.

11. Извадки

11.1. Случаен избор. Разпределение на извадките

Извадката се прави измежду множество елементи. Размерът на това множество е обикновено безкраен. То всъщност се състои от всевъзможните изходи от експериментите от даден вид. Такова множество се нарича *популация* (често се използува и терминът *генерална съвкупност*).

Ако е осъществена извадка от n елемента, казва се, че *извадката е с обем* n. Важно е да се знае, че обемът на извадката е винаги краен.

Нека разпределението на случайната величина $\bar{x}$ в популацията се описва с вероятностната плътност $f(x)$. Ние се интересуваме от стойностите на $\bar{x}$, които се съдържат в елементите на извадката. Предполагаме, че сме извършили последователно l извадки с обем за всяка от тях n и при това сме получили следните резултати:

$$\begin{aligned} &x_1^1, x_2^1, ..., x_n^1 \\ &... \\ &x_1^l, x_2^l, ..., x_n^l \end{aligned} \qquad (1)$$

Резултатът от всяка извадка може да бъде групиран в една n-мерна случайна величина:

$$x^j = (x_1^j, x_2^j, ..., x_n^j) \qquad (2)$$

която може да се разглежда като вектор в пространството на n-мерните извадки (напомняме, че случайната величина $\bar{x}$, която всъщност изучаваме и чиито всевъзможни реализации образуват популацията, е само една). Тази случайна величина има вероятностна плътност:

$$g(x) = g(x_1, x_2, ..., x_n) \qquad (3)$$

Ние казваме, че плътността $g(x)$ описва процеса на *случаен избор* тогава и само тогава, когато са изпълнени следните условия:

- случайните величини $\bar{x}_i$ са независими, т.е.:

  $$g(x) = g_1(x_1)g_2(x_2)...g_n(x_n); \qquad (4)$$

- всяка вероятностна плътност в (4) е равна на вероятностната плътност на случайната величина $\bar{x}$ в популацията:

  $$g_1(x) = g_2(x) = ... = g_n(x) = f(x) \qquad (5)$$

Ако не е уговорено друго, по-нататък ще считаме, че извадките, които разглеждаме, са резултат именно от случаен избор.

Трябва да се отбележи, че в реалния процес на избор е много трудно да се осъществи неговата случайност, а пък още по-трудно е да се убедим, че сме я постигнали. Това означава да премахнем систематичните отклонения в процеса на измерване и да оставим да действуват само случайни фактори (и то едни и същи в целия процес на получаване на извадката). На по-обичаен език това означава да осигурим условия за *повторяемост на резултатите от експеримента*. Това е едно от главните умения на добрия експериментатор.

Сега да предположим, че все пак сме осигурили условията за случаен избор. Нека елементите на извадката да са подредени в редица по възходящите стойности на $\bar{x}$ (съгласно правилата на случайния избор това не е ограничение на общността на разглеждане), и нека n_x е броят на елементите, за които $\bar{x} < x$. Тогава функцията:

$$W_n(x) = n_x / n \qquad (6)$$

може да се разглежда като емпирична функция на вероятностно разпределение на случайната величина $\bar{x}$. Тази функция е стъпаловидна и тя нараства с $1/n$ всеки път, когато $\bar{x}$ стане равно на някоя от стойностите на така подредените елементи на извадката. Тази функция (6) се нарича *разпределение на извадката*. Очевидно, $W_n(x)$ е приближение на функцията на

разпределение на популацията $F(x)$, към която се стреми при неограничено нарастване на n.

Всяка функция на елементите на извадката (2) е случайна величина. Такива случайни величини се наричат ***статистики***. Най-известният пример за статистика е ***средното аритметично на извадката***:

$$\tilde{x} = \frac{1}{n}(x_1 + x_2 + ... + x_n) \qquad (7)$$

11.2. Оценки на параметри. Свойства на оценките

Една типична задача на анализа на данни е следната: Предполагаме, че е известен общият вид на вероятностната плътност на (разпределението на) популацията. От експеримента е получена една извадка, от която трябва да се получи числовата стойност на един (или повече) параметър/ри. Поради крайния обем на извадката резултатът не може да бъде точен. Така възниква задачата за *оценка на параметри*.

Тъй като оценяваната стойност на параметъра се получава с помощта на извадката, то очевидно това е някаква *статистика* от специфичен вид; такава статистика се нарича ***оценка***: $S = S(x_1,...,x_n)$. Казва се още, че статистиката S служи за оценка на параметъра λ или пък че S е оценка за λ.

Една оценка се нарича ***неотместена***, ако за извадка с произволен обем нейното математическо очакване е равно на оценявания параметър:

$$E\{S(x_1,...,x_n)\} = \lambda \quad (\forall n) \qquad (8)$$

Една оценка се нарича ***състоятелна***, ако тя се схожда по вероятност към стойността на оценявания параметър (т.е. при неограничено нарастване на обема на извадката точността на оценката също неограничено нараства). Или, за състоятелните оценки е в сила:

$$\lim_{n\to\infty} \sigma(S) = 0 \qquad (9)$$

Накрая, не всички оценки са еднакво ***ефективни***. За сравняване на ефективността на две оценки за един и същ параметър може да се използува съотношението:

$$\eta = \frac{\sigma^2(S_1)}{\sigma^2(S_2)} \qquad (10)$$

По-ефективната оценка има при един и същ обем на извадката по-малка дисперсия. По-нататък ще видим, че понякога е възможно да се установи точна долна граница на дисперсията за всевъзможните оценки на даден параметър.

Естествено, ако съществува оценка, за която се достига тази долна граница на дисперсията, то тя е за предпочитане пред останалите, тъй като е по-ефективна от всички тях.

11.3. Оценки за средната стойност и дисперсията на вероятностното разпределение

Средната стойност и дисперсията на вероятностното разпределение на популацията са най-често оценяваните параметри при физичните експерименти. Ето защо ние искаме да построим подходящи статистики, които да се използуват за оценки на тези параметри.

Разглеждаме извадка с обем n, получена чрез случаен избор от дадена популация на случайната величина $\bar{x}$, описваща се с вероятностна плътност $f(x)$. Ще установим, че *средното аритметично на извадката*:

$$\tilde{x} = \frac{1}{n}(\bar{x}_1 + \bar{x}_2 + ... + \bar{x}_n) \qquad (1)$$

е неотместена и състоятелна оценка за средното на популацията.

Тъй като $\tilde{x}$ всъщност е функция на случайни величини, то самото е също случайна величина. Тя е статистика, тъй като е функция на елементите на извадката. Нейното математическо очакване е:

$$E(\tilde{x}) = \frac{1}{n}E(\sum \bar{x}_i) = \frac{1}{n}n\hat{x} = \hat{x} \qquad (2)$$

т.е. то е равно на математическото очакване на популацията (и то за всяко n); следователно, (1) представлява неотместена оценка за средното на популацията.

За да установим, че статистиката (1) е и състоятелна оценка, ние пресмятаме нейната дисперсия:

$$\sigma^2(\tilde{x}) = E\{[\tilde{x} - E(\tilde{x})]^2\} = E\left\{\left[\frac{\bar{x}_1 + \bar{x}_2 + ... + \bar{x}_n}{n} - \frac{n\hat{x}}{n}\right]^2\right\}$$
$$= E\left\{\frac{1}{n^2}\left[\sum_{i=1}^{n}(\bar{x}_i - \hat{x})\right]^2\right\} \qquad (3)$$

Тъй като компонентите на извадката $\bar{x}_i$ са взаимно независими (по определение това е случаен избор), то смесените членове в сумата от вида $E\{(\bar{x}_i - \hat{x})(\bar{x}_j - \hat{x})\}$ са всичките нули при $i \neq j$ (това са всъщност ковариациите). Оттук получаваме:

$$\sigma^2(\tilde{x}) = \frac{1}{n^2} n\sigma^2(x) = \frac{1}{n}\sigma^2(\bar{x}) \qquad (4)$$

Тук $\bar{x}$ е кой да е елемент от извадката (те всичките имат еднакво разпределение и следователно, равни дисперсии). От (4) се вижда, че $\lim_{n\to\infty}\sigma^2(\tilde{x}) = 0$, с което се установява състоятелността на оценката (1) за математическото очакване на популацията $\hat{x}$. Ще отбележим още, че стойността на всяко отделно измерване (т.е. всяка случайно избрана стойност $\bar{x}_i$) е също така неотместена оценка за $\hat{x}$. Тя обаче не е състоятелна.

Сега да потърсим оценка за дисперсията на популацията. Определяме (засега) *<u>дисперсията на извадката</u>* като средно аритметично на средноквадратичните отклонения от средното на извадката:

$$s'^2 = \frac{1}{n}\sum_{i=1}^{n}(\bar{x}_i - \tilde{x})^2 \qquad (5)$$

Сега ние искаме да проверим дали тя е неотместена и състоятелна оценка за дисперсията на самата популация.

Пресмятаме първо математическото очакване на тази оценка:

$$\begin{aligned} E(s'^2) &= \frac{1}{n}E\left\{\sum_{i=1}^{n}[\bar{x}_i - \tilde{x}]^2\right\} = \frac{1}{n}E\left\{\sum_{i=1}^{n}[\bar{x}_i - \hat{x} - (\tilde{x} - \hat{x})]^2\right\} \\ &= \frac{1}{n}E\left\{\sum_{i=1}^{n}[\bar{x}_i - \hat{x}]^2\right\} - \frac{2}{n}E\left\{\sum_{i=1}^{n}(\bar{x}_i - \hat{x})(\tilde{x} - \hat{x})\right\} + \frac{1}{n}E\left\{\sum_{i=1}^{n}[\tilde{x} - \hat{x}]^2\right\} \qquad (6) \\ &= \frac{1}{n}n\sigma^2(\bar{x}) - \frac{2}{n}E\left\{\sum_{i=1}^{n}(\bar{x}_i - \hat{x})\frac{1}{n}\sum_{j=1}^{n}[\bar{x}_j - \hat{x}]\right\} + \frac{1}{n}n\left[\frac{\sigma^2(\bar{x})}{n}\right] \end{aligned}$$

Тези преобразования се правят по аналогия с извършеното в предишната точка. Във втория член остават ненулеви само диагоналните членове $(i = j)$, тъй като отново недиагоналните са нули (ковариации на независими поради правилата на случайния избор величини).

Окончателно:

$$\sigma^2(s'^2) = \sigma^2(\bar{x}) - \frac{2}{n^2}n\sigma^2(\bar{x}) + \frac{\sigma^2(\bar{x})}{n} = \sigma^2(\bar{x})\left(\frac{n-1}{n}\right) \qquad (7)$$

Вижда се, че $E(s'^2) \neq \sigma^2(\bar{x})$, т.е. оценката (5) *не е неотместена* оценка за дисперсията на популацията (нейното отместване впрочем клони към нула при $n \to \infty$; такива оценки се наричат *<u>асимптотично неотместени</u>*). Същевременно става ясно, че следната оценка:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{x}_i - \tilde{x})^2 \qquad (8)$$

е неотместена оценка за дисперсията на популацията.

Може да се установи (по аналогичен начин), че (8) е и състоятелна оценка за дисперсията на популацията.

Множителят $(n-1)^{-1}$ в дефиницията (8) изглежда донякъде странен. За да се установи неговият смисъл, си представяме, че $n=1$. Тогава средното аритметично на извадката е равно на стойността на единствения й елемент, и според първоначалната дефиниция за дисперсията (5) тя би станала 0 (т.е. безкрайно точно измерване?), което очевидно не може да е вярно. Според неотместената оценка (8) в случая на извадка с обем единица дисперсията на популацията остава неопределена ($0/0$), т.е. тя не може да бъде оценена. Причината е, че информацията, съдържаща се в стойността на измерването $\bar{x}_1$ вече е използувана за оценка на $\hat{x}$ - средната стойност на разпределението на популацията.

Друга гледна точка за обяснението на този множител е, че при определянето на оценката (5), респ. (8), елементите $\{\bar{x}_1,...,\bar{x}_n\}$ не са вече всичките независими – между тях се е появила една връзка с определянето на средното им аритметично чрез (1). По такъв начин, ефективният брой на елементите в извадката се намалява с един. В такъв случай се казва, че броят на *степените на свобода* на оценката s^2 е равен на $n-1$.

Оценката (1) се нарича още *емпирично средно*, а (8) – *емпирична дисперсия на извадката*. Положителният квадратен корен от (8) се нарича *емпирично средноквадратично отклонение* (за извадката).

Може да се установи, че всяко допълнително уравнение (връзка) между $\{\bar{x}_i\}$ - елементите на извадката, намалява броя на степените й на свобода с единица. Ние ще използуваме този резултат по-нататък без доказателство.

VI. Метод на максималното правдоподобие

12. Функция на правдоподобие

Досега ние разглеждахме въпроса за оценка на параметри. Бяха посочени някои желателни свойства на оценките, но без да се каже как да се построят оценки с такива свойства в конкретните случаи. Оценки бяха получени само за такива (наистина много важни) величини като математическото очакване и дисперсията, но техният вид като че ли пада от небето. Сега ние ще разгледаме една по-обща задача:

Нека разглеждаме една n-мерна случайна величина: $x = (\overline{x}_1,...,\overline{x}_n)$; $\{\overline{x}_i, i = 1,...,n\}$ са нейните компоненти, n на брой, които са също така случайни величини. Популацията в този случай се състои от всевъзможните експерименти, при които се получава някаква конкретна, фиксирана стойност на x. (Например, x може да е цял спектър, т.е. последователност от еквидистантни измервания, при което i е номерът на съответния канал, а $\{\overline{x}_i, i = 1,...,n\}$ са случайни величини, описващи резултатите от измерванията в съответните канали).

Предполагаме още, че съвместното вероятностно разпределение на величините $\{\overline{x}_i, i = 1,...,n\}$ напълно се определя от параметрите $\lambda = (\lambda_1,...,\lambda_p)$, p на брой. Вероятностната плътност на разпределение на случайната величина x се записва така: $f(\overline{x};\lambda)$.

Нека сме извършили единичен експеримент, подчиняващ се на правилата за случаен избор (все едно, сме измерили един спектър, но с n компоненти). На този експеримент може да се съпостави едно число:

$$dP^{(j)} = f(x^{(j)};\lambda)dx \qquad (1)$$

Числото dP показва - *след като резултатът от експеримента е вече известен* - с каква вероятност този резултат е могъл да бъде очакван. То има характер на *апостериорна вероятност*. За извадка с обем N (независими елементи), тази вероятност се определя като произведение:

$$dP = \prod_{j=1}^{N} f(x^{(j)};\lambda)dx \qquad (2)$$

при това dP е функция на параметрите $\{\lambda\}$. Съществуват ситуации, в които е известно, че популацията може да се опише с един от два възможни и различни набора от параметри $\{\lambda_1\}$ и $\{\lambda_2\}$. Образуваме отношението:

$$Q = \frac{\prod_{j=1}^{N} f(x^{(j)};\lambda_1)}{\prod_{j=1}^{N} f(x^{(j)};\lambda_2)} \qquad (3)$$

Резултатът от един конкретен (случаен) избор (измерване) може да се изрази така: наборът от параметри $\{\lambda_1\}$ е Q пъти по-вероятен от набора $\{\lambda_2\}$. Изразът (3) се нарича ***отношение на правдоподобие***, а функцията:

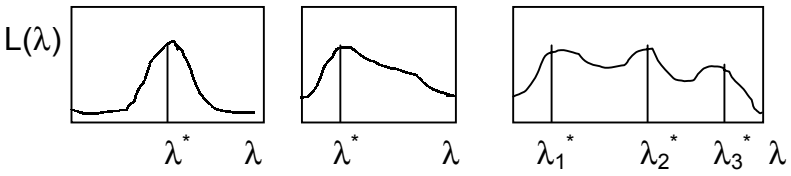
$$L = \prod_{j=1}^{N} f(x^{(j)};\lambda) \qquad (4)$$

се нарича ***функция на правдоподобие***.

Много важно е да се разбере разликата между така определената функция L и обикновената (априорна) вероятност. Именно, при априорната вероятност (*apriori* = преди експеримента) ние предполагаме, че имаме някакви зададени, фиксирани стойности на параметрите $\{\lambda\}$ и като знаем вероятностната плътност $f(x;\lambda)$, ние можем да определим (да предскажем) вероятността да се получи някакъв резултат (например, че $\bar{x} = x^*$), т.е. тогава λ са параметри, а x - аргументи. При апостериорната вероятност резултатите от измерването са *вече* известни (това са векторите $x^{(j)}$, които са фиксирани), и ние можем по тези резултати от експеримента да оценяваме величините λ, които са сега аргумент на функцията на правдоподобие L (а фиксираните $x^{(j)}$ в този случай са, напротив, параметри).

Разглеждайки конструкцията на понятието отношение на правдоподобие, лесно може да се стигне до заключението, че най-голямо доверие заслужават онези стойности на параметрите λ, при които се реализира максимум на функцията на правдоподобие L. В това собствено се състои и принципът на максималното правдоподобие (англ. maximum likelyhood principle).

На фигурата са показани някои най-характерни възможни профили на функцията на правдоподобие (за един параметър).



В първия случай ние сме в най-благоприятна ситуация (функцията на правдоподобие е симетрична камбановидна крива). Несъмнено, стойността λ^*, съответствуваща на максимума на $L(\lambda)$, може да се приеме като най-добра оценка на параметъра λ. Функцията на правдоподобие задава едно разпределение (тя е пропорционална на вероятностната плътност при съответната нормировка), чиято дисперсия е мярка за неопределеността на оценката λ^* за параметъра λ.

Във втория случай (на унимодално асиметрично разпределение) най-вероятната стойност λ^* все още е най-добра оценка за параметъра λ. Дисперсията обаче, въпреки че запазва своя математически смисъл, не е вече така удовлетворителна оценка за описание на поведението на функцията около нейния максимум. В този случай е по-добре заедно с λ^* да се пресмятат и по-висшите моменти или дори цялата съвместна вероятностна плътност L.

Може накрая да се случи и така, че $L(\lambda)$ да има няколко локални максимума. В този случай обикновено се отдава предпочитание на най-големия максимум на функцията на правдоподобие. Но пресмятанията, свързани с намирането на глобалния максимум, могат да бъдат много сложни (и продължителни), особено при многопараметрични задачи. Вероятността за появата на множество локални максимуми на $L(\lambda)$ корелира с възможността за получаване на нееднозначно решение на физическата задача, от която всъщност сме тръгнали. Това може да се случи най-често при:

- малък брой наблюдения;
- лоша статистическа точност на измерванията;
- недобре определени начални предположения за вида на зависимостта $f(x;\lambda)$, която е често само приблизително известна.

В такива случаи е необходимо, очевидно, крайно внимание при интерпретацията на резултатите.

13. Метод на максималното правдоподобие

За определянето на положението на максимума на $L(\lambda)$ могат да бъдат използувани стандартните методи на математическия анализ: да се диференцира спрямо λ, да се приравнят производните на нула и да се решат възникващите при това уравнения (или системи уравнения). Диференцирането обаче на произведения е трудоемко и затова се предпочита следният подход:

$$l = \ln L = \sum_{j=1}^{N} \ln f(x^{(j)}; \lambda) \qquad (5)$$

се въвежда като *логаритмична функция на правдоподобие.* Очевидно l и L имат максимуми при едни и същи стойности на λ, така че е достатъчно да диференцираме $l(\lambda)$, което е много по-лесно.

И така, трябва да се реши системата:

$$\frac{\partial l}{\partial \lambda_i} = \sum_{j=1}^{N} \frac{\partial}{\partial \lambda_i} \ln f(x^{(j)}; \lambda) = \sum_{j=1}^{N} \frac{\frac{\partial f(x^{(j)}; \lambda)}{\partial \lambda}}{f(x^{(j)}; \lambda)} = 0, \;\; i = 1,..., p \qquad (6)$$

Тази система от (изобщо нелинейни) алгебрични уравнения се нарича *система уравнения на правдоподобие*.

По този начин, чрез принципа на максималното правдоподобие, задачата за оценка на неизвестни стойности на физични параметри се свежда до една задача от алгебрата - решаване на системи от уравнения относно неизвестните $\{\lambda_i, i = 1,..., p\}$.

Пример: Разглеждаме важния случай за оценка на средната стойност при многократни измервания. Тези измервания (компонентите на извадката) могат да имат различна точност (напр. да са получени чрез различни методи, с различна апаратура и пр.). Предполагаме, че грешките в измерванията са разпределени нормално, т.е. всяко измерване е извадка с обем 1 от популация със средна стойност λ, (една и съща за всички измервания) и (различни) дисперсии σ_j^2. Тогава вероятностната плътност за $\bar{x}^{(j)}$ е:

$$f(x^{(j)}; \lambda) = \frac{1}{\sigma_j \sqrt{2\pi}} e^{-\frac{(x^{(j)} - \lambda)^2}{2\sigma_j^2}} \qquad (7)$$

Нека сме направили N измервания. Съответната функция на правдоподобие е:

$$L = \prod_{j=1}^{N} \frac{1}{\sigma_j \sqrt{2\pi}} e^{-\frac{(x^{(j)}-\lambda)^2}{2\sigma_j^2}} \tag{8}$$

а логаритмичната функция:

$$l = -\sum_{j=1}^{N} \frac{(x^{(j)} - \lambda)^2}{2\sigma^2} + const \tag{9}$$

Уравнението на правдоподобие е едно и има вида:

$$\frac{dl}{d\lambda} = \sum_{j=1}^{N} \frac{(x^{(j)} - \lambda)}{\sigma_j^2} = 0 \tag{10}$$

Неговото решение е:

$$\lambda^* = \frac{\sum_{j=1}^{N} \frac{x^{(j)}}{\sigma_j^2}}{\sum_{j=1}^{N} \frac{1}{\sigma_j^2}} \tag{11}$$

Следователно, резултатът, следващ от принципа на максималното правдоподобие, може да се изкаже така: най-правдоподобната оценка за средната стойност на популацията при многократни измервания е средното от отделните измервания, взети с тегла, обратно пропорционални на техните дисперсии. Забележете още, че в този конкретен пример функцията на правдоподобие е произведение от Гаусови функции с общ център; следователно ние се намираме в най-благоприятната ситуация от нарисуваните по-горе профили (симетрична камбановидна функция).

Важно е да се отбележи, че методът на максималното правдоподобие е конструктивен метод – той дава възможност да се построяват алгоритми за оценка на произволни параметри във всяка конкретна ситуация.

Методът на максималното правдоподобие (ММП) има и други положителни страни. Както казахме, съществува връзка между отместването на една оценка и нейната дисперсия (т.нар. *неравенство за информацията* или *неравенство на Крамер-Рао (Kramer-Rao)*). То поставя долна граница на дисперсията на една оценка. Очевидно, оценките, при които се постига тази долна граница, са за предпочитане пред останалите, тъй като те са най-ефективни (наричат се още *оценки с минимална дисперсия*). Доказва се, че когато обемът на извадката N клони към безкрайност, оценките, получени чрез метода на максималното правдоподобие (ММП) имат следните (асимптотични) свойства:

- те са неотместени;
- имат минимална дисперсия;

- функцията на правдоподобие е ***асимптотично нормална***.

Последното дава възможност да използуваме средноквадратичното отклонение на ММП-оценката $\sigma(\lambda^*)$ като оценка за вероятността истинската стойност на оценявания параметър λ да лежи в някакъв ***доверителен интервал***, напр.

$$P\{\lambda \in [\lambda^* - 1\sigma(\lambda^*),\ \lambda^* + 1\sigma(\lambda^*)]\} = 0.68,$$
$$P\{\lambda \in [\lambda^* - 2\sigma(\lambda^*),\ \lambda^* + 2\sigma(\lambda^*)]\} = 0.95$$

и пр., което, разбира се, представлява количествена мярка за увереността, с която можем да използуваме тези оценки.

VII. Проверка на статистически хипотези

14. Обща теория

При статистическия анализ на експериментални данни ние се стремим да оценяваме стойностите на параметри, които имат определен физически смисъл. Тези параметри не могат да се считат за съвършено неизвестни. Напротив, много често ние имаме конкретни очаквания за стойностите на тези параметри – въз основа на предварителни ориентировъчни експерименти, някакви модели или пък от теорията. Така че целта на измерванията, които провеждаме, е всъщност проверката на дадена *хипотеза*.

Статистическа хипотеза се нарича всяко твърдение относно статистическите свойства на изучаваните случайни величини.

В много случаи хипотезата, подлежаща на проверка, се състои от предположение за вероятностната плътност на разглежданата случайна величина. Една хипотеза се нарича *проста*, ако тя напълно характеризира вероятностната плътност, т.е. тя предполага напълно определени стойности за всички параметри, от които зависи разпределението на всички случайни величини в дадения експеримент. Хипотезата е *сложна*, ако общият функционален вид на вероятностното разпределение е известен, но точната стойност поне на един от неговите параметри не е определена.

Примери:

- проста хипотеза: $f(x)$ има стандартно нормално разпределение (тук всички параметри са определени – математическото очакване е 0, а дисперсията – 1);
- сложна хипотеза: $f(x)$ има нормално разпределение със средна стойност напр. 5.6 (тук дисперсията остава неуточнена).

Хипотезата, която се проверява в даден момент, се нарича *нулева*. Всяка друга хипотеза относно същия експеримент се нарича *алтернативна*.

Как се проверяват статистическите хипотези? При осъществяване на измерванията ние можем да получим оценки чрез извадката за един или друг параметър на вероятностното разпределение на популацията, който представлява интерес за нас. Тъй като всяка проста нулева хипотеза дава пълното вероятностно разпределение в пространството на извадките, то от нея може да се пресметне априорната вероятност за получаване на именно тази

стойност на оценката, която сме определили въз основа на извадката. Така на всеки резултат от измерването ние можем да съпоставим едно число – вероятността той да се осъществи, *ако нулевата хипотеза е вярна*. Ясно е, че ако тази вероятност е голяма, то и вероятността хипотезата ни да е правилна е също голяма, и обратно. Проблемът обаче се състои в това, че поради статистическия характер на измерванията са възможни по принцип всякакви резултати (с определени вероятности). Така че ние не можем да дадем прост отговор на простия въпрос: вярна ли е в крайна сметка проверяваната хипотеза или не?

Тъй като все пак е необходимо да вземем решение (което може да има много важни последици за по-нататъшните ни действия), ние постъпваме по следния начин:

- предполагаме, че нулевата хипотеза H_0 е вярна;
- предварително (преди експеримента) определяме едно (обикновено малко) число α, което се нарича ***ниво на значимост***;
- построяваме една ***критична област*** S от условието: $P(\bar{x} \in S \mid H_0) = \alpha$. Това е вероятността резултатът да попадне в критичната област S при условие, че H_0 е вярна (забележете, че критичната област S не се определя еднозначно от числото α);
- извършваме експеримента и проверяваме дали резултатът от измерването x^* попада в критичната област;
- Ако $x^* \in S$, ние ***отхвърляме хипотезата H_0 за нивото на значимост α***;
- в противния случай ние не можем да изкажем противоположното твърдение, т.е. не можем да приемем, че хипотезата H_0 е вярна. Ние можем да кажем само, че H_0 ***е съвместима с резултатите от нашия експеримент (при ниво на значимост α)***.

Нивото на значимост α е *мярка за риска* нулевата хипотеза H_0 да е вярна, макар че сме я отхвърлили. Такава грешка се нарича ***грешка от първи род***. Подобна грешка е неизбежна поради вероятностния характер на измерването (иначе трябва да положим $\alpha = 0$; тогава цялото пространство на извадките ще бъде извън критичната област и никакво решение относно валидността на нулевата хипотеза не може да бъде взето). Все пак ние имаме свободата да избираме нивото на значимост α. Този избор се прави в зависимост от

евентуалните последици от възможната грешка от първи род в конкретния случай - колкото те са по-значителни, толкова по-малко трябва да бъде α. В практиката на научните изследвания обикновено се използуват стойностите α = 0.05, 0.01, 0.001 и съществуващите таблици на статистическите разпределения са съобразени с тези стойности.

Има още една възможност да приемем грешно решение, именно да отхвърлим някоя вярна алтернативна хипотеза поради това, че x^* не попада в критичната област. Такава грешка се нарича ***грешка от втори род***. Вероятността за грешка от втори род зависи от конкретната алтернативна хипотеза, която е погрешно отхвърлена.

Близо до ума е, че един критерий за проверка на хипотези ще бъде особено полезен, ако при зададена вероятност α за грешка от първи род критичната област бъде определена така, че грешката от втори род да е минимална. Критерии, които отговарят на условието за минимизиране на вероятността за отхвърляне на *дадена* вярна проста алтернативна хипотеза при зададено ниво на значимост α се наричат ***най-мощни критерии***. Може да съществуват и ***равномерно най-мощни критерии***; това са такива критерии, които са най-мощни спрямо *всяка възможна* алтернативна хипотеза.

Директната проверка за принадлежност на резултата от експеримента към критичната област е неудобна в многомерния случай, тъй като тя може да има много сложна конфигурация (тя е област в многомерно пространство). Вместо това се построява някаква статистика T, която е функция на компонентите на извадката. Тогава проверката $x^* \in S$ се заменя с проверката: $T(x^*) \in U(x^*) = T(S(x^*))$. Последното равенство се отнася до образа на критичната област чрез статистиката T. Тази проверка е много по-лесна, защото се свежда обикновено до проверката за принадлежност на *едно* реално число $T(x^*)$ към някакъв интервал. Такава статистика се нарича ***статистика, на основата на която се строи критерият*** (за проверка на нулевата хипотеза). Разбира се, не всяка статистика T е подходяща за построяване на критерий за проверка.

Общата теория на критериите излиза извън рамките на този курс. Засега ще се ограничим с тези най-общи бележки и ще разгледаме няколко примера за най-често използувани конкретни критерии.

15. Критерий за принадлежност към стандартно нормално разпределение

Нулевата хипотеза е, че *измерваната случайна величина* $\bar{x}$ *има стандартно нормално разпределение*:

$$H_0: \ \bar{x} \in N(0,1) \qquad (1)$$

Нека смятаме, че сме направили извадка от n елемента. Интуицията ни подсказва, че за статистика, лежаща в основата на критерия за проверка на хипотезата (1) е добре да вземем средното на извадката $\tilde{x}$. Ако H_0 е вярна, $\tilde{x}$ трябва да има разпределение $N(0, 1/n)$.

Нека за конкретната извадка сме получили резултат $\tilde{x} = x^*$. Тъй като математическото очакване на разпределението $N(0, 1/n)$ е 0, то ясно е, че ако $|x^*| >> 0$, то проверяваната хипотеза едва ли е вярна.

На практика постъпваме по следния начин:

предполагаме, че (1) е вярна и избираме ниво на значимост α. От свойствата на нормалното разпределение ние знаем вероятността за получаване на резултат, по-голям от някое фиксирано x':

$$P(|\tilde{x}| > x' : H_0) = 2[1 - \psi_0(x'\sqrt{n})] \qquad (2)$$

където:

$$\psi_0(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} du \ \exp(-u^2/2) \qquad (3)$$

(2) е всъщност площта на областта $(-\infty, -x') \ V \ (+x', +\infty)$. И така, ние определяме границите $\pm x'_\alpha$ на критичната област от уравнението:

$$\alpha = 2[1 - \psi_0(x'_\alpha \sqrt{n})] \qquad (4)$$

Избраната от (4) критична област е симетрична; тя се състои от интервалите:

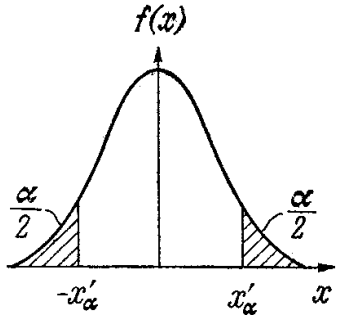
$$S: \ (-\infty, -x'_\alpha) \ V \ (x'_\alpha, \infty) \qquad (5)$$

Добре е да имаме предвид, че за стандартните стойности на нивото на значимост α е в сила:

$$\begin{aligned} 0.05 &= 2[1 - \psi_0(1.96)] \\ 0.01 &= 2[1 - \psi_0(2.58)] \\ 0.001 &= 2[1 - \psi_0(3.29)] \end{aligned} \qquad (6)$$

И така, ако например нашият резултат е $\tilde{x}^* \sqrt{n} = 2.1$, тогава заключението ни може да се изкаже така: *хипотезата (1) (че изходната случайна величина $\bar{x}$ се описва със стандартно нормално разпределение) се отхвърля за ниво на значимост $\alpha = 0.05$*. Ние обаче не можем да отхвърлим тази хипотеза (при този резултат), ако сме избрали ниво на значимост $\alpha = 0.01$ (тъй като тогава резултатът ни не попада в критичната област). Както вече казахме, не можем да изкажем и противоположното съждение. Тогава правилната формулировка на заключението ни би трябвало да бъде, че *резултатите от експеримента не противоречат на хипотезата за стандартно нормално разпределение при ниво на значимост $\alpha = 0.01$*.

За голямо съжаление, дори и резултатите ни да не могат да отхвърлят нулевата хипотеза даже и за много голямо ниво на значимост, това не ни позволява да кажем, че нашата хипотеза е вярна (вярна може да бъде всъщност някоя алтернативна хипотеза, която също не може да бъде отхвърлена при тези стойности на измерването). Такава ситуация е възможна например при измервания с ниска точност (с големи грешки), които са съвместими с много широко множество от различни хипотези. В крайната ситуация, един експеримент с напълно неопределен резултат е съвместим с всички възможни хипотези (но той не може да отдели нито една от тях спрямо останалите)!



В разглеждания случай критичната област изглежда така (защрихованата площ):

Всяка от защрихованите площи има лице $\alpha/2$. В случая имаме т.нар. *двустранен критерий*, тъй като критичната област (5) е несвързана. Понякога ни интересуват едностранни критерии, когато и знакът на отклонението е важен (например когато контролираме температурата в някаква система и желаем тя да не надминава някоя стойност). В такъв случай ние проверяваме отклоненията само в едната посока и гледаме дали се изпълнява неравенството:

$$P(\bar{x} \geq x'_\alpha) < \alpha \qquad (7)$$

Това неравенство определя друга критична област, която е само в дясната страна на Гаусовата крива и има площ α; тя съответствува на едностранен критерий за проверка на хипотезата. Ако резултатът попадне в такава критична област, отхвърля се хипотезата, че измерваната величина има нормално разпределение с неотрицателна средна стойност и дисперсия 1 за ниво на значимост α.

16. F-критерий за равенство на дисперсиите

Проблемът за сравняване на дисперсии възниква обикновено във връзка с повторни измервания на една и съща случайна величина, да речем с различни прибори или по различни методики. Друг пример е влошаването на точността на показанията на някакъв прибор с течение на времето поради различни причини. Може да възникне въпросът дали това се е случило. Това би могло да се установи, ако имаме две серии (извадки) от измервания, правени с този прибор в различни моменти от време на едни и същи еталонни образци. Тогава средните стойности са равни (ако няма систематични грешки), но дисперсиите са, изобщо казано, различни.

И така, предполагаме, че имаме две извадки от съвкупности с нормално разпределение, с обеми съответно N_1 и N_2. Нулевата хипотеза е: *дисперсията на двете серии измервания е еднаква*.

Статистиката, въз основа на която се строи критерият, се избира така:

$$Z = \frac{s_1^2}{s_2^2} \qquad (1)$$

където s_1^2 и s_2^2 са емпиричните дисперсии на извадките N_1 и N_2. Ясно е, че ако двете дисперсии са равни, Z ще е близко до 1. Известно е, че ако разпределението на популацията е нормално, статистиките:

$$X_1^2 = \frac{(N_1 - 1)s_1^2}{\sigma_1^2} \in \chi^2(f_1) \quad f_1 = N_1 - 1$$
$$X_2^2 = \frac{(N_2 - 1)s_2^2}{\sigma_2^2} \in \chi^2(f_2) \quad f_2 = N_2 - 1 \qquad (2)$$

имат разпределение χ^2 със съответния брой степени на свобода. Ако нулевата хипотеза е вярна, то $\sigma_1^2 = \sigma_2^2$ (това е всъщност дисперсията на популацията, която в този случай е една и съща). Тогава:

$$Z = \frac{f_2}{f_1}\frac{X_1^2}{X_2^2} \tag{3}$$

т.е. Z е отношение на две статистики с разпределение χ^2.

Вероятностната плътност на разпределението χ^2 е:

$$f(\chi^2) = \frac{1}{\Gamma\left(\frac{f}{2}\right)2^{\frac{f}{2}}}(\chi^2)^{\frac{1}{2}(f-2)}e^{-\frac{\chi^2}{2}} \tag{4}$$

Сега ние търсим вероятността:

$$V(Q) = P\left(\frac{X_1^2}{X_2^2} < Q\right) \tag{5}$$

Тя ще се получи с интегриране в триъгълната област:

$$V(Q) = \iint\limits_{\substack{x>0\\ y>0\\ x/y<Q}} dxdy\ f_{f_1}(x)\, f_{f_2}(y) \tag{6}$$

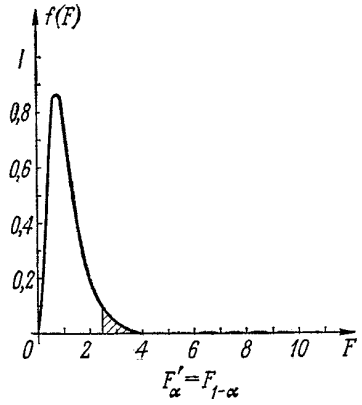
Този интеграл води до:

$$V(Q) = \int_0^Q dF\ f(F) \tag{7}$$

където за вероятностната плътност $f(F)$ имаме:

$$f(F) = \frac{\Gamma(f_1/2 + f_2/2)}{\Gamma(f_1/2)\Gamma(f_2/2)}\left(\frac{f_1}{f_2}F\right)^{f_1/2-1}\left(1+\frac{f_1}{f_2}F\right)^{-(f_1/2+f_2/2)} \tag{8}$$

Това разпределение се нарича ***разпределение на Фишер (Fisher)***. То има два параметъра – двете степени на свобода f_1 и f_2. То е донякъде подобно по форма на разпределението χ^2- има ненулева плътност само за $F > 0$. При $f_2 > 2$ математическото му очакване е: $E(Z) = f_1/f_2 - 2)$.



Обобщавайки, хипотезата за равенство на дисперсиите на две извадки се проверява така:

- избира се нивото на значимост α;
- определяме границата на критичната област (защрихованата на фигурата) от условието: $P(F > F'_\alpha) = \alpha$. Ясно е, че F'_α е квантилът на разпределението на Фишер на ниво $1-\alpha$;
- извършваме измерванията; определяме емпиричните дисперсии s_1^2, s_2^2 и пресмятаме Z от (1);
- Ако $Z > F'_\alpha$, то ние се намираме в критичната област и нулевата хипотеза се отхвърля за нивото на значимост α. Тогава можем да заключим, че $\sigma_1^2 > \sigma_2^2$, тъй като критерият тук е едностранен.

17. Т-критерий за сравняване на средни

17.1. Т-критерий за сравняване на средна стойност с константа

Първо ще разгледаме случая на една случайна величина $\bar{x}$ с нормално разпределение. Нулевата хипотеза е, че *математическото очакване на популацията е* $\hat{x} = \lambda$. Правим извадка с обем N; нека средното на извадката е $\tilde{x}$. Дисперсията на това средно е:

$$\sigma^2(\tilde{x}) = \sigma^2(\bar{x})/N \qquad (1)$$

Съгласно централната гранична теорема, за достатъчно голяма извадка средното $\tilde{x}$ клони към случайна величина с нормално разпределение $N(\hat{x},\sigma^2)$. Следователно, съответната нормирана случайната величина ще има стандартно нормално разпределение:

$$\bar{y} = \frac{(\tilde{x} - \hat{x})}{\sigma(\tilde{x})} \in N(0,1) \qquad (2)$$

С това задачата би се свела до познатата задача за проверка на хипотезата за принадлежност към стандартно нормално разпределение, ако $\sigma(\tilde{x})$ би била известна. Проблемът е, че обикновено $\sigma(\tilde{x})$ не е известна; вместо нея ние разполагаме само с една нейна оценка чрез извадката:

$$s^2(\bar{x}) = \frac{1}{N-1}\sum_{j=1}^{N}(\bar{x}_j - \tilde{x})^2 \qquad (3)$$

Съответната оценка за $\sigma(\tilde{x})$ ще бъде:

$$s^2(\tilde{x}) = \frac{s^2(\bar{x})}{N} = \frac{1}{N(N-1)}\sum_{j=1}^{N}(\bar{x}_j - \tilde{x})^2 \qquad (4)$$

Сега, ние се интересуваме от разпределението на случайната величина:

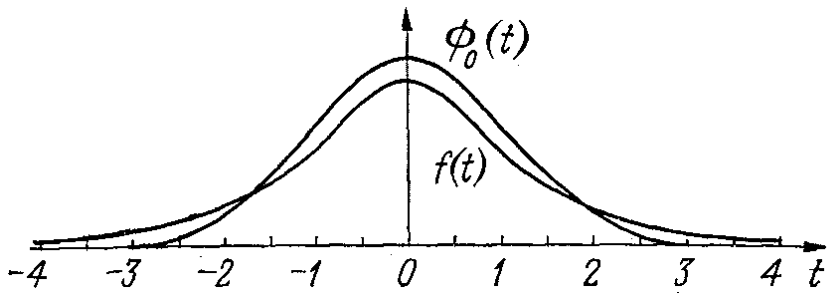
$$\bar{t} = \frac{(\tilde{x} - \hat{x})}{s(\tilde{x})} = \frac{(\tilde{x} - \hat{x})\sqrt{N}}{s(\bar{x})} \qquad (5)$$

Очевидно, това разпределение ще се отличава от разпределението на $\bar{y}$ (което е $N(0,1)$) поради замяната $\sigma(\tilde{x}) \to s(\tilde{x})$.

Без ограничение на общността можем временно да считаме, че $\hat{x} = 0$ (това води до проста транслация по абсцисната ос на търсеното разпределение). Тъй като $(N-1)s^2(\bar{x}) \equiv fs^2(\bar{x}) \in \chi_f^2$ ($f = N-1$ е броят на степените на свобода), то ясно е, че величината $\bar{t}$ представлява частно на две случайни величини: тази в числителя има нормално разпределение, а тази в знаменателя е квадратен корен от величина с разпределение χ^2. Вероятностната плътност на (5) при условие, че $\hat{x} = 0$ е:

$$f(t) = \frac{\Gamma\left(\frac{f+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\sqrt{\pi}\sqrt{f}}\left(1 + \frac{t^2}{f}\right)^{-\left(\frac{f+1}{2}\right)} \qquad (6)$$

Това разпределение се нарича *разпределение на Стюдънт (Student)*. То е симетрично (четна функция на t) и зависи от един параметър – броя на степените на свобода $f = N-1$. При $f \to \infty$ то клони към нормално. На фигурата е показано разпределението на Стюдънт $f(t)$ в сравнение с нормалното разпределение $\Phi_0(t)$.



Т-критерият (известен още като критерий на Стюдънт) за средната стойност на една случайна величина се прилага така:

- Приемаме за вярна нулевата хипотеза, че математическото очакване на популацията е $\hat{x} = \lambda$;
- Избираме ниво на значимост α;
- намираме границите на критичната област $\pm t'_\alpha$ (разпределението е симетрично и тук имаме двустранен критерий, така че те са две симетрични числа) от условието: $\int_{t'_\alpha}^{\infty} dtf(t) = \frac{\alpha}{2}$;
- правим измерванията, получаваме извадката и пресмятаме статистиката: $|t^*| = \frac{|\tilde{x} - \lambda|\sqrt{N}}{s(\bar{x})}$;
- ако $|t^*| > t'_\alpha$, то хипотезата $\hat{x} = \lambda$ се отхвърля за нивото на значимост α.

Този критерий може да се обобщи и да се използува за сравняването на средните стойности на две извадки.

17.2. Т-критерий за сравняване на средните на две извадки

Нека от две популации X и Y са направени две извадки с обеми N_1 и N_2. Хипотезата, която ще проверяваме, е, че $\hat{x} = \hat{y}$.

Средните по извадка имат приблизително нормално разпределение (съгласно централната гранична теорема). Техните дисперсии са:

$$\sigma^2(\tilde{x}) = \sigma^2(\bar{x})/N$$
$$\sigma^2(\tilde{y}) = \sigma^2(\bar{y})/N \qquad (1)$$

Оценките за дисперсиите от извадката са:

$$s^2(\tilde{x}) = \frac{1}{N_1(N_1-1)}\sum_{j=1}^{N}(\bar{x}_j - \tilde{x})^2$$
$$s^2(\tilde{y}) = \frac{1}{N_2(N_2-1)}\sum_{j=1}^{N}(\bar{y}_j - \tilde{y})^2 \qquad (2)$$

Знаем, че разликата:

$$\Delta = \tilde{x} - \tilde{y} \qquad (3)$$

е също приблизително нормално разпределена и оценката за нейната дисперсия е:

$$s^2(\Delta) = s^2(\tilde{x}) + s^2(\tilde{y}) \qquad (4)$$

Ако нашата хипотеза е вярна ($\hat{x} = \hat{y}$), то, очевидно, $\hat{\Delta} = 0$ и ще е в сила:

$$\bar{t} = \frac{\Delta}{\sigma(\Delta)} \in N(0,1) \qquad (5)$$

Тогава веднага може да пресметнем вероятността нулевата хипотеза да е вярна, ако $\sigma(\Delta)$ е известно. Ние обаче разполагаме само с оценката (4).

Обикновено се предполага, че X и Y са от една и съща генерална съвкупност (това във всеки случай е в съгласие с предположението ни за верността на нулевата хипотеза). Тогава:

$$\sigma^2(\tilde{x}) = \sigma^2(\tilde{y}) \qquad (6)$$

Следователно:

$$s^2 = \frac{(N_1-1)s^2(\bar{x}) + (N_2-1)s^2(\bar{y})}{(N_1-1)+(N_2-1)} \qquad (7)$$

което е претегленото средно на $s^2(\bar{x})$ и $s^2(\bar{y})$, е най-добрата оценка за дисперсията на популацията ((7) всъщност е равноправно сумиране на всички индивидуални дисперсии).

В такъв случай:

$$s^2(\Delta) = \frac{s^2}{N_1} + \frac{s^2}{N_2} = \frac{N_1+N_2}{N_1N_2}s^2, \qquad (8)$$

Доказва се, че величината:

$$\bar{t} = \frac{|\Delta|}{s(\Delta)} \qquad (9)$$

се подчинява на разпределението на Student с $f = N_1 - 1 + N_2 - 1$ степени на свобода.

Този критерий се прилага по-нататък напълно аналогично на критерия на Стюдънт от предния раздел.

18. Критерий за съгласие χ^2

Досега разглеждахме критерии за проверки на хипотези, които се отнасят до определянето на един или повече параметри на популацията. Такива критерии се наричат ***параметрични критерии***. Друг клас критерии сравнява директно вероятностната плътност на хипотетичното разпределение с това на резултатите от извадката. Това са така наречените ***критерии за съгласие***. Те са непараметрични критерии, тъй като се отнасят не до някой конкретен параметър, а до вероятностната плътност като цяло. Тук ще разгледаме най-важния от тях – критерия за съгласие χ^2.

Нека $f(x), F(x)$ са съответно вероятностната плътност и функцията на разпределение на популацията на изследваната случайна величина $\bar{x}$.

Цялата област на изменение на случайната величина $\bar{x}$ се разделя на r подинтервала: $\xi_1, \xi_2, ..., \xi_r$. Чрез интегриране на $f(x)$ в тези интервали се получава вероятността за попадане на резултатите от наблюденията на $\bar{x}$ в ξ_k:

$$p_k = P(\bar{x} \in \xi_k) = \int_{\xi_k} dx f(x), \quad \sum_{k=1}^{r} p_k = 1 \qquad (1)$$

Нека направим извадка с обем n и да я сортираме в интервалите $\xi_1, \xi_2, ..., \xi_r$. Нека n_k е броят на елементите на извадката, попадащи в интервала ξ_k. Ясно е, че: $\sum_{k=1}^{n} n_k = n$.

Нулевата ни хипотеза тук е, че *разпределението на случайната величина* $\bar{x}$ *е с вероятностна плътност* $f(x)$. Тогава трябва да очакваме да е изпълнено приблизителното равенство:

$$n_k \cong np_k, \quad k = 1, ..., r \qquad (2)$$

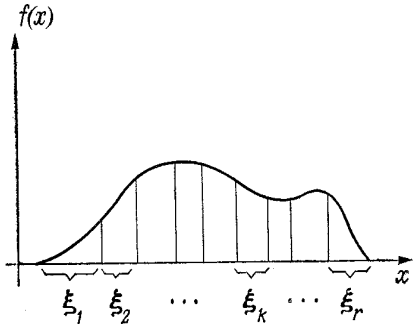
Като мярка за отклоненията построяваме статистиката:

$$X^2 = \sum_{k=1}^{r} \frac{(n_k - np_k)^2}{np_k} \qquad (3)$$

Статистиката X^2 се подчинява (асимптотично) на разпределение χ^2 с $r-1$ степени на свобода. Това твърдение няма да доказваме тук. Само ще отбележим, че намаляването на броя на степените на свобода с единица се дължи на ограничението:

$$\sum_{k=1}^{r} n_k - n\sum_{k=1}^{r} p_k = 0 \qquad (4)$$

Ние вече знаем, че такива ограничения намаляват броя на степените на свобода.



Прилагане на критерия:

- Избира се ниво на значимост α;
- Правят се n наблюдения: $x_1, x_2, ..., x_n$ на случайната величина $\bar{x}$. Разбиването на дефиниционната област на $\bar{x}$ на подинтервали $\{\xi\}$ се прави така, че:
  - ❖ броят им r да е достатъчно голям, така че да се описва добре профилът на разпределението;
  - ❖ броят на елементите във всяко ξ_k да е достатъчно голям, $n_k >> 1$, за да бъде X^2 близо до асимптотичното си разпределение;
- пресмята се статистиката X^2 от (3) и се сравнява с границата на критичната област χ'_α, която се определя от условието: $\int_{\chi'_\alpha}^{\infty} d\chi^2 f(\chi^2) = \alpha$ (за $r-1$ степени на свобода). Ако $X^2 > \chi'_\alpha$, хипотезата, че вероятностната плътност на $\bar{x}$ е $f(x)$, се отхвърля за нивото на значимост α.

Критерият за съгласие може да се обобщи за сравняване на вероятностните плътности на две случайни величини $\bar{x}$ и $\bar{y}$. В такъв случай нулевата ни хипотеза е, че *двете вероятности плътности са равни*: $f(x) = g(y)$. За проверка на тази хипотеза ние извършваме n_x наблюдения на величината $\bar{x}$ и n_y наблюдения на $\bar{y}$. След това разпределяме двете серии наблюдения *в едно и също множество интервали* $\{\xi\}$. Статистиката (3) се модифицира така:

$$X^2 = \sum_{k=1}^{r} \frac{[n(x)_k - n(y)_k]^2}{n(x)_k + n(y)_k} \qquad (5)$$

Величината X^2 се подчинява (асимптотично) на разпределение χ^2 с $r-1$ степени на свобода.

Казаното дотук е валидно при проста хипотеза, т.е. когато $f(x)$ е напълно известна. Може обаче да е известен общият вид на $f(x)$, но стойностите на някои параметри да не са известни, т.е. ние трябва да проверим една сложна хипотеза. Сега ние ще покажем как неизвестните стойности на параметрите на разпределението могат да се получат от наличните наблюдения, като използуваме метода на максималното правдоподобие:

Нека неизвестните параметри са $\lambda=(\lambda_1,\lambda_2,...,\lambda_p)$, $p<r$. Тогава вероятностите p_k са функции на λ: $p_k=p_k(\lambda)$. Функцията на правдоподобие за тази извадка е:

$$L(\lambda)=\frac{n!}{\prod_{k=1}^{r} n_k!}\prod_{k=1}^{r}[p_k(\lambda)]^{n_k} \qquad (5)$$

Този израз се получава направо от съвместната вероятностна плътност за наблюдаване на n_k елемента в интервала ξ_k, $k=1,...,r$ (това е всъщност известното полиномно разпределение).

Системата уравнения на правдоподобие е:

$$\frac{\partial l}{\partial \lambda_i}=\sum_{k=1}^{r}\frac{n_k}{p_k}\frac{\partial p_k}{\partial \lambda_i}=0,\quad i=1,...,p \qquad (6)$$

Решавайки (6), ние можем да определим най-правдоподобните оценки за стойностите на неизвестните параметри $\lambda=(\lambda_1,\lambda_2,...,\lambda_p)$.

Уравненията (6) налагат върху компонентите на X^2 още p на брой ограничения, които намаляват броя на степените на свобода. В такъв случай:

$$X^2=\sum_{k=1}^{r}\frac{(n_k-np_k(\lambda))^2}{np_k(\lambda)} \underset{n\to\infty}{\in} \chi^2_{r-p-1} \qquad (7)$$

VIII. Метод на най-малките квадрати

19. Общи бележки

Методът на най-малките квадрати (МНК, least-squares) е създаден от Льожандър (Legendre) и Гаус (1805-1809) първоначално въз основа на интуитивни съображения. В първоначалния си вид това правило е изглеждало така:

Резултатите от многократните наблюдения $\{y_j, j = 1,...,n\}$ на неизвестната величина x се разглеждат като сума от истинската величина x и някаква добавъчна грешка:

$$y_j = x + \varepsilon_j \quad j = 1,...,n \tag{1}$$

Правилото гласи, че *стойността на неизвестната величина x трябва да се определи от условието за минимум на сумата от квадратите на грешките*:

$$\sum_{j=1}^{n} \varepsilon_j^2 = \sum_{j=1}^{n} (y_j - x)^2 = \min_x \tag{2}$$

В много случаи това правило може да бъде получено от метода на максималното правдоподобие (ММП, който е развит доста по-късно), което вече служи за сериозна обосновка на МНК. Но дори и в случаите, когато МНК не може да се получи чрез ММП, МНК-резултатите притежават редица благоприятни свойства (в сравнение с други методи за оценка). Поради това МНК е най-често използуваният от всички статистически методи за обработка на данни и оценка на параметри.

Трябва да се разграничават следните случаи на употреба на МНК:

- при *преки наблюдения* - когато ние директно измерваме интересуващата ни величина;
- при *косвени наблюдения* - когато наблюдаваните величини са свързани с оценяваните параметри не пряко, а чрез някакви (известни или предполагаеми) функционални зависимости. С други думи, наблюдаваните величини са някакви функции - *линейни (ЛМНК, linear least-squares)* или *нелинейни (НМНК, non-linear least squares)* на търсените параметри;

- накрая, говорим за *наблюдения с ограничения* - когато между неизвестните параметри съществуват (или се предполагат) съотношения от определен функционален тип (пак се разграничават случаите на *линейни* и *нелинейни ограничения*);
- по-нататък, ограниченията могат да бъдат *строги (ограничения от тип равенства)* (напр. ограничението сумата от измеряемите ъгли на един триъгълник да е 180^0) или пък *ограничения от тип неравенства* (напр. площите на спектралните линии да са неотрицателни).

Ние ще разглеждаме последователно различните варианти на МНК, започвайки от по-простите.

20. Метод на най-малките квадрати при преки наблюдения

Този случай е най-прост, но ние го разглеждаме специално, за да въведем някои означения и да очертаем основната схема на МНК.

Нека са извършени n *независими* измервания на една (неизвестна) величина x. Измерените стойности $\{y_j, j=1,...,n\}$ съдържат грешки от измерванията:

$$y_j = x + \varepsilon_j \quad j = 1,...,n \tag{1}$$

За грешките ε_j предполагаме, че *са разпределени нормално около нулата*:

$$\varepsilon_j \in N(0, \sigma_j^2) \quad j = 1,...,n \tag{2}$$

Това предположение се оправдава в много случаи от централната гранична теорема.

Вероятността в резултат на измерването да получим стойност в интервала $(y_j, y_j + dy)$ е:

$$f_j dy = \frac{1}{\sigma_j \sqrt{2\pi}} \exp\left[-\frac{(y_j - x)^2}{2\sigma_j^2} \right] dy \tag{3}$$

Сега можем да конструираме функцията на правдоподобие и логаритмичната функция на правдоподобие. Те са, съответно:

$$L = \prod_{j=1}^{n} \frac{1}{\sigma_j \sqrt{2\pi}} \exp\left[-\frac{(y_j - x)^2}{2\sigma_j^2} \right] \tag{4}$$

$$l = \ln L = -\frac{1}{2}\sum_{j=1}^{n}\frac{(y_j - x)^2}{\sigma_j^2} + const(x) \qquad (5)$$

Следователно, условието за максимум на $l = \ln L$ съвпада с условието:

$$M = \sum_{j=1}^{n}\frac{(y_j - x)^2}{\sigma_j^2} \equiv \sum_{j=1}^{n}\frac{\varepsilon_j^{\,2}}{\sigma_j^2} = \min_x \qquad (6)$$

което е точно условието на МНК. Виждаме, че в този най-прост случай МНК се получава непосредствено от ММП.

В интерес на единството на по-нататъшните означения ще въведем величините:

$$g_j \overset{def}{=} \frac{1}{\sigma_j^2} \quad j = 1,...,n \qquad (7)$$

Те се наричат ***тегла*** на измерванията $\{y_j, j = 1,...,n\}$. При това получаваме:

$$M = \sum_{j=1}^{n} g_j \varepsilon_j^2 \qquad (8)$$

Вижда се, че теглото на всеки от членовете в сумата (8) е обратно пропорционално на дисперсията на съответното измерване; следователно, по-малко точните измервания участвуват с по-малко тегло при определянето на резултата x.

Самият резултат (т.е. оценката за стойността на неизвестната величина x), разбира се, може да се получи в този случай непосредствено чрез пресмятане на минимума на функционала M:

$$\frac{\partial M}{\partial x} = 2\sum_{j=1}^{n}\frac{(x - y_j)}{\sigma_j^2} = 2\sum_{j=1}^{n} g_j (x - y_j) = 0 \qquad (9)$$

Оттук:

$$\tilde{x} = \frac{\sum_{j=1}^{n} g_j y_j}{\sum_{j=1}^{n} g_j} \qquad (10)$$

това е претегленото средно аритметично на отделните измервания.

Ако всички дисперсии биха били равни $\{\sigma_j^2 = \sigma^2, j = 1,...,n\}$ (което би съответствувало точно на условията за случаен избор от една и съща популация), тогава $\tilde{x}$ би съвпаднало точно със средното аритметично на извадката. Изразът (10) е по-общ, доколкото той отчита евентуалната нееднаква точност на измерванията.

Използувайки известните ни свойства за дисперсията на сума от случайни величини, за дисперсията на $\widetilde{x}$ получаваме:

$$\sigma^2(\widetilde{x}) = \left(\sum_{j=1}^{n} \frac{1}{\sigma_j^2} \right)^{-1} = \left(\sum_{j=1}^{n} g_j \right)^{-1} \qquad (11)$$

(при равни дисперсии на измерванията получаваме известния ни резултат $\sigma^2(\widetilde{x}) = \frac{\sigma^2}{n}$).

Сега, тъй като грешките $\{\varepsilon_j\}$ са случайни величини, те могат да бъдат оценени чрез извадката $\{y_j\}$ в съответствие с вече полученото решение $\widetilde{x}$:

$$\widetilde{\varepsilon}_j = y_j - \widetilde{x}\ ,\ \ j = 1,...,n \qquad (12)$$

Оценките $\widetilde{\varepsilon}_j$ са също случайни величини и имат разпределение $\widetilde{\varepsilon}_j \in N(0, \sigma_j^2)$ (същото като това на $\widetilde{x}$). Следователно:

$$\frac{\widetilde{\varepsilon}_j}{\sigma_j} \in N(0,1) \qquad (13)$$

Но тогава ние знаем, че в точката на решението:

$$\widetilde{M} = \sum_{j=1}^{n} \frac{(y_j - \widetilde{x})^2}{\sigma_j^2} = \sum_{j=1}^{n} g_j (y_j - \widetilde{x})^2 \in \chi^2_{n-1} \qquad (14)$$

има χ^2 разпределение с $n-1$ степени на свобода (една степен се отнема от замяната $x \to \widetilde{x}$). *Това свойство на величината M може да се използува за проверка на първоначалната ни хипотеза*, а именно, че (припомняме):

- $\{y_j = x + \varepsilon_j \quad j = 1,...,n\}$ са наблюдения на неизвестната величина x;
- $\{\varepsilon_j \in N(0, \sigma_j^2) \quad j = 1,...,n\}$

И така, статистиката (14) може да се използува за проверка на горната хипотеза. Как ставаше това? Ако за избраното ниво на значимост α, $\widetilde{M} = \sum_{j=1}^{n} g_j \widetilde{\varepsilon}_j^2 > \chi^2_{1-\alpha}$, то ние се намираме в критичната област и трябва да отхвърлим хипотезата. Това означава или че (някои от) y_j не са наблюдения на неизвестната величина x, или че грешките $\{\varepsilon_j \notin N(0, \sigma_j^2) \quad j = 1,...,n\}$. Първата възможност може да се дължи на някакво грубо въздействие върху апаратурата (смущение, "изхвърчали" точки и др.), когато резултатите от измерването не може да се считат за наблюдения само на неизвестната величина x. Дори само едно $y_j \neq x + \varepsilon_j$ води до много голямо M, $M >> \chi^2$ и по-

нататъшната проверка става безсмислена. В такъв случай се постъпва така: взема се измерването с най-голямо отклонение ($\tilde{\varepsilon}_j = \max\{\tilde{\varepsilon}\}$) и му се приписва тегло нула, $g_j = 0$, т.е. това измерване се изключва от процеса на оценка. След това построяваме нова МНК-оценка за x, $\tilde{\tilde{x}}$. Ако за нея $M_{\tilde{\tilde{x}}}$ е значително по-малко, то трябва да изхвърлим измерването y_j и да продължим с останалите наблюдения.

Понякога обаче е възможно първият пункт от изказаната по-горе хипотеза да е в сила, но грешките от измерванията да не са разпределени нормално; в частност, наблюденията могат да имат отместване спрямо нулата (напр. "пълзене" на апаратурата и пр.). Това вече говори за *систематични грешки* в измерванията. В такъв случай оправдано от статистическа гледна точка е (ако не могат да се "набедят" само 1-2 измервания като неточни) да не се правят опити за МНК-оценки и позитивни заключения за резултатите от измерванията, докато не бъдат извършени по-нататъшни експерименти.

21. Метод на най-малките квадрати при косвени наблюдения

21.1. Линеен случай

Сега предполагаме, че имаме няколко неизвестни величини: $x = (x_1,...,x_r)$. Обикновено ние не измерваме пряко самите величини, които ни интересуват, а някакви техни функции. Означаваме тези функции с η и предполагаме засега, че те са линейни:

$$\eta_j + a_{j0} + a_{j1}x_1 + ... + a_{jr}x_r = 0 \quad j = 1,...,n \qquad (1)$$

В матричен вид системата (1) се записва така:

$$\eta + a_0 + Ax = 0 \qquad (2)$$

Тук η и a_0 са n-мерни вектори (стълбове), а $A - n \times r$ матрица.

Ние пак смятаме, че всяко измерване съдържа грешка и че $\{\varepsilon_j \in N(0,\sigma_j^2) \quad j = 1,...,n\}$. Самите резултати от измерванията означаваме с y:

$$y = \eta(x) + \varepsilon \qquad (3)$$

Предполагаме, че измерванията са *независими*. Следователно, ковариационната матрица на y (същата е и на ε) е диагонална и има вида:

$$C_y = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ \dots & \dots & \dots \\ 0 & 0 & \sigma_n^2 \end{pmatrix} \qquad (4)$$

Обратната матрица на C_y се нарича *тегловна матрица на измерванията*. В нашия случай (ур.(4)) тя се получава лесно:

$$G_y \overset{def}{=} C_y^{-1} = \begin{pmatrix} g_1 & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & g_n \end{pmatrix} \qquad (5)$$
$$g_j = 1/\sigma_j^2, \quad j = 1,\dots,n$$

Заместваме (3) в (2) и получаваме:

$$y - \varepsilon + a_0 + Ax = 0 \qquad (6)$$

Това е система линейни уравнения спрямо $x = (x_1,\dots,x_r)$. Ще търсим решение на тази система в духа на метода на максималното правдоподобие (тъй като тази система е преопределена при брой на неизвестните по-малък от този на измерванията, $r < n$, което е обичайният случай. Ако $r = n$, решението е единствено и няма какво повече да се прави). По предположение грешките ε_j (а от (3) това следва и за y_j) са разпределени нормално, т.е.:

$$f(y_j) = \frac{1}{\sigma_j\sqrt{2\pi}} \exp\left[-\frac{(y_j - \eta_j)^2}{2\sigma_j^2}\right] = \frac{1}{\sigma_j\sqrt{2\pi}} \exp\left[-\frac{\varepsilon_j^{\,2}}{2\sigma_j^2}\right] \quad j = 1,\dots,n \qquad (7)$$

Функцията на правдоподобие, респ. логаритмичната функция на правдоподобие ще бъдат:

$$L = \prod_{j=1}^{n} f(y_j) = (2\pi)^{-\frac{n}{2}} \left(\prod_{j=1}^{n} \sigma_j^{-1}\right) \exp\left(-\frac{1}{2}\sum_{j=1}^{n} \frac{\varepsilon_j^2}{2\sigma_j^2}\right)$$
$$l = \ln L = -\frac{n}{2}\ln(2\pi) - \sum_{j=1}^{n} \ln\sigma_j - \frac{1}{2}\sum_{j=1}^{n} \frac{\varepsilon_j^2}{\sigma_j^2} \qquad (8)$$

Очевидно, ММП изисква да търсим максимума на l, който съвпада с минимума на M, където:

$$M = \sum_{j=1}^{n} \frac{\varepsilon_j^2}{\sigma_j^2} = \sum_{j=1}^{n} \frac{(y_j + a_{jk}x_k + a_{j0})^2}{\sigma_j^2} = \min_{\{x_k\ k=1..r\}} \qquad (9)$$

(сумирането по k се подразбира). Или, в матричен вид M се записва така (този запис е валиден и когато измерванията y не са независими, т.е. тегловната матрица няма простия диагонален вид (5)):

$$M = \varepsilon^T G_y \varepsilon = (y + Ax + a_0)^T G_y (y + Ax + a_0) \qquad (10)$$

Минимумът на M се постига, когато едновременно са изпълнени условията:

$$\frac{\partial M}{\partial x_i} = 0 \quad i = 1,...,r \tag{11}$$

Това е равносилно на:

$$2A^T G_y (y + a_0 + Ax) = 0 \tag{12}$$

(последният резултат може да се провери чрез диференциране по компоненти). При $r \le n$ това матрично уравнение (еквивалентно на система от линейни уравнения) може да се реши спрямо x. За целта то се записва във вида:

$$(A^T G_y A)x = -A^T G_y (y + a_0) \tag{13}$$

Тогава решението му е:

$$\tilde{x} = -(A^T G_y A)^{-1} A^T G_y (y + a_0) \tag{14}$$

По такъв начин най-добрите ММП-оценки за неизвестните величини x се изразяват чрез измерванията y, тяхната ковариационна матрица $C_y = G_y^{-1}$ и известните коефициенти (A, a_0) на функциите $\eta(x)$.

За съгласуване на означенията с по-нататъшните раздели въвеждаме съкращението:

$$c \overset{def}{=} y + a_0 \tag{15}$$

което не зависи от x. Тогава решението (14) се записва така:

$$\boxed{\tilde{x} = -(A^T G_y A)^{-1} A^T G_y c} \tag{16}$$

Тук случаят на преки измервания се получава като частен случай.

Сега ще се спрем на въпроса за влиянието на грешките от измерванията върху оценките на неизвестните параметри x. Тъй като (16) дава линейна връзка между неизвестните x и измерванията y, ние можем да използуваме закона за разпространение на грешките за определяне на дисперсиите (и ковариациите) на x.

Нека $C_{\tilde{x}}$ е ковариационната матрица на x. Тогава имаме:

$$C_{\tilde{x}} = [(A^T G_y A)^{-1} A^T G_y] G_y^{-1} [(A^T G_y A)^{-1} A^T G_y]^T \tag{17}$$

За опростяване на израза вземаме предвид известното свойство на матриците $(AB)^T = B^T A^T$ и факта, че матриците G_y, G_y^{-1}, $(A^T G_y A)$ са симетрични. Резултатът е:

$$C_{\tilde{x}} = (A^T G_y A)^{-1} \qquad (18)$$

И тъй, получихме проста форма за ковариационната матрица на най-добрите МНК-оценки $\tilde{x}$ за неизвестните x. Квадратните корени от диагоналните елементи на (18) могат да се разглеждат като мерки за неопределеностите ("грешките") на неизвестните x, макар и те да не се измерват непосредствено в експеримента. Недиагоналните елементи са, разбира се, мярка за корелациите на x_i, x_j по двойки.

По-нататък, МНК-оценка за x, т.е. (16) може да се използува за определяне на най-добрата (в смисъл на МНК) оценка за вектора на грешките ε:

$$\tilde{\varepsilon} = A\tilde{x} + c \qquad (19)$$

Статистиката M:

$$M = \tilde{\varepsilon}^T G_y \tilde{\varepsilon} \qquad (20)$$

се подчинява на разпределение χ^2 с $n - r$ степени на свобода.

Пример: Полиномната апроксимация като линейна задача на МНК

Нека t е една независима променлива, а величините, които се измерват, са полиноми на t:

$$\eta = \sum_{i=0}^{r-1} x_i t^i \qquad (1)$$

Нека сме извършили n измервания на η за $t = t_j, \quad j = 1,...,n$. Това са:

$$y_j = \eta_j + \varepsilon_j \quad j = 1,...,n \qquad (2)$$

Предполагаме, че освен y ние знаем и тяхната ковариационна матрица C_y (най-често предполагаме, че y са независими; тогава е достатъчно да познаваме $\{\sigma^2(y_j) \quad j = 1,...,n\}$).

Системата линейни уравнения за x ще има вида:

$$\eta + a_0 + Ax = 0 \qquad (3)$$

Очевидно, сега ще имаме: $a_0 = 0$ и:

$$A = \begin{pmatrix} -1 & -t_1 & -t_1^2 & ... & -t_1^{r-1} \\ -1 & -t_2 & -t_2^2 & ... & -t_2^{r-1} \\ ... & ... & ... & ... & ... \\ -1 & -t_n & -t_n^2 & ... & -t_n^{r-1} \end{pmatrix} \qquad (4)$$

Това е задачата за прекарване на най-добрия полином от степен $r-1$ през n точки. Тази задача е най-популярното приложение на МНК.

21.2. Нелинеен случай

Досега предполагахме, че функционалните съотношения между измеряемите величини и неизвестните параметри са линейни. Сега ще се освободим от това ограничение и ще разгледаме общия случай.

И тъй, имаме:

$$f_j(x,\eta) = \eta_j - g_j(x) = 0 \quad j = 1,...,n \tag{1}$$

Идеята е, както много често се прави, нелинейният случай да се сведе към линеен чрез развитие на нелинейните функции в ред на Тейлър (Taylor) и ограничаването на това разложение до линейни членове включително.

Нека $x_0 = (x_{10},...,x_{r0})$ е някакво начално приближение към стойностите на неизвестните параметри x. Тогава:

$$f_j(x,\eta) = f_j(x_0,\eta) + \sum_{k=1}^{r} \left.\frac{\partial f_j}{\partial x_k}\right|_{x_0} (x_k - x_{k0}) + O(x_k - x_{k0})^2 \ , \ j = 1,...,n \tag{2}$$

Означаваме:

$$\xi = x - x_0$$

$$a_{jk} = \left.\frac{\partial f_j}{\partial x_k}\right|_{x_0} \tag{3}$$

$$A = \begin{pmatrix} a_{11} & ... & a_{1r} \\ ... & ... & ... \\ a_{n1} & ... & a_{nr} \end{pmatrix}$$

Предполагаме, както и преди, че за фактическите резултати от измерванията y е в сила:

$$y_j = \eta_j + \varepsilon_j \quad j = 1,...,n \tag{4}$$

Тогава:

$$0 = f(x,\eta) = f(x_0, y-\varepsilon) + A\xi = f(x_0, y) - \varepsilon + A\xi \tag{5}$$

Означаваме още:

$$c = f(x_0, y) \tag{6}$$

Следователно:

$$A\xi + c - \varepsilon = 0 \tag{7}$$

Системата линейни уравнения (7) е напълно аналогична на линейния случай. Решението й е:

$$\tilde{\xi} = -(A^T G_y A)^{-1} A^T G_y c \qquad (8)$$

Това обаче сега е МНК-оценката за ***вектора на корекциите*** ξ. Съответно, ковариационната матрица на $\tilde{\xi}$ ще бъде:

$$C_{\tilde{\xi}} = (A^T G_y A)^{-1} \qquad (9)$$

По същия начин (както за линейния случай) може да се получи оценка за уточнените грешки $\tilde{\varepsilon}$.

Тъй като началното приближение x_0 е фиксиран вектор, за вектора $x_1 = x_0 + \tilde{\xi}$ е в сила:

$$C_{x_1} = C_{\tilde{\xi}} = (A^T G_y A)^{-1} \qquad (10)$$

Сега вече можем да заменим x_0 с $x_1 = x_0 + \tilde{\xi}$, да развием $f(x,\eta)$ около x_1 и да повторим цялата процедура. Този итерационен процес може да се повтаря, докато поправките $\tilde{\xi}$ станат толкова малки, че уточненията за поредната стъпка да станат незначителни.

Тук не е показано, че такава процедура е сходяща и води към единствено решение. Нещо повече, това в общия случай не може и да се докаже (засега?) Сходимостта трябва да се проверява следователно във всеки конкретен случай.

Ясно е, че функциите η не трябва да се отличават силно от линейни функции в околността на точките $x_0, x_1, \ldots$ Или пък, алтернативно, ако η са силно нелинейни, началното приближение x_0 трябва да е достатъчно близо до решението $\tilde{x}$. Това е същността на изкуството за прилагане на МНК в нелинейния случай.

21.3. Алгоритъм на МНК за измервания без ограничения

Тук ще представим последователността на пресмятанията, необходими за получаването на решението на задачата за оценка на параметри по метода на най-малките квадрати за измервания без ограничения.

Дадени са:

- n измервания, $\{y_j,\ j = 1,\ldots,n\}$;
- тяхната ковариационна матрица C_y;

- n на брой измеряеми величини $\{\eta_j,\ j=1,...,n\}$, които са известни функции на (евентуално) някакви независими променливи и на неизвестните параметри $\{x_k, k=1,...,r\}$: $f(x,\eta)=0$

Дадено е също едно начално приближение за стойностите на неизвестните параметри: $x=x_0$ (r-мерен вектор).

Следва схемата на итерационния процес за решаване на МНК-проблема:

0. Означаваме номера на итерацията с s и полагаме $s=0$. Пресмятаме тегловната матрица на измерванията: $G_y=C_y^{-1}$;
1. Пресмятаме матрицата $A=\left(\frac{\partial f}{\partial x}\right)_{x=x_{s-1}}$ и вектора $c=f(x_{s-1},y)$;
2. Пресмятаме $C_{x_s}=(A^TG_yA)^{-1}$ - ковариационната матрица на неизвестните в това приближение;
3. $\widetilde{\xi}=-C_{x_s}A^TG_yc$, $\widetilde{\varepsilon}=A\widetilde{\xi}+c$ - вектора на корекциите и оценките за грешките на измерванията;
4. $x_s=x_{s-1}+\widetilde{\xi}$ - неизвестните параметри в това приближение;
5. $M_s=\widetilde{\varepsilon}^TG_y\widetilde{\varepsilon}$.
6. Ако е постигната сходимост (вж. по-долу), приключваме пресмятанията. Текущото решението се съдържа в x_s, C_{x_s}. В противен случай, увеличаваме номера на итерацията $s=s+1$ и се връщаме към точка 1.

Критериите за постигане на сходимост на итерационната процедура могат да са следните:

- $|M_s-M_{s-1}|<\delta M_s$: (в две последователни итерации стойностите на статистиката M се различават незначително); това е белег, че минимумът на M е практически достигнат;
- $\left\|\widetilde{\xi}\right\|<\delta$ (нормата на вектора на корекциите за дадена итерация е незначителна; тук нормата е мярка за дължината на вектора, напр. Евклидовата норма е: $\|a\|=\sqrt{\sum_{k=1}^{r}a_k^2}$); това означава, че корекциите вече са достатъчно малки.

Тези два критерия могат да се прилагат поотделно или в някаква комбинация помежду им. Освен това трябва да се предвидят и резервни изходи:

- изход при достигане на някакъв предварително определен максимален брой итерации, независимо от това дали е постигната сходимост;
- изход при разходимост ($M_s > M_{s-1}$), тъй като, както отбелязахме, МНК-процедурата не е гарантирано сходяща; това се случва нерядко при лоши начални приближения.

22. Свойства на решенията, получени чрез метода на най-малките квадрати

Досега ние разглеждахме (и получавахме) МНК-решения на задачата за статистическа оценка на параметри само като приложение на метода на максималното правдоподобие за решаването на линейни и линеаризуеми задачи. Ние считахме, че грешките на наблюденията са разпределени нормално. При това положение е лесно да се построи функцията на правдоподобие, а оттам изискването на МНК за минимум на сумата на квадратите на отклоненията се получава непосредствено и съвпада с резултата от максимизацията на функцията на правдоподобие.

Ако се освободим от предположението (което е най-спорният пункт), че резултатите от наблюденията са разпределени нормално около истинските стойности, то все още е възможно да се използуват съотношения от типа на МНК като "рецепти" за построяване на оценки на неизвестните параметри. Може да изглежда, че в такъв случай няма достатъчно основания от теоретичен характер за използуването на МНК. Съществува обаче една теорема (на Гаус-Марков), която установява, че дори и в такива случаи ($\varepsilon_j \notin N(0,\sigma_j^2)$) МНК-решенията притежават ред благоприятни свойства.

Първо ще дадем едно определение:

Съвкупността от предположения:

$$\begin{aligned} &y = Ax + \varepsilon \\ &\mathrm{cov}(\varepsilon) = \sigma^2 I \\ &\det(A^T A) \neq 0 \end{aligned} \qquad (1)$$

се нарича ***общ линеен модел*** (тук I е единичната матрица от ранг n).

Сега ще сумираме *свойствата на МНК-оценката* $\tilde{x}$ в рамките на общия линеен модел:

1. МНК-оценката $\tilde{x}$ е неотместена;
2. Ако $\lim_{n\to\infty}\frac{1}{n}(A^TA)$ е неизродена матрица, то $\tilde{x}$ е състоятелна оценка, а векторът $\sqrt{n}(\tilde{x}-x)$ е асимптотично нормален;
3. МНК-оценката $\tilde{x}$ притежава минимална дисперсия в класа на всички възможни неотместени линейни (спрямо y) оценки за параметрите x; при това векторите $\tilde{x}$ и $\tilde{\varepsilon}=y-A\tilde{x}$ са некорелирани (теорема на Гаус-Марков);
4. Математическото очакване на величината $M=\tilde{\varepsilon}^TC_y^{-1}\tilde{\varepsilon}$ е $E(M)=n-r$ (т.е. е равно на броя на степените на свобода на задачата).

Ако в допълнение на горните предположения векторът на грешките ε е с нормално разпределение ($\varepsilon\in N(0,\sigma^2)$), то МНК-решението има следните *допълнителни свойства*:

5. МНК-оценката $\tilde{x}$ съвпада с оценката по метода на максималното правдоподобие;
6. $\tilde{x}$ има нормално разпределение с математическо очакване равно на x и с ковариационна матрица: $C_x=(A^TG_yA)^{-1}$;
7. Случайната величина (статистиката) $M=\tilde{\varepsilon}^TC_y^{-1}\tilde{\varepsilon}$ има χ^2-разпределение с $n-r$ степени на свобода;
8. Оценките $\tilde{x}$ и $s^2=\frac{1}{n-r}\tilde{\varepsilon}^T\tilde{\varepsilon}$ са достатъчни оценки съответно за x и за σ^2 (т.е. за вектора на неизвестните параметри и за неговите дисперсии);
9. Векторите $\tilde{x}$ и $\tilde{\varepsilon}=y-A\tilde{x}$ са независими.

Очевидно, условието 9) е по-силно от условието 3b); също така условието 7) е по-силно от условието 4) - 4) се отнася само до математическото очакване на M, а 7) - до цялата вероятностна плътност на M.

Когато грешките ε на измерванията са разпределени нормално, ние можем да използуваме статистиката M за проверка на нулевата хипотеза по критерия χ^2. Именно, ако за едно избрано ниво на значимост α се случи

$M = \tilde{\varepsilon}^T C_y^{-1} \tilde{\varepsilon} > \chi^2_{1-\alpha}(n-r)$, то ние трябва да отхвърлим хипотезата, от която сме тръгнали (за даденото ниво на значимост). Освен от неизбежната възможност за грешка от I-ви род (т.е. отхвърляне на все пак вярна хипотеза), рискът за което е известен (той е точно равен на α), отказът ни от нулевата хипотеза може да бъде обусловен от някоя от следните причини:

a) предполагаемата функционална зависимост $f(x,\eta) = 0$ между измеряемите величини и неизвестните параметри не се реализира (т.е. говорим за *грешен модел, грешна моделна функция*). Или видът на тази функция не е правилно избран с оглед на конкретния експеримент, или някои константи в модела, които сме считали за точно известни, са зададени неточно. В такъв случай е необходимо моделът да се преразгледа (евентуално да се допълни), както и стойностите на константите (понякога положението се подобрява, ако те се разглеждат също като неизвестни параметри, подлежащи на оценка);

b) функционалната зависимост $f(x,\eta) = 0$ е правилна, но нейното Тейлърово разложение с точност само до линейни членове не е достатъчно, за да представи адекватно тази функция в разглежданата област от изменение на параметрите;

c) началното приближение x_0 е твърде далеч от истинското решение x. По-добро начално приближение може да доведе до по-приемлива стойност на M (ако сме попаднали на локален минимум). Очевидно, този пункт е тясно свързан с предишния, тъй като всяка развиваема в ред на Тейлър функция в достатъчно близка околност на центъра може да бъде представена с линейни членове с произволна точност. От тази гледна точка трябва да отбележим, че си струва да се пожертвуват усилия (и време) за търсене на добро начално приближение, тъй като иначе биха се употребили повече на брой итерации за достигане на сходимост. От друга страна, изчислителното време за една итерация е значително по-голямо от това за пресмятане на M при $x = x_0 = fix$;

d) Ковариационната матрица на измерванията $C_y = G_y^{-1}$, която често се основава само на груби оценки или даже на догадки, е невярна. В тази връзка едно сериозно подобрение на оценката на C_y може да се постигне чрез извършването на многократни измервания при същите

условия (т.е. да получим извадка) и да оценим C_y чрез емпиричните дисперсии и ковариации на основата на тази извадка.