Софийски Университет Св. Климент Охридски
Факултет по математика и информатика
Катедра "Вероятности и Статистика"

доц. ДИМИТЪР Л. ВЪНДЕВ

Записки
по
СТАТИСТИКА

СОФИЯ, 2000

Съдържание

Увод

Названието Статистика цели да обхване, както стандартно преподаваните в ФМИ методи на математическата статистика, така и да даде някаква представа на студентите на популярните статистически процедури и техните изчислителни варианти.

Цел на тези записки е да се даде едно допълнително пособие на студентите по математика и информатика, което да ги снабди със сведенията, отсъствуващи или трудно откриваеми в стандартните български учебници.

Първата лекция, която има обзорен характер за методите на статистиката, би могла чудесно да бъде допълнена с достъпното и леко за четене ръководство (Проданова 1998) по количествени методи.

Тези части, които са отразени в учебника (Димитров и Янев 1990) са дадени в максимално съкратен вид. В частност това е теорията на точковото оценяване. Полезен и достатъчно пълен набор от задачи по тези лекции има в (Н. Янев 1989).

Книгата (Уилкс 1967) ще си позволим да препоръчаме като най - представителен справочник по математическа статистика. Широк спектър от статистически методи е представен в (Дрейпер и Смит 1973). Материалът по дискриминантен и дисперсионен анализи е почти преписан от книжката (Вындев и Матеев 1988).

Основният материал е използуван във Факултета по Математика и Информатика при четене на курсове и за специалностите Информатика и Приложна математика.

Това е текущ вариант на записките. Той все още съдържа много непълноти и постоянно се променя.

Авторът е много благодарен на колегите си от катедра Вероятности и Статистика, които си направиха труда да прочетат внимателно първия вариант на записките и да отбележат многобройните грешки.

Тема 1

Статистика.

В тази лекция ще разгледаме основните методи на статистиката и ще се опитаме да намерим мястото на математиката в нея.

Понятието статистика е твърде широко. То включва в себе си както методи на просто преброяване и сгъстяване на информацията, така и методи за взимане на решения, основани на строги математически разсъждения. Да не говорим, че със същата дума "статистика" често означаваме и събраната информация - статистика за футбола, например.

1.1 Събиране на данни

Думата статистика произлиза от латинския корен stata означаващ държава. В частност, statist това е държавен служител. Събирането на данни за населението (с цел "осъвременяване" на данъците) е било важна държавна работа от както съществува държавата. Известни са такива записи за всички изследвани цивилизации в миналото. И сега всяка държава поддържа съответния орган, който е длъжен да я снабдява с такава информация. В България това е Централния статистически институт (ЦСИ), в САЩ - Central Statistical Office.

Основно понятие в статистиката е понятието "генерална съвкупност".

Определение 1.1 *Генерална съвкупност наричаме множеството от обекти на изследване.*

За едно изследване на ЦСИ това могат да бъдат:

- всички държавни учреждения в България;
- домакинствата в планинските райони;

- семействата без деца;
- всички жители на страната и т.н.

За един орнитолог, който също използува методите на статистиката, това е популацията от щъркели, например. За преподавателя по статистика - това могат да бъдат студентите от неговия курс.

1.1.1 Изчерпателни данни

Наричаме изчерпателни данни, които напълно описват дадено явление. Такива са например данните получени при едно преброяване на населението в ЦСИ. За геолога, интересуващ се от съдържанието на желязо в Кремиковското находище, това ще е самото находище разделено на някакви малки обеми.

За съжаление такива данни рядко са достъпни, пък и струват прекалено скъпо. Когато не е възможно такова изследване и данните за интересуващото ни явление не са достъпни. Така че генералната съвкупност става абстрактно множество от обекти представляващо цел на нашето изследване.

1.1.2 Извадки

В практиката често се работи с т.нар. извадка, част от генералната съвкупност. По този начин, търсените характеристики на генералната съвкупност се оценяват по данните от извадката.

Основна цел е по даден непълен обем данни да се направи някакво правдоподобно заключение за генералната съвкупност като цяло. Този набор от обекти, който всъщност се изследва (премерва, разпитва) се нарича извадка. Извадките биват систематични, случайни или подходящи за целите на изследването комбинации от двата метода.

Определение 1.2 *Извадка наричаме подмножеството от обекти на генералната съвкупност, достъпно за премерване.*

Например, една систематична извадка на дадено находище предполага сондажи разположени равномерно по площта му. От друга страна при случайната извадка се предполага, че шанса на всеки обект от генералната съвкупност да попадне в извадката е равен - всички обекти са равноправни и изборът е напълно случаен. Далеч не винаги е възможно да си избираме с кой от двата метода да конструираме извадката.

1.1.3 Планиране на експеримента

В селското стопанство и техниката често възниква задачата да максимизираме добива или оптимизираме даден производствен процес.

Това става с помощта на така наречения планиран (селскостопански) експеримент. Избираме няколко полета, засяваме ги с различни сортове пшеница и ги торим с различни видове тор. Така ще подберем подходящата за нашите цели комбинация (сорт и тор). Как обаче да избегнем влиянието на различните видове почва и може би природни условия? Как да намалим максимално броя на експерименталните полета за и без това скъпия и продължителен опит? На това ни учи планирането на експеримента. Математическа наука - част от математическата статистика.

Както се вижда, едва ли можем да гледаме на резултатите от такъв опит като на извадка от нещо.

1.1.4 Временни редове

Често нашите наблюдения са над някакво явление или процес, който се променя във времето. Това може да бъде курса на долара в поредни дни или средната температура на въздуха в София.

Наблюдението и тук не е извадка. Въпреки това, както ще видим по нататък, теорията на случайните процеси ни дава достатъчни математически средства да анализираме такива данни и правим (понякога разумни) прогнози.

1.2 Дескриптивна статистика

1.2.1 Числови и нечислови данни

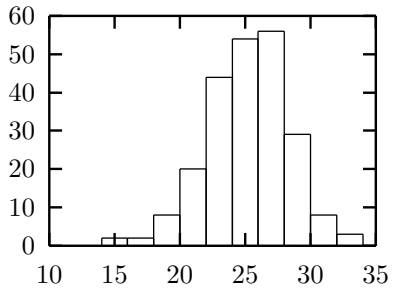
Информацията, която представляват данните обикновено се различава по това как се записва - понякога това са числа: размери, тегло, бройки и т.н. Друг път това са нечислови характеристики като цвят, форма, вид химическо вещество и т.н. Ясно е, че даже и да кодираме с числа подобни данни, при тяхното изучаване и представяне трябва да се отчита тяхната нечислова природа.

1.2.2 Графични методи

Представянето на данни всъщност е основна задача както на изчерпателната така и на извадъчната статистика. Информацията, която се

съдържа в милионите числа трябва да бъде представена в обозрима форма, така че всеки да си представи основните качества на множеството обекти. Главна роля в това кондензиране на информация има графичното представяне. То е ефектно и в минимална степен при него се губи информация.

Хистограмата е основният вид за представяне на информацията за наблюдения върху числов признак. Тя се строи по просто правило. Избират се обикновено еднакво големи е не много на брой (5 - 20) еднакво големи прилежащи интервала покриващи множеството от стойности на наблюдавания признак. Те се нанасят върху оста x. След това всеки от обектите на извадката се премерва и получената стойност попада в някой от интервалите.



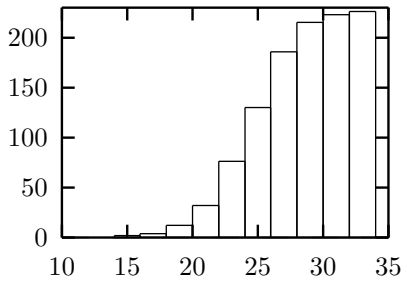
Фиг. 1.1: **Съдържания на апатит**

Ако интервалът $[x_{min},\ x_{max}]$ се раздели на k еднакви части с ширина h, т.е. $h\ = \frac{x_{max}\ -\ x_{min}}{k}$ и за всяко h се преброят попаданията на стойностите, то полученото число n се нарича честота на срещане. Последната, нормирана спрямо общият брой на данните N, е известна като относителна честота на срещане $f_i\ =\ \frac{n_i}{N}$, където с i е означен съответния интервал.

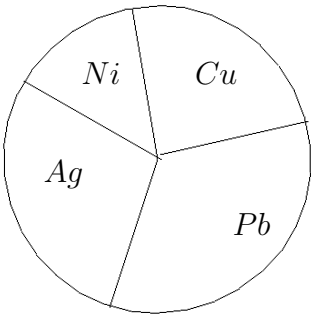
При графично маркиране на f_i с помощта на стълбчета, с височина стойността на f_i и ширина h, се получава хистограма, която служи за описание на изследваната съвкупност от данни (фиг.1.1).

Също така много удобна е така наречената кумулативна хистограма (фиг. 1.2). Тя се строи по натрупаните данни и позволява лесен отговор на въпроси от вида:
- каква е частта от наблюденията, попаднали под дадена граница;
- кое е числото под което са половината наблюдения – т.н. медиана.

Когато изследваме нечислови признаци, най - подходящото представяне е като процентно съдържание, например на гласовете подадени за различните партии в едно гласуване. Това може да се направи и с хис-



Фиг. 1.2: **Кумулативно представяне**



Фиг. 1.3: **Секторна диаграма**

тограма, но не е прието, тъй като разместването на стълбовете отговарящи на различните типове обекти променя общият вид на рисунката. Затова се използуват така наречените *секторни диаграми* или торти (piechart).

Отделните сектори отговарят по лице на пропорциите на различните типове и понякога са разноцветни.

1.2.3 Описателни статистики

Категорни данни

Нека отначало се занимаем с един нечислов признак - например пол. Ясно е, че цялата информация за пола в едно множество от n изследвани обекта е разделянето на обема това число на две слагаеми n_1 и n_2, съответно, броят на обектите от мъжки и женски пол.

Когато разгледаме нечислов признак на един случайно избран обект

от генералната съвкупност, то той съгласно предположенията ни за равнопоставеност на обектите в извадката би трябвало да попадне в дадена категория с вероятност равна на пропорцията на обектите в тази категория от генералната съвкупност. Ако съвкупността е голяма (или извадката ни е с връщане), то броят на обектите от извадката n_i с признак от категория i би трябвало да се окаже биномно разпределен с $B(n, p_i)$.

Количествени данни

За не много на брой количествени данни е прието да се използува така наречения вариационен ред. Освен това, той е много удобен и за теоретични изследвания, както ще видим по-нататък.

Определение 1.3 *Наредените по големина стойности на $x_1, x_2, \ldots, x_n$ се наричат вариационен ред $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, а елементите на реда — порядкови статистики.*

Така първата порядкова статистика $x_{(1)} = \min_I(x_i)$, а последната $x_{(n)} = \max_I(x_i)$. Интуитивно е ясно, че информацията за генералната съвкупност, която се съдържа в извадката, е представена изцяло във вариационния ред. Същата информация може да се представи и в следната форма.

Определение 1.4 *Извадъчна функция на разпределение наричаме функцията:*

$$F_n(x) = \begin{cases} 0 & x < x_{(1)} \\ \frac{k}{n} & x_{(k-1)} \leq x < x_{(k)} \\ 1 & x_{(n)} \leq x \end{cases}$$

В приложната статистика често се използуват следните дескриптивни (описателни) статистики:

- средна стойност: $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$.
- дисперсия: $D = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$.

Те лесно се изразяват чрез извадъчната функция на разпределение:

$$\overline{x} = \mu_1 = \int_{-\infty}^{\infty} x dF_n(x), \quad \mu_2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 = \int_{\infty}^{\infty} x^2 dF_n(x),$$

$$D = \mu_2(n) - \mu_1(n)^2.$$

Функциите μ_i наричаме извадъчни моменти. Извадъчните моменти μ_k са "състоятелни" оценки на моментите на сл.в. $\mathbf{E}\,\xi^k$. Същото твърдение важи и за други характеристики на извадъчното разпределение - квантили, медиана и т.н. Всички такива функции на извадъчното разпределение наричаме дескриптивни статистики. Например, порядковата статистика $x_{(k)}$ клони към квантила q_α, ако $k/n \to \alpha$.

Определение 1.5 *Медиана се определя като решение на уравнението:* $F(\mu) = 1/2$. *Медиана на извадка (извадъчна медиана) е наблюдението, което разделя вариационния ред на две равни части (когато обемът е четен се взима средното на двете централни наблюдения).*

Медианата описва положението на средата на разпределението върху числовата ос. В случая на големи отклонения от нормалност или при наличие на твърде отдалечени, съмнителни наблюдения, това е предпочитана оценка за "средата" на разпределението.

В много случаи се използува и положението на други характерни точки от разпределението.

Определение 1.6 *Извадъчен квантил* q_α *с ниво* α *на дадена извадка с ф.р.* F_n *се определя като приближено решение на уравнението:* $F_n(q_\alpha) = \alpha$.

Така медианата $\mu = q_{1/2}$.

1.2.4 Математическа таксономия

Математическа таксономия или многомерен анализ на данни се нарича серия от методи, даващи възможност да се описват задоволително големи по обем съвкупности от данни (както по брой на участвуващите обекти, така и по брой на измерваните параметри) Много от тези методи са емпирични, други са основани на математически или алгоритмични похвати. Целта им е да се даде представа за наблюдаваните множество числа (или кодове) във възможно най-сбита форма. Тук ще споменем следните методи:

- клъстерен анализ;
- многомерно скалиране;
- факторен анализ;
- логлинейни модели.

Обикновено тези методи се сумират под общото название *анализ на данни.*

1.3 Математическа статистика.

Основна цел на математическата статистика е изграждането на математически модели с помощта на теория на вероятностите. Освен това, тя дава средства за тяхната проверка върху реални данни, както и за интерпретация на резултатите от тях.

В математическата статистика винаги се разглежда следния вероятностен модел. За всяко наблюдение се предполага, че то е сл.в. Ако наблюденията са много, то обикновено те са независими сл.в. Когато независимостта е съмнителна, се предполага някакво съвместно разпределение на тези сл.в. Въз основа на така направените предположения се изследват вероятностните свойства на различни "полезни" функции от наблюденията - техните разпределения, моменти и т.н.

Когато в тези функции се поставят реалните наблюдения, се получават конкретни стойности - числа. Въз основа на тези числа се правят заключения - статистически изводи - за самия модел. Идеята на тези изводи е, че при верен модел ние знаем доколко те са "вероятни".

- какви са неговите параметри;
- доколко той е правдоподобен (адекватен);

Когато моделът се окаже неадекватен, се строи друг и т.н., както изобщо в науката математическо моделиране.

Тема 2

Прости непараметрични методи

Идеята на тази лекция е да илюстрира някои прости статистически разсъждения. Въз основа на данните ще правим заключения за неизвестните параметри (или други качества) на генералната съвкупност.

Основната идея на математическата статистика е разглеждане на наблюденията (и различни функции от тях) като сл.в. Това безусловно налага използването на чисто вероятностни методи на разсъждение.

Така при числови наблюдения се вижда, че вероятностен модел на вариационния ред е векторна случайна величина — функция от вектора $\xi_1, \xi_2, \ldots, \xi_n$, а извадъчната функция на разпределение става случайна функция.

В тази лекция ще разгледаме няколко примера на възможно най-прости вероятностни разсъждения в статистиката. Тези примери не се нуждаят от особено силни предположения и, съответно, не притежават други добри качества освен простотата си.

2.1 Тест на знаците

Нека са дадени две извадки от различни съвкупности с еднакъв обем $x_1, x_2, \ldots, x_n$ и $y_1, y_2, \ldots, y_n$. При това се предполага, че *наблюденията са сдвоени*, т.е. на всяко x_i съответствува y_i.

Такава ситуация възниква често в практиката. Например, когато мерим някаква характеристика върху едни и същи обекти преди и след въздействието с някакъв химикал или състоянието на болни преди и след лечението с определено лекарство. Често наричаме такива наблюдения повторни.

По-естествено е да се говори за една извадка от генерална съвкупност, на която всеки обект притежава два (или повече) параметъра от един и същи тип подлежащи на измерване.

Да си поставим задачата да отговорим на въпроса за наличието или не на съществена разлика между двете измервания (преди и след даването на лекарство, например). Да разгледаме статистиката (функция от наблюденията) $Z = \#\{i : y_i > x_i\}$) - броят на положителните разлики между наблюденията "след" и "преди". Да се опитаме да проверим хипотезата, че лекарството не оказва съществено влияние. Тогава за всеки случаен конкретно избран пациент вероятността неговото измерване y да е по-голямо от x би трябвало да бъде равна на $1/2$. Нека свържем с такова измерване сл.в. ξ приемаща стойности 1 (когато $y > x$) и 0 (в противен случай). Тъй като в математическата статистика се предполага, че извадката е от безкрайна съвкупност и резултатите от измерване на отделните обекти в извадката са независими, получаваме че статистиката Z е сума на n (броят на елементите в извадката) независими сл.в., т.е. има биномно разпределение $B(n, 1/2)$, ако хипотезата е верна.

Сега нека се спрем на целта на нашето лекарство - например, да повиши стойността на изследвания параметър. Ако то наистина действува, би трябвало $\mathsf{P}(\xi = 1) > 1/2$. Значи и в извадката би трябвало да има повече позитивни резултати - Z би трябвало да нарастне.

Следователно, критична за нашата хипотеза област ще бъде локализирана в дясната част на биномното разпределение:

$$W = \{Z_n : Z_n \geq i\}, \quad \mathsf{P}(W) = \sum_{k=i}^{n} b(n, k, 0.5) \leq \alpha.$$

При големи стойности на n се използува интегралната теорема на Муавър - Лаплас. Това ни дава лесна възможност да намерим необходимото i. Така, ако броят на наблюденията с положителен знак $Z > 0.5(n + 1.68\sqrt{(}n))]$, би трябвало да отхвърлим хипотезата, че в двете измервания няма разлика. Вероятността да сбъркаме при такова твърдение е малка - $\alpha = 0.05$.

2.2 Доверителен интервал за медиана

Нека си поставим за цел по n наблюдавани стойности да кажем нещо за неизвестната медиана μ на разпределението. Да означим с $\xi_{(1)} \leq \xi_{(2)} \leq \cdots \leq \xi_{(n)}$ наредените по големина стойности на наблюденията (сл.в.).

Теорема 2.1 *За всяко* $i < n/2$

$$\mathsf{P}(\xi_{(i)} \leq \mu \leq \xi_{(n-i+1)}) = 1 - 2(\frac{1}{2})^n \sum_{k=0}^{i-1} \binom{n}{k} \tag{2.1}$$

Доказателство: Имаме равенствата:

$$\begin{aligned} P(\xi_{(i)} \leq \mu \leq \xi_{(n-i+1)}) = &\quad 1 - P(\mu < \xi_{(i)}) - P(\xi_{(n-i+1)} < \mu) \\ P(\mu < \xi_{(i)}) = &\quad P(\xi_{(n-i+1)} < \mu) = (\tfrac{1}{2})^n \textstyle\sum_{k=0}^{i-1} \binom{n}{k}, \end{aligned}$$

от които следва търсената формула. Вторият ред е всъщност изразяване на вероятността като сума от Биномни вероятности. Наистина, при n-те експеримента по - малко от i са успешни, т.е. под медианата. □

Така като заместим във формулата (2.1) стойностите на наблюденията, ние получаваме *доверителен интервал* за неизвестната медиана. Вероятността в дясно се нарича *ниво на доверие*, например, 0.95. При големи стойности на n е затруднително пресмятането на суми от биномни коефициенти. Тогава се използува интегралната теорема на Муавър - Лаплас. Това ни дава лесна възможност да намерим необходимото i. Така при ниво на доверие 0.95 получаваме: $i = [.5(n - 1.96\sqrt{n})]$ Например, при $n = 100$ получаваме, че неизвестната медиана с вероятност 0.95 се намира между 40 и 61 членове на вариационния ред.

2.3 Тест на Ман-Уитни или Уилкоксън

Нека са дадени две независими извадки от различни съвкупности $x_1, x_2, \ldots, x_{n_x}$ и $y_1, y_2, \ldots, y_{n_y}$ възможно с различен обем. Проверяваме хипотезата, че двете съвкупности са еднакви — с еднакви медиани $H_0 : \mu_x = \mu_y$ — срещу алтернативата, че едната медиана е по - голяма от другата: $H_1 : \mu_x < \mu_y$.

Въвеждаме статистиката

$$U_x = \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \delta_{ij}, \tag{2.2}$$

където

$$\delta_{ij} = \begin{cases} 1 & x_i > y_j; \\ \frac{1}{2} & x_i = y_j; \\ 0 & x_i < y_j. \end{cases}$$

Аналогично се пресмята U_y, при това се оказва, че

$$U_x + U_y = n_1 n_2$$

. Когато искаме да проверим хипотезата H_0 очевидно доверителната област ще има вида:

$$P(U_{1-\alpha} \leq U_x) = 1 - \alpha.$$

При малки $min(n_x, n_y) < 20$ стойностите на $U_{1-\alpha}$ се вземат от специална таблица, а при големи се използува асимптотичното нормално разпределение на тази статистика:

$$\mathbf{E}\, U_x = \frac{n_x n_y}{2}, \qquad \mathbf{D}\,(U_x) = \frac{n_x n_y (n_x + n_y + 1)}{12}.$$

Тема 3

Методи на математическата статистика

В тази лекция ще се спрем само на модели за проста независима извадка, когато наблюденията се интерпретират като независими сл.в. с еднакво разпределение. В статистиката обикновено се предполага, че това разпределение е неизвестно.

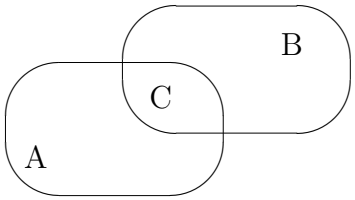
Когато разпределението е неизвестно с точност до определянето на някои параметри, методите на статистиката се наричат *параметрични*. В противен случай те са *непараметрични*.

Ще си поставим следните цели:

- да напомним идеята за статистическите изводи и хипотези;
- да поставим задачите на точковото оценяване;
- да уеднаквим понятията си за доверителен интервал и проверка на статистическа хипотеза;
- да напомним някои стари и дадем някои нови примери.

3.1 Статистически изводи и хипотези

Статистическите изводи са заключения за различни свойства на генералната съвкупност направени въз основа на наблюденията и различни предположения за генералната съвкупност. Така ако предположенията са верни, нашите твърдения стават функции на извадката, т.е. придобиват случаен характер — стават сл. в. Тъй като твърденията приемат две "стойности" — истина и неистина, задачата всъщност е да намерим вероятността едно заключение да бъде верно.



Фиг. 3.1: Лема на Нейман-Пирсън

Най-популярната и коректна форма за построяване на статистически извод е статистическата хипотеза. Много често имаме основания да предположим за неизвестното разпределение на генералната съвкупност, че то притежава плътност $f(x)$. Така е и по - лесно да построим "оптимална" критична област. За основен инструмент ни служи следната знаменита лема.

3.1.1 Лема на Нейман–Пирсън

Лема 3.1 *(Нейман–Пирсън) Нека са дадени две плътности $f_0(x), f_1(x)$. Тогава решението на разпределителната задача:*

$$\sup_W \int_W f_1(x)dx \quad \textit{при фиксирано} \quad \alpha = \int_W f_0(x)dx \tag{3.1}$$

се дава от условието $W = \{x : f_1(x) \geq cf_0(x)\}$ при подходящо подбрано c.

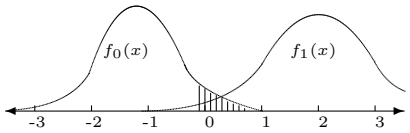
Доказателство: Нека $W = \{x : f_1(x) \geq cf_0(x)\}$ и $\alpha = \int_W f_0(x)dx$. Нека W' е такова, че $\alpha = \int_{W'} f_0(x)dx$. Да разгледаме разликата:

$$\begin{array}{ll} \int_W f_1(x)dx - \int_{W'} f_1(x)dx & = \int_A f_1(x)dx - \int_C f_1(x)dx \quad \geq \\ \int_A cf_0(x)dx - \int_C cf_0(x)dx & = c(\int_W f_0(x)dx - \int_{W'} f_0(x)dx) \quad = 0. \end{array}$$

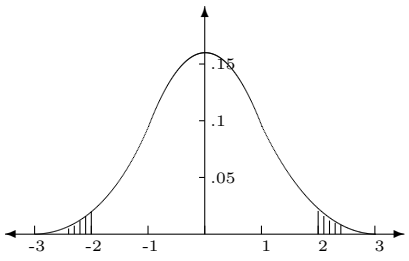
Тук сме означили $A = W \setminus W'$,$B = W' \setminus W$,$C = WW'$ или $W = A + C$,$W' = B + C$, както това е показано на фигурата.□

3.1.2 Критерий за проверка на хипотеза

Резултатът се използува по следния начин. Искаме да проверим хипотезата H_0, че наблюдението има плътност $f_0(x)$ срещу контра хипотезата



Фиг. 3.2: Едностранен критерий



Фиг. 3.3: Двустранен критерий

или алтернативата H_1, че то има плътност $f_1(x)$. Решението, което ще вземем съответно е, че хипотезата ни H_0 е вярна или не. Когато наблюдението попадне в критичната област W, отхвърляме хипотезата и, обратно, когато попадне извън нея, я приемаме. Естествено си задаваме критичното ниво $\alpha = \int_W f_0(x)dx$, което всъщност представлява вероятността да отхвърлим вярна хипотеза, като малко число – например 0.05.

Числото α наричаме *грешка от първи род*, а числото $1 - \alpha$ - *ниво на доверие*. На фиг. 3.2 защрихованата площ под кривата е равна на α. Тук алтернативата f_1 е отдясно на основното разпределение и критичната област е съответно в дясната част на основното разпределение f_0. Естествено, ако f_1 беше отляво, критичната област щеше да бъде на ляво.

Когато алтернативата е със значително по - голяма дисперсия, съгласно лемата на Нейман - Пирсън ще получим двустранна критична област. Същата област ще се получи и, когато "нямаме алтернатива".

Възможна е и обратната грешка β– *грешка от втори род* – да приемем хипотезата, когато тя не е вярна. Естествено е нашето желание да търсим критичната си област така, че запазвайки α да минимизираме

β. Лемата на Нейман - Пирсън ни дава средство лесно да строим *оптимални* критични области. Тя може да се използува и за произволни функции от наблюденията. Числото $1-\beta$ се нарича *мощност* на критерия (критичната област) и е различно за всяка конкретна алтернатива.

Пример 3.1 *Нека $H_0 : \xi \in N(0,1)$, а $H_1 : \xi \in N(1,1)$. Нека сме направили n наблюдения. Намерете оптималната критична област.*

Решение. Векторното наблюдение x ще има за плътности и при двете хипотези многомерната нормална плътност с единична ковариационна матрица, но различни средни стойности. От лемата 3.1 следва, че оптималната критична област има вида:

$$\begin{aligned} \sum (x_i - 0)^2 + c \quad &\geq \sum (x_i - 1)^2 \\ \bar{x} = \frac{1}{n} \sum x_i \qquad &\geq c. \end{aligned}$$

Определяме константата от уравнението $1 - \alpha = \Phi(c\sqrt{n})$. До същия извод щяхме да стигнем ако бяхме използували направо статистиката средна стойност и нейното разпределение. □

Пример 3.2 *Нека $H_0 : \xi \in N(0,1)$, а $H_1 : \xi \in N(0,\sigma^2)$, $\sigma > 1$ - неизвестен параметър. Нека сме направили n наблюдения. Намерете оптималния критерий за всяка от тези алтернативи.*

3.1.3 Равномерно най-мощен критерий

Когато нямаме възможност да изберем разумна проста алтернатива построяването на критерий (критична област) с максимална мощност е затруднително. В някои случаи, обаче, това става лесно. В пример 3.1 се вижда, че за всички алтернативи (със средна стойност по - висока от 0) решението ще бъде същото.

Определение 3.1 *Казваме, че критерият е равномерно най-мощен за дадено множество алтернативи, ако той е оптимален за всяка алтернатива поотделно.*

Така на фигура 3.2 е показан критерий, който е равномерно най - мощен за всички "десни" алтернативи.

Пример 3.3 *Нека $H_0 : \xi \in N(0,1)$, а $H_1 : \xi \in N(\theta,1)$, θ - неизвестен параметър с произволен знак. Нека сме направили n наблюдения. Не съществува равномерно най - мощен критерий за това множество алтернативи.*

Докажете го.

3.2 Доверителни области и интервали

От горните примери се вижда, че в крайна сметка и двата разгледани критерия се изразяват чрез функции от наблюденията на извадката — прието е всички такива функции да се наричат *статистики*. Много често имаме основания да предположим за неизвестното разпределение на генералната съвкупност, че то притежава плътност $f(x, \theta)$, зависеща от неизвестен параметър θ. Такава форма на представяне на нашите априорни познания ще наричаме *параметрична*.

Тогава възниква необходимостта да направим статистически изводи за този параметър. Едно естествено заключение за числов параметър би било твърдение за принадлежността на неизвестния параметър към някоя област. Наричаме такава област *доверителна*, а вероятността на твърдението *доверителна*. Ясно е, че колкото по - широка е областта, толкова по - вероятно е неизвестния параметър да попадне в него. Естествено би било да поискаме и тук някаква оптималност — например, областта да има минимален обем при фиксирана вероятност. Когато говорим за едномерен параметър, се интересуваме от доверителни интервали с минимална дължина.

В такава постановка задачата много прилича на лемата на Нейман - Пирсън. Първоначално ще построим доверителна област за наблюдението, така че тя да има минимален обем. В последствие (при подходящи условия) тя ще се превърне в доверителна област за параметъра.

Лема 3.2 *Нека е дадено семейството плътности $f(x, \theta)$. Тогава решението на разпределителната задача:*

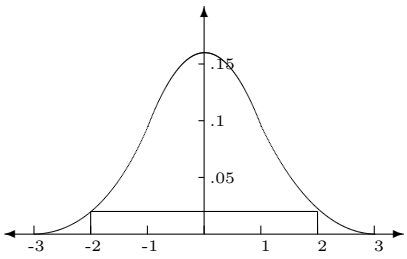
$$\inf_U \int_U dx \quad \textit{при фиксирано} \quad \alpha = \int_U f(x, \theta) dx \tag{3.2}$$

се дава от условието $U = \{x : f(x, \theta) \geq c\}$ при подходящо подбрано c.

Доказателство: Абсолютно същото като на оригиналната лема.□

Нека сега решаваме задачата в случая, когато $f(x, \theta) = f(x - \theta)$ — т.е. разпределението е известно с точност до неизвестен параметър на локация. От лемата следва, че в едномерния случай, когато имаме унимодално разпределение, трябва да построим доверителния интервал така, че плътността да бъде равна в двата края. Обикновено това е достатъчно за проверка на оптималността (минималната дължина) на така построения доверителен интервал.

Пример 3.4 *Нека $\xi \in N(\theta, 1)$. Нека сме направили n наблюдения. Намерете оптимална доверителна област за θ.*



Фиг. 3.4: Доверителен интервал

Решение 1. Векторното наблюдение x ще има за плътности многомерната нормална плътност $c_n e^{-\frac{1}{2}||x-\theta e||^2}$. От лемата 3.2 следва, че оптималната доверителна област за x има вида:

$$\sum_{i=1}^{n}(x_i-\theta)^2 \leq c.$$

Тъй като статистиката $\sum_{i=1}^{n}(x_i-\theta)^2$ има Хи-квадрат разпределение с n степени на свобода, определяме константата от уравнението $1-\alpha = \chi_n(c)$. Но при зададени наблюдения това е твърдение за θ. □Сега нека разгледаме внимателно това решение. Имаме равенството:

$$\sum_{i=1}^{n}(x_i-\theta)^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\theta)^2.$$

Така нашата доверителна област зависи главно от статистиката $\bar{x}$ и всъщност е симетричен около $\bar{x}$ интервал относно θ.

3.3 Статистики

Така във всички разгледани до сега примери ние стигнахме до изучаването на статистики, които са свързани с определени параметри на разпределението в генералната съвкупност. Като сл.в. те притежават разпределение и при правилни предположения могат да се смятат някои характеристики на тези разпределения. Тъй като в момента говорим за неизвестни параметри, естествено е да наречем статистиките *оценки* . За да избегнем безмислената оценка θ, ще разглеждаме като оценки на неизвестния параметър само такива функции на наблюденията, в аналитичния израз на които не участвува този неизвестен параметър.

Ще се върнем отново към пример 3.4.

Решение 2. Статистиката $\bar{x}$ има разпределение $N(\theta, \frac{1}{n})$. Доверителна област за $\bar{x}$ може да бъде

$$|\bar{x} - \theta| < z/(n)^{1/2}$$

Тук z се определя от уравнението $\Phi(x) - \Phi(-x) = 1 - \alpha$ и се нарича *двустранен квантил* на нормалното разпределение за критично ниво α. Така построения доверителен интервал удовлетворява равенството: $\phi(x) = \phi(-x)$, което следва от лема 3.2 и е с минимална дължина. □

Двете решения, които предложихме, са очевидно различни. Кое от тях е по - добро и как да търсим възможно най - добрите оценки и строим най - правдоподобни твърдения ни учи т.н. теория на оценяване на параметри, която ще разгледаме в следващите лекции.

Да разгледаме още един пример.

Пример 3.5 *Оценка на радиоактивността. Нека ξ е експоненциално разпределена. Нека сме направили n наблюдения. Намерете горна доверителна граница за неизвестния параметър λ с ниво на доверие γ.*

Решение. Да разгледаме статистиката $T = \sum_{i=1}^{n} \xi_i$. Тя има т.н. разпределение на Ерланг:

$$f(x) = \frac{\lambda^n x^n e^{-\lambda x}}{(n-1)!}, \tag{3.3}$$

което е частен случай на Гама разпределението. Следователно, статистиката 2λ ще има χ^2 разпределение с $2n$ степени на свобода. Значи

$$\gamma = \mathsf{P}(2\lambda T < q_\gamma) = \mathsf{P}(\lambda < \frac{q_\gamma}{2T}).$$

□

Тема 4

Оценяване на параметри

Тук ще дадем само някои елементи на теорията на точковите оценки. Ще определим някои техни приятни качества – неизместеност, ефективност и състоятелност. Ще покажем, че неизместените оценки с минимална дисперсия са единствени. Ще докажем знаменитото неравенство на Рао - Крамер.

В цялата лекция ще предполагаме, че наблюденията са извършени над сл.в. с плътност на разпределение $f(x, \theta)$ известна с точност до неизвестен параметър θ.

4.1 Определения

Определение 4.1 *Казваме, че статистиката $\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$ е оценка на параметъра θ, ако $\hat{\theta}$ не зависи от стойността на параметъра.*

Определение 4.2 *Казваме, че оценката $\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$ на параметъра θ е неизместена, ако $\mathbf{E}\hat{\theta} = \theta$.*

Разбира се, в това определение се счита, че математическото очакване се смята при стойност на неизвестния параметър точно равна на θ. Да разгледаме статистиката $\bar{x}$. Тя очевидно е неизместена оценка на математическото очакване при произволно разпределение на генералната съвкупност.

Определение 4.3 *Казваме, че оценката $\hat{\theta}$ на параметъра θ е ефективна, ако е с минимална дисперсия сред всички неизместени оценки на този параметър.*

Определение 4.4 *Казваме, че редицата от статистики $\hat{\theta}_n$ е състоятелна оценка на параметъра θ, ако $\hat{\theta}_n \xrightarrow{p} \theta$ при увеличаване на броя n на наблюденията.*

Съществува и по - силен вариант, строга състоятелност, където сходимостта е п.с.

4.2 Неизместени оценки с минимална дисперсия (н.о.м.д.)

В тази секция ще докажем две теореми за неизместените оценки, които отразяват тяхното значение.

Теорема 4.1 *(Рао - Блекуел) Неизместената оценка с минимална дисперсия (н.о.м.д.) е единствена (п.с.).*

Доказателство: Следва лесно от свойствата на проекцията. Достатъчно е да определим върху всички оценки Хилбертово пространство със скаларно произведение $(U, V) = \mathbf{E}\, UV$ и да разгледаме афинното подпространство на неизместените оценки: $\mathbf{E}\, V = \theta$. Нека сега V е н.о.м.д. Да допуснем че съществува друга неизместена оценка U. Нека $H = U - V$. Тогава

$$\mathbf{E}\,(V + \lambda H) = \theta,$$

т.е. оценката $V + \lambda H$ е неизместена.

$$||V + \lambda H - \theta E||^2 = ||V - \theta E||^2 + 2\lambda \mathbf{E}\, H(V - \theta E) + ||H||^2\lambda^2.$$

Тъй като V е н.о.м.д., горната квадратична функция на λ трябва да има минимум при $\lambda = 0$. Ако V е също с минимална дисперсия, то тогава $||H|| = 0$ и двете оценки съвпадат п.с. □

4.3 Неравенство на Рао - Крамер

Тук ще предполагаме, че наблюдаваната сл.в. притежава плътност и ще докажем едно знаменито неравенство.

Определение 4.5 *Наричаме функция на правдоподобие $f(x, \theta)$ плътността на наблюдаваната сл.в. ξ, когато тя зависи от неизвестен параметър.*

Теорема 4.2 *(Рао - Крамер) Ако θ е едномерен параметър и*

1. $f(x,\theta) > 0, \quad x \in X;$
2. $f(x,\theta)$ *притежава производни по* $\theta, \quad x \in X;$
3. *съществува* $\mathbf{E}\,((\frac{\partial \log f}{\partial \theta})^2) < \infty$
4. $\hat{\theta}$ *е неизместена оценка на* θ, *такава че* $\mathbf{E}\,\hat{\theta}^2 < \infty,$

то е валидно следното неравенство:

$$\mathbf{D}(\hat{\theta}) \geq \frac{1}{\mathbf{E}\,((\frac{\partial \log f(x,\theta)}{\partial \theta})^2)}. \tag{4.1}$$

При това, равенство се достига само ако

$$\frac{\partial \log f(x,\theta)}{\partial \theta} = k(\theta)(\hat{\theta} - \theta). \tag{4.2}$$

Доказателство: В следващите сметки ще използуваме равенството:

$$\frac{\partial \log f}{\partial \theta} = \frac{\partial f}{\partial \theta}/f,$$

което е изпълнено винаги, когато функцията f е положителна. Първо да покажем, че съществува следният интеграл:

$$\int_X \hat{\theta}\frac{\partial f(x,\theta)}{\partial \theta}dx = \int_X \hat{\theta}\sqrt{f(x,\theta)}\frac{\partial \log f(x,\theta)}{\partial \theta}\sqrt{f(x,\theta)}dx \leq$$
$$(\mathbf{E}\,|\hat{\theta}^2|)^{\frac{1}{2}}(\mathbf{E}\,((\frac{\partial \log f}{\partial \theta})^2))^{\frac{1}{2}} < \infty.$$

Тогава можем да диференцираме по θ двата интеграла:

$$\frac{\partial}{\partial \theta}\int_X f(x,\theta)dx = \int_X \frac{\partial f(x,\theta)}{\partial \theta}dx = \mathbf{E}\,\frac{\partial \log f(x,\theta)}{\partial \theta} = 0. \tag{4.3}$$

$$\frac{\partial}{\partial \theta}\mathbf{E}(\hat{\theta}) = \frac{\partial}{\partial \theta}\int_X \hat{\theta}(x)f(x,\theta)dx = \int_X \hat{\theta}\frac{\partial f(x,\theta)}{\partial \theta}dx = \mathbf{E}\,\hat{\theta}\frac{\partial \log f(x,\theta)}{\partial \theta} = 1.$$

Тогава имаме:

$$1 = \int_X (\hat{\theta} - \theta)\frac{\partial \log f(x,\theta)}{\partial \theta}f(x,\theta)dx = \mathbf{E}\,(\hat{\theta} - \theta)\frac{\partial \log f(x,\theta)}{\partial \theta} \leq$$

$$\mathbf{D}(\hat{\theta})^{\frac{1}{2}}(\mathbf{E}\,(\frac{\partial \log f}{\partial \theta})^2)^{\frac{1}{2}}.$$

Неравенството следва веднага след вдигане на квадрат. Обратно, когато се достига равенство двете подинтегрални функции трябва да са пропорционални. Това значи, че съществува константа по x (може би зависеща от θ) $k(\theta)$, така че е изпълнено равенство (4.2). □

Горните две теореми дават лесно средство за проверка на ефективността на оценките - достатъчно е да се достигне равенство в неравенството на Рао - Крамер,т.е. да бъде изпълнено равенството (4.2).

Пример 4.1 *Докажете, че в случая с пример (3.4) оценката $\bar{x}$ е ефективна и строго състоятелна оценка на математическото очакване.*

Тема 5

Методи за построяване на оценки

В тази лекция ще продължим разглеждането на основните методи на математическата статистика.

- ще разгледаме основните методи за конструиране на точкови оценки: метода на моментите и метода на максимално правдоподобие;
- ще изведем някои от свойствата на максимално правдоподобните оценки;
- ще дадем идея за метода на конструиране на критерии с отношение на правдоподобия;
- да напомним някои стари и дадем някои нови примери.

5.1 Метод на моментите

Извадъчната функция на разпределение $F_n(x)$ е сума от независими сл.в. и има просто разпределение при фиксирана стойност на x. Ако x е точка на непрекъснатост на $F(x)$, то $nF_n(x)$ е биномна сл.в., т.е. $nF_n(x) \in B(n, F(x))$. Имаме $\mathbf{E}\, F_n(x) = F(x)$, $\mathbf{D}F_n(x) = F(x)(1-F(x))/n$. Следователно, $F_n(x)$ клони п.с. към теоретичната $F(x)$ за всяко фиксирано x. Вярно е обаче по - силното твърдение на Гливенко – Кантели (теорема 13.1):

Теорема 5.1

$$\mathbf{P}(\lim_{n\to\infty} \sup_x |F_n(x) - F(x)| = 0) = 1. \tag{5.1}$$

Веднага от тази теорема следва, че всички линейни комбинации от порядкови статистики, зададени във формата

$$\int f(x)dF_n(x) \longrightarrow \int f(x)dF(x), \tag{5.2}$$

са (силно) състоятелни оценки на съответните параметри. Стига, разбира се, интегралът от дясно да съществува — това е така, например, за всяка ограничена непрекъсната функция.

Пример 5.1 *Оценка на моментите.*

В частност,

$$\mu_1(n) = \bar{x} = \int_{-\infty}^{\infty} xdF_n(x), \quad \mu_2(n) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \int_{\infty}^{\infty} x^2 dF_n(x).$$

Тези функции (дескриптивни статистики) наричаме извадъчни моменти. Следователно, извадъчните моменти $\mu_k(n)$ са състоятелни оценки на моментите на сл.в.

Нека разгледаме сл.в., за която съществува $E|x|^r < \infty$. Нека неизвестния параметър $\theta = f(\mathbf{E}\,\xi, \mathbf{E}\,\xi^2, ..., \mathbf{E}\,\xi^r)$, където $f(.)$ е непрекъсната функция. Тогава

$$F(\mu_1(n), \mu_2(n), \ldots, \mu_r(n)) \longrightarrow F(\mu_1, \mu_2, \ldots, \mu_r).$$

5.2 Метод на максималното правдоподобие

Това е най - популярния метод за конструиране на точкови оценки в теоретичната статистика. Неговата популярност се дължи на две неща:

1. изключително стройна и завършена теория;

2. добри асимптотични качества на построените оценки.

Нека предположим, че разпределението в генералната съвкупност има плътност $f(x, \theta)$ (по отношение на Лебеговата мярка) известна с точност до неизвестен едномерен или многомерен параметър $\theta \in \Theta$. Тогава извадката $\{\xi_1, \xi_2, ..., \xi_n\}$ като вектор от независими сл.в. ще има плътност в извадъчното пространство R^n от вида $L_n(\vec{x}, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$, която наричаме *функция на правдоподобие.*

Определение 5.1 *Казваме, че оценката $\hat{\theta}(x)$ удовлетворява принципа на максимално правдоподобие, ако*

$$L_n(\vec{x}, \hat{\theta}(x)) = \max_{\theta} L_n(\vec{x}, \theta) \tag{5.3}$$

за почти всяко x по мерките определени от плътностите $L_n(\vec{x}, \theta), \theta \in \Theta$.

Всъщност достатъчно е в случая да се разглежда Лебегова мярка, но само там където някоя от плътностите притежава положителни стойности. Максимум на правдоподобието L_n се достига в същата точка и за логаритъма $LL_n(\vec{x}, \theta) = \log L_n(\vec{x}, \theta)$. Затова е удобно при намирането му да решаваме "уравненията на правдоподобие":

$$\frac{\partial LL_n(\vec{x}, \theta)}{\partial \theta} = 0. \tag{5.4}$$

Определение 5.2 *Наричаме оценката максимално-правдоподобна, ако функцията на правдоподобие е диференцируема и оценката удовлетворява уравненията на правдоподобие (5.4).*

Пример 5.2 *Нека $\xi \in N(m, \sigma^2)$. Нека сме направили n наблюдения. Да намерим максимално - правдоподобните оценки.*

Решение. Определяме $\Theta = R_1 \times R_1^+$. Логаритъмът на правдоподобие има вида:

$$LL_n(x, \theta) = \log L_n(x, \theta) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2.$$

За да намерим максимума на тази функция по m и σ я диференцираме и получаваме уравненията на правдоподобие:

$$\begin{aligned} 0 = \frac{\partial LL_n(x,m,\sigma)}{\partial m} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - m) \\ 0 = \frac{\partial LL_n(x,m,\sigma)}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - m)^2. \end{aligned}$$

Така лесно получаваме двете оценки:

$$\hat{m} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Разбира се, редно е да се убедим, че това са истински максимуми на LL_n. Оценката $\bar{x}$ е неизместена и ефективна, но $\hat{\sigma}$ е изместена. И двете оценки обаче са състоятелни. □

5.2.1 Ефективност

Теорема 5.2 *Ако са изпълнени условията на неравенството на Рао - Крамер и съществува ефективна оценка, то тя съвпада с оценката по метода на максимално правдоподобие.*

Доказателство: Наистина, условието за ефективност (равенството 4.2) изглежда така:

$$\frac{\partial LL_n(\vec{\xi},\theta)}{\partial\theta} = k(\theta)(\hat{\theta}(\vec{x}) - \theta). \tag{5.5}$$

Ако сега искаме да проверим дали тази оценка е максимално- правдоподобна, трябва да заместим θ с $\hat{\theta}$. Тогава очевидно получаваме уравнението на правдоподобие. □

5.2.2 Асимптотика

Асимптотичното поведение на м.п. оценките е най-хубавото им качество.

Теорема 5.3 *Ако плътността $f(x,\theta), \theta \in \Theta \subset R^m$ удовлетворява следните условия:*

1. *притежава непрекъснати производни по θ до втори ред включително;*
2. *максимално - правдоподобната оценка $\hat{\theta}$ е състоятелна в R^m;*
3. *съществуват и са крайни следните интеграли:*

$$c_{i,j} = \mathbf{E}\,(\frac{\partial \log f(\theta)}{\partial\theta_i}\frac{\partial \log f(\theta)}{\partial\theta_j})$$

.

4. *матрицата C притежава ненулева детерминанта;*

то асимптотичното разпределение на $\hat{\theta}$ при $n \to \infty$ е нормално със средна стойност θ и ковариационна матрица $\frac{1}{n}C^{-1}$.

Доказателство: Ще докажем тази теорема за едномерен параметър. Поради състоятелността можем да напишем разложението на производната на функцията на правдоподобие в ред на Тейлор около истинския параметър:

$$0 = \frac{\partial LL_n(\vec{x},\hat{\theta})}{\partial\theta} = \frac{\partial LL_n(\vec{x},\theta)}{\partial\theta} + (\hat{\theta}-\theta)\frac{\partial^2 LL_n(\vec{x},\theta)}{\partial\theta^2} + o(\hat{\theta}-\theta)^2. \tag{5.6}$$

Да разгледаме поотделно трите елемента на това равенство. Първият член:

$$\frac{\partial LL_n(\vec{x},\theta)}{\partial\theta} = \sum_{i=1}^{n} \frac{\partial \log f(x_i,\theta)}{\partial\theta}$$

е сума от еднакво разпределени независими сл.в. и съгласно ЦГТ има асимптотично нормално поведение, когато е правилно нормиран. Съгласно равенството (4.3) имаме

$$\mathbf{E}\,\frac{\partial \log f(\xi,\theta)}{\partial\theta} = 0, \qquad \mathbf{D}\,\frac{\partial \log f(\xi,\theta)}{\partial\theta} = c_{1,1} = C.$$

Значи трябва да нормираме първия член като разделим цялото равенство с $(nC)^{1/2}$. Така първия член клони към $N(0,1)$. Да отбележим, че от равенството $\int f''dx = 0$ следва равенството:

$$-\mathbf{E}\,\frac{\partial^2 \log f(\xi,\theta)}{\partial\theta^2} = \mathbf{E}\,(\frac{\partial \log f(\xi,\theta)}{\partial\theta})^2 = C.$$

Наистина, $\int l''f = \int (f'/f)'f = \int (f''f - f'^2)/f = -\int l'^2 f$.

Вторият член разделяме на два съмножителя:
първият $(nC)^{1/2}(\hat{\theta}-\theta)$ е, което ни трябва,
а вторият по закона на големите числа клони п.с. към 1:

$$\frac{1}{nC}\sum_{i=1}^{n} \frac{\partial^2 \log f(x_i,\theta)}{\partial\theta^2} \longrightarrow \frac{1}{C}\mathbf{E}\,\frac{\partial^2 \log f(\xi,\theta)}{\partial\theta^2} = 1.$$

Последният член очевидно клони към нула:

$$\frac{1}{(nC)^{1/2}} o(\hat{\theta}-\theta)^2 \to 0.$$

Следователно:

$$-(nC)^{1/2}(\hat{\theta}-\theta) \longrightarrow N(0,1).$$

□

5.3 Конструиране на критерии с отношение на правдоподобия

В този параграф отново ще се върнем към проверката на статистически хипотези.

В много случаи не сме в състояние да формулираме просто както хипотезата си, така и алтернативата. В този случай директното използуване на лемата на Нейман-Пирсън е затруднително. Това се получава, например, при сложна алтернатива, както видяхме в предната лекция.

Нека формулираме нашата обща постановка така. Нека $\Theta_0 \subset \Theta$.

- Основна хипотеза $H_0 : \theta \in \Theta_0$,
- Алтернатива $H_1 : \theta \in \Theta - \Theta_0$.

Нека определим следното отношение на правдоподобия:

$$\lambda(x) = \frac{\sup\limits_{\theta \in \Theta_0} L(x, \theta)}{\sup\limits_{\theta \in \Theta} L(x, \theta)}.$$

Ясно е, че $\lambda(x) \leq 1$. Определяме си критичната област по правилото $W = \{x : \lambda(x) \leq c\}$. Особено удачно този тип критерии работи при нормалната теория на регресията, където ще видим и много примери за използуването им.

Тема 6

Многомерно нормално разпределение

Тази лекция съдържа факти от теория на вероятностите, необходими за строгото обосноваване на многомерните статистически процедури. Резултатите ще бъдат изложени тук като следствия от свойствата на нормалното разпределение.

Сведенията от тази лекция могат да бъдат намерени без особени затруднения във всяка книга по математическа статистика или теория на вероятностите.

6.1 Нормално Разпределение

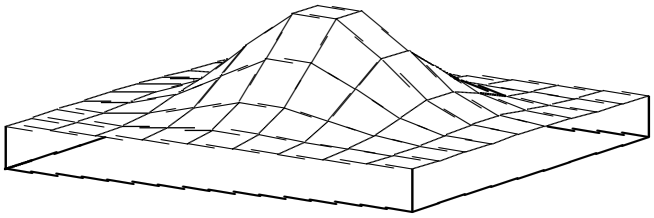
Определение 6.1 *Плътността на стандартното нормално разпределение в R^n има вида:*

$$\phi(x) = \frac{1}{(2\pi)^{n/2}} e^{-||x||^2/2}, \tag{6.1}$$

където $x \in R^n$.

От определението се вижда, че тази плътност зависи само от нормата на вектора x и, следователно, е инвариантна относно всякакви ортогонални трансформации — те запазват нормата и имат якобиян равен на 1. Също така е ясно, че тя може да се представи като произведение на n едномерни стандартни нормални плътности (виж фиг.6.1).

До края на тази лекция ще предполагаме, че случайната величина ξ има *стандартно нормално разпределение* в R^n. Плътността на многомерното нормално разпределение от по - общ вид $N(m, C)$ в R^n има вида:



Фиг. 6.1: N(0,I) в R^2

$$\phi(x,m,C) = \frac{1}{(2\pi)^{\frac{n}{2}}(det(C))^{\frac{1}{2}}} e^{-(x-m)'C^{-1}(x-m)/2}, \tag{6.2}$$

където $x \in R^n$, $m \in R^n$ е средната стойност, а C – ковариационната матрица. На фиг.6.2 е показана линия на постоянно ниво на двумерна гаусова плътност – тя е елипса.

И тук както в едномерния случай имаме връзка между параметрите на закона и моментите на сл.в.:

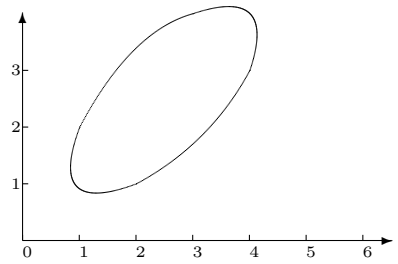
Теорема 6.1 *Ако $\eta \in N(m,C)$, то*

$$E\eta = m, \qquad E(\eta - m)(\eta - m)' = C.$$

Теоремата може да се докаже с проста смяна на променливите или като следствие на следната по - обща теорема.

Теорема 6.2 *Случайната величина $\eta = T\xi + a$, където T е неизроден линеен оператор от R^n в $R^k (n \geq k)$, има разпределение $N(a, TT')$ в R^k.*

От тази теорема следва, че всички маргинални разпределения (или проекции в произволна размерност) са нормални. Следва също, че произволна линейна функция от (зависими или независисми) нормални сл.в. е нормална сл.в.



Фиг. 6.2: Линия на ниво

Верно е също, че и условните разпределения (при линейни ограничения от типа на равенството) са гаусови.

Доказателство: Ще разгледаме само случая $\eta = T\xi$. В случая операторът T се представя просто като матрица с $k \leq n$ реда и n колони и трябва да притежава пълен ранг k. Това означава, че нейните редове са линейно независими вектори в R^n. Нека означим с S подпространството от линейните им комбинации. То очевидно има размерност k. Нека допълним редовете на T с $n-k$ ортогонални помежду си и на S единични вектори и означим така получената матрица с $\tilde{T}$. Тогава по формулата за смяна на променливите $\tilde{\eta} = \tilde{T}\xi$ ще има разпределение с плътност:

$$f(x) = \frac{1}{(2\pi)^{n/2} det(\tilde{T})} e^{-\frac{1}{2}x'(\tilde{T}\tilde{T}')^{-1}x}.$$

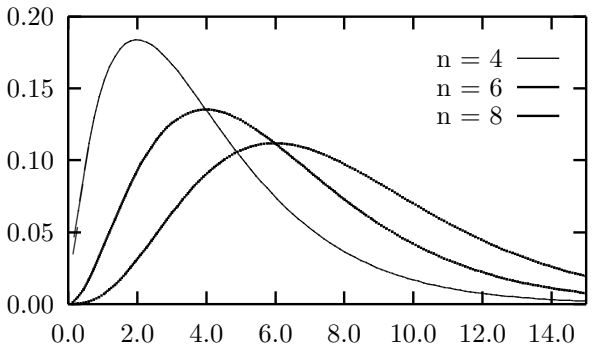
Но $det(\tilde{T}) = det(T)$ и матрицата $\tilde{T}\tilde{T}'$ е блочно диагонална. Такава е и нейната обратна. Следователно, разпределението се разпада в произведение на две плътности:

$$f(x) = f_1(x_1)f_2(x_2) = \frac{1}{(2\pi)^{\frac{k}{2}} det(T)} e^{-\frac{1}{2}x_1'(T'T)^{-1}x_1} \quad \frac{1}{(2\pi)^{\frac{n-k}{2}}} e^{-\frac{1}{2}||x_2||^2}.$$

Тук разлагането $x = \{x_1, x_2\} = x_1 + x_2$ представя вектора в неговите проекции (координати) в подпространството S и неговото ортогонално допълнение. От тук лесно следва твърдението на теоремата. □

6.2 Теорема на Кокрън

Определение 6.2 *Случайната величина $\chi_n^2 = \xi'\xi$ има разпределение $\chi^2(n)$ с n степени на свобода.*



Фиг. 6.3: χ^2 разпределение

Теорема 6.3 *Случайната величина χ^2_n има плътност:*

$$f(x,n) = C(n)x^{n/2-1}e^{\frac{-x}{2}}. \tag{6.3}$$

Тук $C(n) = (\frac{1}{2})^{n/2}\Gamma(\frac{n}{2})$ е нормираща константа.

Веднага се вижда, че това е Гама-разпределение $\Gamma(\frac{n}{2}, \frac{1}{2})$ и изводът може да се направи по индукция от $n = 1$ и възпроизводящите свойства на i t x-разпределението: $\Gamma(a,\lambda) + \Gamma(b,\lambda) = \Gamma(a+b,\lambda)$.

Средната стойност на $\chi^2_n = \xi'\xi$ е очевидно n, а дисперсията лесно се пресмята и е равна на $2n$. От следната проста лема непосредствено се вижда, че разпределението $\chi^2_n = \xi'\xi$ не може да се получи от друга квадратична форма на n гаусови сл.в., освен тривиалната.

Лема 6.1 *Нека $\xi_i, i = 1, ..., n$ са независими сл.в. с еднаква дисперсия и $\lambda_i, i = 1, ..., n$ са такива, че $\sum_{i=1}^n \lambda_i = 1$. Тогава сл.в. $\eta = \sum_{i=1}^n \lambda_i\xi_i$ има минимална дисперсия, когато $\lambda_i = 1/n, i = 1, 2, ..., n$.*

Доказателство: Да предположим, че $D(\xi_i) = 1$. Тогава $D(\eta) = \sum_{i=1}^n \lambda_i^2$. В сила е обаче неравенството:

$$(\sum_{i=1}^n \lambda_i)^2 \leq n\sum_{i=1}^n \lambda_i^2.$$

При това, равенство се достига тогава и само тогава, когато $\lambda_i = C$. □

Лема 6.2 *Ако дадена квадратична форма Q има ранг q и сл.в. $\xi'Q\xi$ има разпределение е χ^2_q, то Q е проектор.*

Доказателство: Да напомним, че проекторите са неотрицателно - определени оператори (т.е. са самоспрегнати $P' = P$), а освен това са и идемпотенти ($P^2 = P$). Това значи, че собствените им числа могат да бъдат само 0 или 1. Естествено, броят на ненулевите собствени числа е равен на ранга.

За доказателството е достатъчно да сравним дисперсиите на двете разпределения и да се възползуваме от лема 6.1. Ние обаче ще го изведем директно — така ще пресметнем и дисперсията на χ^2_q разпределение.

Действително, Q е неотрицателно определена и значи може да се представи като $Q = UDU'$, където U е ортогонална матрица, а D - диагонална. Тогава сл.в. $\xi' Q \xi$ и $\xi' D \xi$ имат едно и също $\chi^2(q)$ разпределение.

$$\mathbf{D}\,(\xi' D \xi) = \mathbf{E}\,[(\sum_{i=1}^{n} d_i(\xi^2 - 1))^2] = 2\sum d_i^2 \geq 2n.$$

Тъй като $tr(Q) = \sum_{i=1}^{n} d_i = n = tr(I)$, последното неравенство става равенство само когато $d_i = 1$,$i = 1, 2, ..., n$. □

Теорема 6.4 Теорема на Кокрън. *Нека Q, R, S са неотрицателно определени матрици с рангове q, r, s съответно, $Q = R + S$ и случайната величина $\xi' Q \xi$ има разпределение $\chi^2(q)$. Случайните величини $\xi' R \xi$ и $\xi' S \xi$ са независими и имат разпределения $\chi^2(r)$ и $\chi^2(s)$ тогава и само тогава, когато $q = r + s$.*

Доказателство: Първо да отбележим, че съгласно лема 6.2 матрицата Q е проектор и можем да се ограничим в пространство с размерност q, когато $Q = I$.

Достатъчност. Имаме: $I = R + S = U(D_R + D_S)U'$. Следователно, $I = D_R + D_S$. Ако $q = r + s$, то D_R и D_S имат съответния брой ненулеви елементи, значи R и S са проектори и $RS = SR = 0$. За доказателството на независимостта използуваме равенството: $||Qx||^2 = ||x||^2 = ||Rx||^2 + ||Sx||^2$ за всяко x и $x'R'Rx = ||Rx||^2 = x'Rx$. Остава да приложим теорема 6.2 за операторите R и S и определението на Хи-квадрат разпределение.

Необходимост. Равенството $q = r + s$ следва директно от лема 6.2. □

Частен случай от теоремата на Кокрън е независимостта на $\bar{x} = 1/n \sum x_i$ и $S^2 = \sum (x_i - \bar{x})^2$. Наистина,

$$||\xi||^2 = n\bar{\xi}^2 + S^2 = \xi' B \xi + \xi'(I - B)\xi.$$

Но тогава и съответните квадратични форми са породени от ортогонални проектори. Т.е. $B\xi \perp (I - B)\xi$. Тук

$$B = \left\| \begin{array}{cccc} 1/n & 1/n & \dots & 1/n \\ 1/n & 1/n & \dots & 1/n \\ \dots & \dots & \dots & \dots \\ 1/n & 1/n & \dots & 1/n \end{array} \right\|.$$

Пример 6.1 *Условно математическо очакване и коефициент на корелация.*

Нека случайната величина $\xi \in R^2$ и има разпределение $N(m, S)$. Тогава условното математическо очакване и коефициентът на корелация се получават по формулите: $\mathbf{E}\,(\xi_2|\xi_1) = a\xi_1 + b$, $r(\xi_1, \xi_2) = a/S_{22}$, където $b = m_2 - am_1$, $a = S_{12}(S_{22}/S_{11})^{1/2}$. С S_{ij} сме означили елементите на ковариационната матрица на двумерната сл.в. ξ. В частност $S_{22} = \sigma^2(\xi_2)$.

Проверете тези формули.

Тема 7

Тестове на Стюдент и Фишер

Тук ще дадем рецептите на няколко най-популярни статистически извода, базирани на разпределенията свързани с нормалното. Ще разгледаме определени статистики, доверителни интервали за неизвестните параметри и хипотези свързани с тях

7.1 Доверителен интервал за дисперсия

Нека $\xi \in N(\mu, \sigma^2)$. Нека сме направили n наблюдения. Намерете оптимална доверителна област за σ^2.

Статистиката $S^2 = \sum_{i=1}^{n}(x_i - \mu)^2$ има разпределение $\sigma^2\chi^2_n$ (определение 6.2). От лема 3.2 следва, че доверителна област с минимална дължина за S^2 се дава от равенството:

$$\mathsf{P}(q_l \leq \frac{S^2}{\sigma^2} \leq q_u) = 1 - \alpha.$$

Тук квантилите на χ^2-разпределението q_l, q_u се определят от уравненията

$$F(q_l) + 1 - F(q_u) = \alpha, \qquad f(q_l) = f(q_u),$$

където F и f са съответно функцията на разпределение и плътността на χ^2-разпределение с n степени на свобода. Окончателно получаваме:

$$\frac{S^2}{q_u} \leq \sigma^2 \leq \frac{S^2}{q_l}. \tag{7.1}$$

Когато м.о. μ е неизвестно, се използува статистиката:

$$S^2(x) = \sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{7.2}$$

която не зависи от μ и има разпределение $\sigma^2\chi^2_{n-1}$. Останалото е същото.

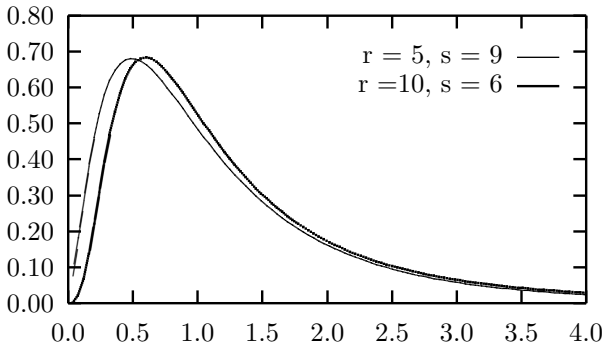
Да отбележим, че на практика така определения интервал (с минимална дължина) се използува рядко. По-често се приравняват вероятностите на двете опашки: $F(q_l) = 1 - F(q_u) = \alpha/2$. Така квантилите се вземат направо от таблицата.

7.2 Разпределение на Фишер

Определение 7.1 *Разпределение на Фишер - Снедекор с r и s степени на свобода има частното:*

$$f_{r,s} = \frac{\chi^2_r/r}{\chi^2_s/s},$$

където χ^2_r, χ^2_s са независими сл.в. с χ^2 разпределение.



Фиг. 7.1: Разпределение на Фишер - Снедекор

Теорема 7.1 *Разпределението на Фишер $F_{r,s}$ има плътност:*

$$f(x, r, s) = C(r, s)x^{r/2-1}(1 + rx/s)^{-(r+s)/2}, \tag{7.3}$$

където $x > 0$ и $C(r, s)$ - нормираща константа.

Докажете го. Можете да използувате направо определението. Иначе може да използувате връзката между Гама, Бета и F - разпределение:

$$b(2a, 2b) = \frac{\chi^2_a}{\chi^2_a + \chi^2_b}.$$

7.2.1 Критерий на Фишер за независими извадки

Нека проверим хипотезата за равенство на дисперсии на две различни генерални съвкупности (г.с.) с нормално разпределение на основата на независими извадки от тях с размери n_1 и n_2 съответно. Ще предположим, че средните на двете г.с. са неизвестни. Да образуваме статистиките:

$$S^2(x) = \sum_{i=1}^{n_1}(x_i - \bar{x})^2, \qquad S^2(y) = \sum_{i=1}^{n_2}(y_i - \bar{y})^2. \tag{7.4}$$

Съгласно определение 6.2 и теорема 6.4 (на Кокрън), всяка от тези статистики има Хи-квадрат разпределение, умножено със съответната σ^2. При това те са независими. Ако за нулева изберем хипотезата: $H_0 : \sigma(x) = \sigma(y)$, то съгласно определение 7.1 частното:

$$f = \frac{S^2(x)/(n_1 - 1)}{S^2(y)/(n_2 - 1)} = \frac{s^2(x)}{s^2(y)}$$

ще има разпределение на Фишер с $(n_1 - 1)$ и $(n_2 - 1)$ степени на свобода съответно. Така с използването на тази статистика можем да проверяваме нулевата хипотеза срещу различни алтернативи:

- За алтернативите $H_1 : \sigma(x) > \sigma(y)$ (или $H_1 : \sigma(x) < \sigma(y)$) критерият ще бъде равномерно най - мощен. Критичната област е $W = \{f > z\}$; $F(z) = 1 - \alpha$.

- За алтернативата $H_1 : \sigma(x) \neq \sigma(y)$ не съществува равномерно най - мощен критерий. На практика се използва следната процедура. Извадката с по-малка извадъчна дисперсия s^2 се означава с y. Критичната област отново е $W = \{f > z\}$; $F(z) = 1 - \alpha$. Така става възможно използването на таблици за квантилите само от горната половина на F разпределението.

Можем да обобщим теста на Фишер и със помощта на следното твърдение, което е директно следствие от теоремата на Кокрън 6.4.

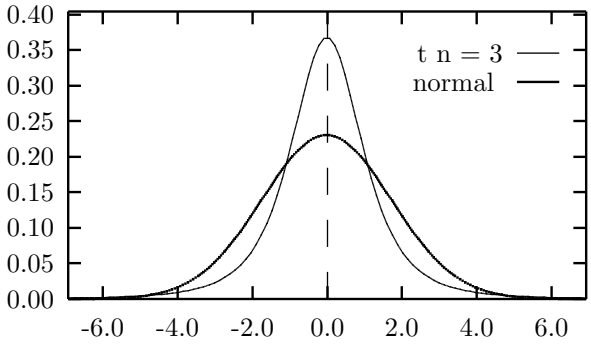
Теорема 7.2 *Нека P, Q са два проектора (матрици) в R^n такива, че $PQ = QP = P$ и $P \neq Q$ и ξ е стандартна нормална в R^n. Тогава частното $f = (\xi'(Q - P)\xi)/(\xi' P \xi)$ има разпределение на Фишер с $dim(Q) - dim(P)$ и $dim(P)$ степени на свобода.*

7.3 Разпределение на Стюдент

Определение 7.2 *Разпределение на Стюдент $T(n)$ с n степени на свобода има частното:*

$$t_n = \frac{\xi}{(\chi_n^2/n)^{1/2}},$$

където ξ, χ_s^2 са независими сл.в. с $N(0,1)$ и $\chi^2(n)$ разпределение с n степени на свобода съответно.



Фиг. 7.2: Разпределение на Стюдент

Теорема 7.3 *Плътността на разпределение на t_n се задава с формулата*

$$f(x,n) = C_n(1 + x^2/n)^{-(n+1)/2}. \tag{7.5}$$

Тук C_n е нормираща константа.

Докажете го. Може да се изведе направо. Може да се използува връзката между T и F разпределения: $(t_n)^2 = f_{1,n}$.

От закона за големите числа следва, че граничното разпределение на t_n при $n \to \infty$ е $N(0,1)$.

7.3.1 Доверителен интервал за м.о. μ

Нека $\xi \in N(\mu, \sigma^2)$. Нека сме направили n наблюдения. Да намерим оптимална доверителна област за μ. Тук втория параметър σ се смята за неизвестен.

Статистиката

$$t = n^{1/2}(\bar{x} - \mu)/s \tag{7.6}$$

има T-разпределение с $(n-1)$ степени на свобода и това разпределение не зависи от σ. Тук $s^2(x) = S^2(x)/(n-1)$.

От лема 3.2 следва, че оптималната доверителна област за $\bar{x}$ е

$$|\bar{x} - \mu| < zs/n^{1/2}).$$

Тук z се определя от уравнението $F(z) - F(-z) = 1 - \alpha$ и се нарича *двустранен квантил* на T-разпределение с $n-1$ степени на свобода за критично ниво α. Така построения доверителен интервал удовлетворява равенството: $F'(x) = F'(-x)$, което следва от симетричността на плътността (7.5) и, следователно, е с минимална дължина.

7.3.2 Критерий на Стюдент

Нека пак предположим, че наблюденията са от $N(\mu, \sigma^2)$. Статистиката t (виж. 7.6) ще има T разпределение независимо от σ. Ако за нулева изберем хипотезата: $H_0 : \mu = 0$, то това ни дава възможност да проверяваме нулевата хипотеза срещу различни алтернативи:

- За алтернативата $H_1 : \mu > 0$ (или $H_1 : \mu < 0$) критерият ще бъде равномерно най - мощен. Критичната област е $W = \{t > z\}$; $F(z) = 1 - \alpha$.

- За алтернативата $H_1 : \mu \neq 0$ не съществува равномерно най - мощен критерий. Критичната област е $W = \{|t| > z\}$; $F(z) = 1 - \alpha/2$.

При голям брой на наблюденията ($n > 100$) разпределението на Стюдент клони към стандартно нормално, така че се упростява пресмятането на квантила.

7.3.3 Критерий на Стюдент за независими извадки

Нека са ни дадени независими извадки $x_1, x_2, \ldots, x_n$ и $y_1, y_2, \ldots, y_m$ от независими наблюдения, съответно, на $\xi_1 \in N(\mu_1, \sigma_1)$ и $\xi_2 \in N(\mu_2, \sigma_2)$. Задачата е по тези наблюдения да проверим хипотезата, че двете генерални съвкупности съвпадат. Т.е. $\mu_1 = \mu_2$ и $\sigma_1 = \sigma_2$.

От начало се прилага теста на Фишер за проверка на равенството на двете дисперсии (виж.7.2.1). Разпределението на Фишер не зависи от стойностите на неизвестните м.о. на двете популации. При отхвърляне на нулевата хипотеза $H_0 : \sigma_1 = \sigma_2$ задачата е решена - двете популации

се приемат за различни. Ако обаче нямаме основания да я отхвърлим, преминаваме към проверката за равенство на м.о.

Да разгледаме хипотезата $H_0 : \mu_1 = \mu_2$ срещу алтернативата $H_0 : \mu_1 \neq \mu_2$ при условие, че $\sigma_1 = \sigma_2 = \sigma$. Да разгледаме обединената (виж.7.4) оценка на дисперсията σ^2:

$$s^2 = \frac{1}{n+m-2}(S^2(x) + S^2(y)) \tag{7.7}$$

Тя е неизместена и разпределението на $(n+m-2)s^2/\sigma^2$ е очевидно χ^2_{n+m-2}. От друга страна сл.в.

$$\bar{x} - \bar{y} \in N(\mu_1 - \mu_2, \sigma^2(\frac{1}{n} + \frac{1}{m})).$$

Двете сл.в. са независими и, следователно, при изпълнена $H_0 : \mu_1 = \mu_2$ статистиката:

$$t = (\frac{nm}{n+m})^{1/2}\frac{\bar{x} - \bar{y}}{s} \tag{7.8}$$

ще има разпределение T_{n+m-2}. Това ни дава възможност, както и по-горе да си проверяваме нулевата хипотеза при различни алтернативи. Така на втората стъпка окончателно ще можем да отговорим за равенството на двете популации.

Със същия критерий може да се проверява и равенство само на м.о. на две г.с. Но формулите за пресмятане на съответната статистика се различават, когато не приемаме за равни дисперсиите им.

Тема 8

Регресионен анализ

Тази статистическа процедура е най - старата и, може би, най - популярната. Терминът "регресия" е въведен от английския антрополог Ф.Галтон във връзка с откритата от него тенденция синовете на родители с ръст по - висок от нормалния, да имат ръст по - близо до средната стойност. Този факт Галтон нарекъл "regression to mediocrity".

Регресионният анализ намира най - често приложение за изследване на причинно - следствени връзки. Той ни позволява да проверяваме хипотези за наличието на такава връзка и да я оценяваме количествено.

Изложеното в тази лекция е незначителна част от теорията, посветена на линейната регресия и пояснява донякъде само това, което е заложено в най - простите регресионни процедури. На интересуващия се читател горещо препоръчваме класическите книги (Себер 1976) и (Дрейпер и Смит 1973).

Нека наблюдаваните променливи са много и една от тях е натоварена с по - особено смислово съдържание. Отделената променлива ще наричаме зависима или отклик. Останалите – независими или предиктори. Поставяме си следните въпроси:

1. Дали стойностите на отклика се влияят или зависят от останалите променливи?

2. Каква е функционалната връзка между стойностите на променливите (т.е. може ли да се избере модел на зависимостта и оценят параметрите му)?

3. Доколко получената връзка отговаря на действителността (или доколко моделът е адекватен)?

4. Какво можем да очакваме от отклика при зададени нови стойности на предикторите (задача за прогноз)?

Ние ще изведем всички свойства на линейната регресия от общите свойства на гаусовото разпределение. Болшинството статистически програми работят по тези формули, изведени в предположение за гаусово разпределение на грешката. Практиката, обаче, показва, че това ограничение далеч не винаги е правдоподобно, пък и резултатите получени с него – не винаги удовлетворителни.

8.1 Линейни модели с гаусова грешка

В цялата лекция нататък ще предполагаме, че $\epsilon \in N(0, \sigma^2 I)$, т.е. че грешките от наблюденията са независими, еднакво разпределени гаусови сл.в. с нулева средна. За наблюденията y ще предполагаме,че е изпълнен следният модел:

$$y \quad = \quad z + \epsilon. \tag{8.1}$$

За неизвестното $z = \mathbf{E}\, y$ се предполага, че $z \in Z$ — линейно подпространство на R^n с размерност k. Това на пръв поглед странно предположение се оказва много удобно от теоретична гледна точка — всички линейни модели лесно се вписват в него.

В долната теорема са сумирани свойствата на оценките, които следват от гаусовото разпределение на ϵ.

Теорема 8.1 *За модела (8.1) са изпълнени свойствата:*

а. максимално-правдоподобните оценки на z и σ^2 се получават по метода на най - малките квадрати:

$$\hat{y} = \underset{z \in \hat{Y}}{argmin}\, ||z - y||^2;$$

$$\hat{\sigma}^2 = \frac{1}{n}||\hat{y} - y||^2;$$

б. оценките $\hat{y}$ и $y - \hat{y}$ са независими.

в. оценката $\hat{y}$ има изродено върху Z разпределение $N(0_Z, \sigma^2 I_Z)$.

г. статистиката $||\hat{y} - y||^2$ има разпределение $\sigma^2 \chi^2(n - k)$;

Доказателство: Всички твърдения са пряко следствие от определенията на максимално - правдоподобните оценки в гаусовия случай. □Ако се

наложи да предположим различни дисперсии за наблюденията, например, $\epsilon \in N(0, \sigma^2 W)$, то в горните твърдения просто трябва да заменим скаларното произведение и нормата:

$$x'y = x'W^{-1}y, \qquad ||x||^2 = x'W^{-1}x.$$

Тогава твърденията на теоремата и всички последващи твърдения остават без изменение.

В практиката често възниква необходимостта от сравняване на различни модели. Едно средство за това ни дава следната теорема от нормалната теория. Ще означим с H_Z линейния проектор върху подпространството Z: $H_Z(y) = \hat{y}$.

Теорема 8.2 *Нека се налага да проверим хипотезата*

$$H_0 : z \in Z_0 \quad \textit{срещу хипотезата} \quad H_1 : z \in Z_1 \backslash Z_0,$$

където $Z_0 \subset Z_1$ са линейни подпространства на R^n с различни размерности $k < m$ съответно. Тогава критичната област се определя от неравенството:

$$f_{m-k,n-m} = \frac{||y_1 - y_0||^2/(m-k)}{||y - y_1||^2/(n-m)} > F_{1-\alpha}, \tag{8.2}$$

като статистиката $f_{m-k,n-m}$, при изпълнена H_0, има разпределение на Фишер с $m-k$ и $n-m$ степени на свобода, а $F_{1-\alpha}$ е квантил на това разпределение. С y_i сме означили проекциите на y върху $Z_i, (i = 0, 1)$.

Доказателство: Формата на областта следва от принципа за отношение на правдоподобия:

$$\lambda(y) = \frac{\sup\limits_{z \in Z_0, \sigma} L(y - z, \sigma)}{\sup\limits_{z \in Z_1, \sigma} L(y - z, \sigma)} = (\frac{||y - y_1||}{||y - y_0||})^n.$$

Проверката на неравенството $\lambda(y) > c$ е еквивалентна на критичната област определена от неравенството (8.2). Твърдението за разпределението е пряко следствие от теоремата на Кокрън. □

Когато към модела (8.1) добавяме предположения за параметризация на Z, получаваме различните форми на, т.н. в литературата, общ линеен модел с гаусова грешка. Някои от тях ще разгледаме в следващите лекции.

8.2 Нормална линейна регресия

Нека изследваният модел е от вида

$$y = Xa + e, \tag{8.3}$$

където $y, e \in R^n, a \in R^m, X \in R^n \times R^m$, грешките $e \in N(0, \sigma^2 I)$. Тук y и X са наблюденията, а σ^2 и a са неизвестни.

Теорема 8.3 *(Гаус - Марков) Ако X има пълен ранг m, оценката за неизвестните параметри a по метода на най - малките квадрати е*

$$\hat{a} = (X'X)^{-1}X'y \tag{8.4}$$
$$cov(\hat{a}) = \sigma^2(X'X)^{-1} \tag{8.5}$$

Оценката $\hat{a}$ е неизместена, ефективна и съвпада с оценката по метода на максимално правдоподобие.

Доказателство: Методът на най - малките квадрати в случая ни учи да търсим минимум на $||y - Xa||^2$, което съвпада с твърдение а. на теорема 8.1 и, следователно, решенията на двата метода съвпадат. Подпространството $Z = Xa$ е линейна комбинация на колоните на X. Тогава проекторът H_Z има вида $H_Z = X(X'X)^{-1}X'$. Оценката $\hat{a}$ за a е просто решение на уравнението $\hat{y} = X\hat{a}$,т.е. съвпада с равенството (8.4). Това решение съществува и е единствено поради пълния ранг на X.

Като заместим y в (8.4) получаваме

$$\hat{a} = a + (X'X)^{-1}X'\epsilon,$$

което влече неизместеността на $\hat{a}$. От същото представяне следва и представянето на $cov(\hat{a})$ в (8.5). □

От теорема 8.1 веднага получаваме, че неизместена оценка на σ^2 ще получим по формулата:

$$\hat{\sigma^2} = \frac{1}{n-k}||y - X\hat{a}||^2. \tag{8.6}$$

Тази оценка, обаче, не е максимално правдоподобна.

Тема 9

Проверки на хипотези в регресията

В тази лекция ще експлоатираме безпощадно теорема 8.2 и ще конструираме множество популярни хипотези в линейната регресия. В някои частни случаи конструираните доверителни области (поради естествените "широки" алтернативни хипотези) ще станат и доверителни интервали за неизвестните параметри.

9.1 Коефициент на детерминация

Коефициент на детерминация или проверка на наличието на линейна връзка между X и y.

Нека разгледаме сега регресионен модел със свободен член:

$$y = Xa + b\vec{1} + e, \tag{9.1}$$

където b е "нов" неизвестен параметър, а $\vec{1}$ е n-мерен вектор от единици. Да се опитаме да проверим наличието на линейна връзка между X и y.

Нека е вярна хипотезата $H_o : a = 0$. Естествената контра хипотеза е $H_1 : a \neq 0$. Следователно, Z_0 има размерност $k = 1$, а Z_1 е с размерност $m = dim(a) + 1$. От теорема 8.2 получаваме, че критичната област за проверка на хипотезата $H_0 : z \in Z_0$ срещу хипотезата $H_1 : z \in Z_1 \backslash Z_0$ се определя от неравенството:

$$F = \frac{||y_1 - y_0||^2/(m-1)}{||y - y_1||^2/(n-m)} > F_{1-\alpha},$$

като при изпълнена H_0 статистиката $F \in F(m-1, n-m)$.

В приложната статистика съответните суми от квадрати имат популярни наименования, разкриващи тяхната роля в тази проверка:

$$SSR = ||y - y_1||^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad \text{- Sum of Squares of Residuals}$$

$$SSM = ||y_1 - y_0||^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \qquad \text{- Sum of Squares due to the Model}$$

Частното

$$R^2 = \frac{SSM}{SSM + SSR}$$

се нарича коефициент на детерминация и има смисъла на коефициент на корелация — колкото по - близко е до единицата, толкова по "детерминиран" е моделът.

9.2 Проверка за равенство на нула на някой от коефициентите

Нека е вярна хипотезата $H_0 : a_1 = 0$. Естествената контра хипотеза е $H_1 : a_1 \neq 0$. Следователно, Z_0 има размерност $k = dim(a) - 1$, а Z_1 - размерност $m = dim(a)$. От теорема 8.2 получаваме, че оптималната критична област за проверка $H_0 : z \in Z_0$ срещу хипотезата $H_1 : z \in Z_1 \backslash Z_0$ се определя от неравенството:

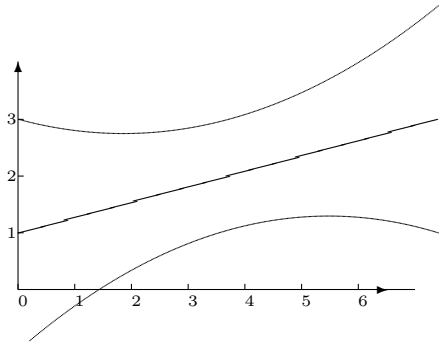
$$F = \frac{||y_1 - y_0||^2}{||y - y_1||^2/(n-m)} > F_{1-\alpha},$$

като при изпълнена H_0 статистиката $F \in F(1, n-m)$. Но това е квадрат на t-разпределение, от където получаваме, че статистиките

$$t_i = \frac{\sqrt{(n-m)}\hat{a}_i}{\hat{\sigma}((X'X)^{-1}_{ii})^{1/2}} \tag{9.2}$$

имат разпределение на Стюдент с $n - m$ степени на свобода при изпълнена хипотеза $H_0 : a_i = 0$. Естествено, със същото разпределение се пресмятат и доверителните интервали около оценките за неизвестните параметри (при изпълнена H_1). Това следва от неизместеността им и от това, че оценките на параметрите не зависят от оценката на дисперсията.

Изведете строго разпределението на статистиките (9.2).



Фиг. 9.1: Проста линейна регресия

9.3 Доверителен интервал за прогноза

За произволни стойности x на предикторите от областта, за която е верен модела (9.1), случайната величина $\hat{y} = x'\hat{a} + \hat{b}$ е неизместена оценка за $E(y|x)$ и

$$D(\hat{y}|x) = \sigma^2(\frac{1}{n} + (x - \bar{X})'(\tilde{X}'\tilde{X})^{-1}(x - \bar{X})). \tag{9.3}$$

Тук с $\bar{X}$ сме означили вектора $\frac{1}{n}X'E$ и E е $(n \times m)$ матрица от единици, а с $\tilde{X}$ сме означили матрицата от центрирани данни (с извадена средна стойност). Следователно, грешката на прогнозираната стойност на конкретното наблюдение ще бъде

$$\sigma_y^2(x) = \sigma^2(1 + \frac{1}{n} + (x - \bar{X})'(\tilde{X}'\tilde{X})^{-1}(x - \bar{X})). \tag{9.4}$$

Проверете уравнения (9.3) и (9.4).

На фигурата е нарисувана апроксимиращата права при простия линеен модел $y = ax + b + \epsilon$. С двете параболи са отбелязани доверителните граници за наблюдаваната стойност съгласно формула (9.4). С аналогична форма, но значително по - тесен е коридорът за модела – формула (9.3). Така се вижда колко опасни (и понякога безсмислени) могат да бъдат прогнози за далечното бъдеще, основани на тенденция, наблюдавана в краен интервал от време.

9.4 Проверка на адекватността на модела

Проверката за адекватност на модела в регресионния анализ е възможна само в два случая: ако е известна σ^2 или ако разполагаме с независима

от SSR и от параметрите на модела нейна оценка.

В първия случай можем да пресметнем статистиката SSR, която има разпределение $\sigma^2\chi^2$ със степени на свобода $n-m$, ако моделът е адекватен, и отместено надясно разпределение при неадекватен модел. Така проверката е лесна – критичната област се определя от неравенството:

$$SSR > \sigma^2\chi^2_{1-\alpha}.$$

Във втория случай, когато не знаем σ^2, се налага да използуваме някоя нейна оценка.

Най-популярния начин за получаване на независима оценка за σ^2 е да се провеждат повторни наблюдения при фиксирани стойности на предикторите. При такива наблюдения сумата SSR също се разлага на две независими събираеми, от които се конструира статистика, която има разпределение на Фишер, в случай че моделът е адекватен. Обикновено тази задача се решава със средствата на еднофакторния дисперсионен анализ. Отделните експериментални точки x се разглеждат като нива на фактор (групираща променлива). За всяко x имаме по n_x наблюдения $y_i(x)$. Имаме равенството:

$$SSR = \sum_x (y_i(x) - \bar{y}(x))^2 + \sum_x n_x(\bar{y}(x) - \hat{y}(x))^2 = SSI + SSM. \quad (9.5)$$

Първата сума не зависи от модела, а втората има разпределение $\sigma^2\chi^2$ със съответен брой степени на свобода, ако моделът е адекватен, и отместено надясно разпределение при неадекватен модел. Така критичната област ще се определи от неравенството:

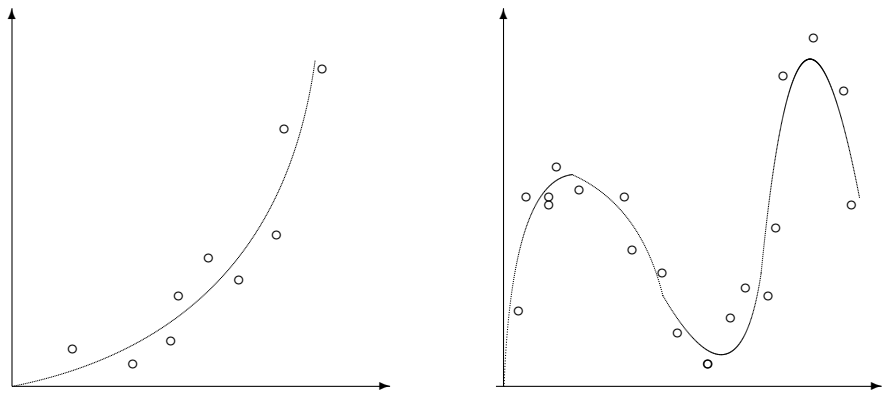
$$\frac{SSM/k}{SSI/j} > F_{1-\alpha}, \qquad j = n-m-k, k = \sum_x (n_x - 1).$$

Опишете подпространствата Z_0 и Z_1 в този случай и изведете уравнение (9.5). Постройте критичната област.

Тема 10

Полиномна регресия

Това е една много популярна форма на линейната регресия, при която за регресионен модел се използуват полиноми. При нея се прибягва, когато нямаме априорни познания за аналитичната форма на модела. На долните картинки са показани типични данни от този случай.



Фиг. 10.1: Криволинейни данни

В тази лекция си поставяме следните цели

- да въведем полиномната регресия;
- на примера на ортогоналните полиноми да покажем изчислителните и статистически удобства на метода на ортогонализацията;

1	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1
1	2	4	8	16	32	64	128	256
1	3	9	27	81	243	729	2187	6561
1	4	16	64	256	1024	4096	16384	65536
1	5	25	125	625	3125	15625	78125	390625
1	6	36	216	1296	7776	46656	279936	1679616
1	7	49	343	2401	16807	117649	823543	5764801
1	8	64	512	4096	32768	262144	2097152	16777216

Таблица 10.1: Матрица X на полиномна регресия

- да разгледаме един популярен пример с данни.

Полиномният модел може да се запише във формата:

$$y_i = a_0 + a_1 x_i + a_2 x_i^2 + \ldots + a_n x_i^n + \epsilon_i. \tag{10.1}$$

Доказахме в предните лекции, че най - доброто решение за апроксимация на модела спрямо данните е методът на най - малките квадрати (МНК).

10.1 Населението на САЩ

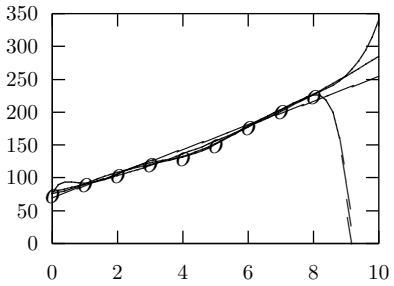
Ще разгледаме един пример. Числата 75.995 91.972 105.711 123.203 131.669 150.697 179.323 203.212 226.505, са публикувани от Американския статистически институт и представляват населението (в милиони хора) на САЩ за периода от 1900 до 1980 г. Нека си поставим за задача да прогнозираме населението за две последователни десетилетия напред – 1990 и 2000. Като базисен ще разгледаме модела (10.1). Ясно е, че ще трябва да се ограничим с $n \leq 8$, тъй като разполагаме с 9 наблюдения и последния полином става интерполиращ.

За конкретния случай матрицата X показана на таблица 10.1 и се вижда, че е почти изродена.

Оценката (8.4) $X'X\hat{a} = X'y$ се получава като решение на системата уравнения:

$$\begin{aligned}
\sum y_i &= a_0 n + a_1 \sum x_i + a_2 \sum x_i^2 +, \ldots, +a_n \sum x_i^n \\
\sum x_i y_i &= a_0 \sum x_i + a_1 \sum x_i^2 + a_2 \sum x_i^3 +, \ldots, +a_n \sum x_i^{n+1} \\
&\ldots \\
\sum x_i^n y_i &= a_0 \sum x_i^n + a_1 \sum x_i^{n+1} + a_2 \sum x_i^{n+2} +, \ldots, +a_n \sum x_i^{n+n}.
\end{aligned}$$

Приготвянето на такава матрица изисква пресмятането на твърде голям брой суми. Така изглежда $X'X$, даже ако забравим последните няколко колони, т.е. разглеждаме полином от по - ниска степен. Затова по - лесно е решението системата $Xa = y$, например, чрез използуването на (Singular Value Decomposition). Именно по този начин в демонстрационната програма census.m на системата MATLAB се решава този пример.



Фиг. 10.2: Населението на **САЩ** 1900 - 1980

На фигурата 10.2 са представени оригиналните данни, заедно с няколко регресионни полинома – от степени 1,2,6,8. Най-очевидно е несъответствието на прогнозираната стойност за полинома от 8 степен, който предсказва изчезване на цялото население на САЩ преди 2000 г. Пресмятанията са вършени с двойна точност така, че на резултатите от изчисленията може да се вярва. На долните таблици ще видим най - същественото от тези числени сметки. Главната цел, обаче е да въведем математическия апарат, който ще ни помогне да изберем "оптималния" от тези полиноми.

10.2 Ортогонални полиноми

Полиномите от произволна степен образуват линейно пространство. Нека разгледаме върху това пространство скаларно произведение зададено във формата:

$$(P, Q) = \sum_{i=1}^{N} P(x_i)Q(x_i).$$

Тук с N голямо сме означили броя на данните.

Определение 10.1 *Казваме, че два полинома са ортогонални ($P \perp Q$), ако* $(P, Q) = 0$.

Теорема 10.1 *Ще построим конструктивно редица от ортогонални полиноми:* $P_0(x) = 1$, $P_1(x) = x - \overline{x}$, *а всички останали (при* $n < N$*) със следната рекурентна формула:*

$$P_n(x) = (x - \alpha_n)P_{n-1}(x) + \beta_n P_{n-2}(x). \tag{10.2}$$

Доказателство: Да отбележим, че $(P_0, P_1) = 0$. Ще покажем първо, как може да се определят числата α_n, β_n.

$$0 = (P_n, P_{n-1}) = (xP_{n-1}, P_{n-1}) - \alpha_n(P_{n-1}, P_{n-1})$$

$$0 = (P_n, P_{n-2}) = (xP_{n-1}, P_{n-2}) + \beta_n(P_{n-2}, P_{n-2})$$

Да отбележим, че от индукцията следва формулата:

$$(xP_{n-1}, P_{n-2}) = (P_{n-1}, xP_{n-2}) = (P_{n-1}, P_{n-1}).$$

От тук получаваме равенствата:

$$\alpha_n = \frac{\sum_{i=1}^{n} x_i P_{n-1}^2(x_i)}{\sum_{i=1}^{n} P_{n-1}^2(x_i)} \tag{10.3}$$

$$\beta_n = -\frac{\sum_{i=1}^{n} P_{n-1}^2(x_i)}{\sum_{i=1}^{n} P_{n-2}^2(x_i)} \tag{10.4}$$

Сега ще покажем, че така получения полином е ортогонален и на всички полиноми с по - ниска степен ($j < n - 2$):

$$(P_n, P_j) = (xP_{n-1}, P_j) = (P_{n-1}, xP_j) = (P_{n-1}, P_{j+1}) = 0.$$

□Така построената редица има смисъл, разбира се, докато степента на полинома е малка по сравнение с броя на данните N. Не е трудно да се провери, че $P_j, j \geq N - 1$ има за корени числата x_i.

Сега моделът (10.1) може да се препише във формата:

$$y_i = b_0 P_0 + b_1 x_i + b_2 P_2(x_i) + \ldots + b_n P_n(x_i) + \epsilon_i. \tag{10.5}$$

Матрицата $X'X$ за този модел е диагонална и съдържа числата $d_{ii} = \sum P_{i-1}^2(x_i)$.

За нашия случай матрицата от стойности на ортогоналните полиноми върху данните е дадена на таблица 10.2.

Коефициентите на ортогоналните полиноми са показани в таблица 10.3 (редовете са степени на полинома (0 - 8), на диагонала е коефициентът пред максималната степен):

Коефициентите на разлагането на отклика b_k по този нов базис са дадени в таблица 10.4.

P_0	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
1	-4	9.3333	-16.8	24.0000	-26.6667	21.8182	-11.7483	3.1329
1	-3	2.3333	8.4	-36.0000	73.3333	-92.7273	70.4895	-25.0629
1	-2	-2.6667	15.6	-18.8571	-26.6667	120.0000	-164.4755	87.7203
1	-1	-5.6667	10.8	15.4286	-60.0000	5.4545	164.4755	-175.4406
1	0	-6.6667	-0.0	30.8571	0.0000	-109.0909	-0.0000	219.3007
1	1	-5.6667	-10.8	15.4286	60.0000	5.4545	-164.4755	-175.4406
1	2	-2.6667	-15.6	-18.8571	26.6667	120.0000	164.4755	87.7203
1	3	2.3333	-8.4	-36.0000	-73.3333	-92.7273	-70.4895	-25.0629
1	4	9.3333	16.8	24.0000	26.6667	21.8182	11.7483	3.1329

Таблица 10.2: Стойности на ортогоналните полиноми

1.								
-4.	1.							
9.333	- 8.	1.						
-16.8	36.2	-12.	1.					
24.0	-124.57	79.5714	-16.	1.				
-26.666	372.88	-393.33	139.44	-20.	1.			
21.818	-1077.82	1664.91	-894.55	215.91	-24.	1.		
-11.748	3277.01	-6623.468	4848.16	-1701.53	309.077	-28.	1.	
3.132	-10953.08	26423.99	-24217.85	11220.28	-2889.6	419.067	-32.	1.

Таблица 10.3: Коефициенти на ортогоналните полиноми

10.3 Оптимална степен

Нека съгласно секция 9 се опитаме да оценим "най - добрия" регресионен полином. Ще построим редицата от подпространства $Z_i, i = 0, 1, 2, \ldots, n$. Подпространството Z_i ще съответствува на полином от степен i и неговата размерност е $i + 1$. Ясно е, че проверката на хипотезата $H_0 : \theta \in Z_{n-1}$ срещу сложната алтернатива $H_1 : \theta \in Z_n \setminus Z_{n-1}$ е еквивалентна на проверката $a_n = 0$ в модела (10.1) или на проверката $b_n = 0$ в модела (10.5), когато, обаче този модел е верен. Това налага да търсим "верната" степен, започвайки отгоре – от максималната възможна степен. Това е $n = N - 2$. За щастие изчислителните формули в модела (10.5) са изключително прости. За да ги изведем, нека въведем норма $||P||^2 = (P, P)$ и означим с $\tilde{P}_k = P_k/||P_k||$, $k = 0, 1, 2, \ldots, N - 1$ ортонормираните полиноми. Тогава N – мерното векторно пространство на наблюденията ($y \in R^N$) се представя в нова координатна система с вектори – стойностите на ортогоналните полиноми:

$$y = \sum_{k=0}^{N-1} \tilde{b}_k \tilde{P}_k, \quad \tilde{b}_k = (y, \tilde{P}_k)$$

$$||y||^2 = \sum_{i=1}^{N} y_i^2 = \sum_{k=0}^{N-1} \tilde{b}_k^2.$$

Последното равенство е всъщност равенството на Парсевал:

$$||y||^2 = \sum_{i=1}^{N} x_i^2,$$

където x_i са координатите на вектора x в който и да е ортонормиран базис. При това връзката между коефициентите е тривиална:

$$\tilde{b}_k = ||P_k||\hat{b}_k.$$

От тези равенства следват и търсените формули:

$$\hat{b}_k = \frac{\sum_{i=1}^{N} y_i P_k(x_i)}{\sum_{i=1}^{N} P_k^2(x_i)}, \tag{10.6}$$

$$\sigma^2(\hat{b}_k) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{N} P_k^2(x_i)}, \tag{10.7}$$

$$\hat{\sigma}^2 = \frac{1}{N-n-1}\sum_{k=n+1}^{N-1} \tilde{b}_k^2 = \frac{1}{N-n-1}\sum_{k=n+1}^{N-1} \hat{b}_k^2 \sum_{i=1}^{N} P_k^2(x_i). \tag{10.8}$$

Тъй като частното:

$$f_n = \frac{\tilde{b}_n^2}{1/(N-n-1)\sum_{k=n+1}^{N-1} \tilde{b}_k^2} = \frac{(N-n-1)\hat{b}_n^2 \sum_{i=1}^{N} P_n^2(x_i)}{\sum_{k=n+1}^{N-1} \hat{b}_k^2 \sum_{i=1}^{N} P_k^2(x_i)}$$

съгласно теорема 7.2 има F - разпределение с 1 и $N-n-1$ степени на свобода, правилото за проверка се свежда до намиране на минималното n, за което е изпълнено неравенството:

$$f_n > F_{\text{кр.}}(1, N-n-1).$$

В таблица 10.4 са изведени данните, необходими за намирането на "оптималния" полином за нашия пример. В последната колона за удобство са поставени критичните стойности за съответното F-разпределение. Вижда се, че максималната статистически значима степен е 2.

От същата таблица се вижда също, че никаква статистическа проверка не е възможна за интерполационния полином от 8 степен. С указаните данни изобщо не е възможна проверка на адекватността на регресионния модел (с полином от втора степен), така че използването му за прогноза едва ли е оправдано.

Числата от последната колона са взети от таблица на квантилите на $F-$разпределение. Те се използват за да се сравнят с тях стойностите

	b_k	$\tilde{b}_k^2$	F-value	Df	$F_{0.95}$
P_0	143.143	184409.3	61.525	1 8	5.32
P_1	18.508	20552.7	287.8528	1 7	5.59
P_2	1.04582	336.87	18.358	1 6	5.98
P_3	.104426	15.546	0.81392	1 5	6.60
P_4	-.07695	34.838	2.52819	1 4	7.70
P_5	-.02555	13.575	0.97661	1 3	10.13
P_6	.00955	5.373	0.479701	1 2	18.51
P_7	.01277	19.31	6.23907	1 1	161.45
P_8	-.00495	3.095			

Таблица 10.4: Оптимална степен

на съответните статистики, дадени в колона 3. Когато става въпрос за програми, пресмятането на квантили е обикновено по - трудно от пресмятането на ф.р. Затова обикновено е автоматизирано пресмятането на вероятността: $\alpha_n = \mathsf{P}(\xi < f_n)$, ($\xi$ е сл.в. със съответното разпределение). Тя носи названието F - probability и така лесно можем да проверим за дадено ниво на доверие (например, $\alpha = 0.95$) дали съответната хипотеза се отхвърля. Правилото за избор на оптимална степен съответно става: $n = \max\{k : \alpha_k > \alpha\}$.

Тема 11

Дисперсионен и ковариационен анализи

Дисперсионният анализ е част от статистиката, изучаваща влиянието на една или няколко групиращи променливи върху една количествена. Както и в регресията, е прието тази зависима променлива да се нарича отклик. Предикторите, обаче тук се наричат фактори. В основата на дисперсионния анализ лежи възможността сумата от квадрати на отклонения на отклика SSY да бъде разложена на няколко независими суми от квадрати, като по този начин става възможна проверката на различни хипотези за влияние на факторите върху отклика.

В дисперсионния анализ е възприето групиращата променлива да се нарича "фактор", стойностите ѝ - "нива" на фактора, а отклоненията на средните стойности на групата от общата средна - "ефекти". Така с всяко ниво на фактора е свързан един ефект. Ако групите са определени от една групираща променлива, казваме, че се извършва "еднофакторен" анализ. Когато факторите са няколко, определянето на групите е по - сложно. Анализът се нарича "многофакторен". При двуфакторния анализ, например, се разглеждат, както прости ефекти, свързани с влиянието на всеки фактор поотделно, така и смесени ефекти. Двете групиращи променливи определят толкова групи, колкото е произведението от броя на нивата на двата фактора. Толкова са на брой и смесените ефекти, които отразяват съвместното влияние на факторите върху отклика. Ако се окаже, че такова съвместно влияние отсъствува, т.е. съвместните влияния са малки, следва да се проверяват за значимост простите ефекти.

Основната задача, която се решава с помощта на дисперсионния анализ, може да се формулира най - просто така: да се провери хипотезата дали съвпадат средните стойности на отклика в няколко различни групи от наблюдения. Ако тази хипотеза се отхвърли, необходимо е да се

оценят различните средни стойности за всяка група. В този случай се казва, че търсим фиксирани ефекти или разглеждаме модел I.

Друг подход в дисперсионния анализ е оценката на така наречените случайни ефекти или модел II. Приема се, че факторът определя ефекти, които са независими, нормално разпределени, със средни стойности нула и дисперсия, една и съща за всички нива на фактора. Хипотезите, които се проверяват при използуване на такъв модел се отнасят до стойността на тази дисперсия. Въпреки че хипотезите за двата модела са различни, статистиките, с които те се проверяват понякога съвпадат - например, при един фактор. При повече фактори нещата се усложняват неимоверно. Ограниченото място не позволява подробното им излагане. При желание читателят може да се запознае подробно с тях в (Шеффе 1963) и по - популярно в (Афифи и Айзен 1982).

Прието е резултатите от дисперсионния анализ да се представят в така наречените таблици на дисперсионния анализ. В тези таблици за всеки прост или смесен ефект се представя съответната сума от квадрати на отклоненията заедно със степените си на свобода. Така, сравнявайки в определен ред нормираните суми от квадрати с критерия на Фишер, може да се получи представа за влиянието на ефектите.

Най-голяма популярност дисперсионният анализ е придобил в областта на селскостопанския експеримент. С негова помощ се изучава влиянието на различни видове торове и почви върху добива при различни природни условия и под въздействието на редица ненаблюдаеми фактори. Това приложение на дисперсионния анализ в област, където отделно взетия експеримент е скъп и продължителен, още при самото му възникване е поставило пред математиците задачата за оптимизиране на броя на провежданите експерименти. Една голяма част от литературата по дисперсионен анализ е посветена на планирането. В решаването на този проблем са привлечени много математически резултати от други области на математиката, а за експериментаторите се публикуват сборници от планове удовлетворяващи широк кръг изисквания, произвеждат се програмни системи генериращи такива планове и т.н.

В много случаи прилагането на дисперсионния анализ е еквивалентно на прилагането на регресионния (например, когато всички групиращи променливи - фактори притежават само по две нива), но даже и в този случай поради вложените в себе си възможности да изучава съвместното влияние на факторите той с лекота отговаря на въпроса, кои фактори и в каква комбинация влияят на отклика.

Често се използуват думите дисперсионен анализ и за редица тестове, провеждани като част от други статистически процедури (вж.например, проверка на адекватност на регресионен модел) и то с пълно основание.

11.1 Основен модел

Математическата литература по дисперсионен анализ е почти необозрима. Това се дължи главно на факта, че в основата му лежи планирането на многофакторни експерименти, тяхното оптимизиране за задачите поставени от експериментатора. Тук ние ще приведем само елементарните формули за еднофакторен експеримент. Анализът на двуфакторен експеримент, даже и с равен брой наблюдения в клетка, се разклонява в зависимост от типа на ефектите - фиксирани и случайни, прости и смесени и т.н. Класическата книга (Шеффе 1963) би представлявала полезно пособие за едно сериозно навлизане в тази област.

Моделът на еднофакторния дисперсионен анализ с фиксирани ефекти се записва като регресионен модел по следния начин:

$$\begin{aligned} y &= Z\mu + e \\ y_{ij} &= m + a_i + \epsilon_{ij}. \end{aligned} \tag{11.1}$$

Тук с a_i сме означили ефектите - влиянията съответстващи на нивата на фактора, а грешките с ϵ - независими случайни величини с разпределение $N(0, \sigma^2)$. Индексите i описват възможните нива на фактора, а j - наблюденията в рамките на едно фиксирано ниво. Ясно е, че ако се опитаме да поставим като предиктори изкуствени вектори състоящи се от нули и единици, тази задача би съвпаднала напълно със задачата на регресионния анализ. Съществува обаче проблем в нейното решаване, тъй като рангът на получената матрица е по - малък от необходимия. Затова се налагат (повече или по - малко естествени) ограничения върху оценяваните параметри. В случая това е ограничението

$$\sum_i a_i = 0. \tag{11.2}$$

Сега вече сме в състояние да извършим оценяване на параметрите на този модел по метода на най - малките квадрати и, (при положение, че имаме достатъчно наблюдения за всяко ниво на фактора) да проверим, например, хипотезата $H_0 : a = 0$.

Съответното разлагане на SSy в този случай изглежда така

$$\sum_i \sum_j (y_{ij} - y_{..})^2 = \sum_i \sum_j (y_{i.} - y_{..})^2 + \sum_i \sum_j (y_{ij} - y_{i.})^2, \tag{11.3}$$

или $SSy = SSm + SSr$. С точки вместо индекси (по традиция в дисперсионния анализ) са означават усредняванията по съответните индекси.

Тук SSr е остатъчната сума от квадрати, а SSm отговаря за влиянието на фактора върху отклика. При изпълнена хипотеза $H_0 : \alpha = 0$ двете събираеми са пропорционални на Хи-квадрат със степени на свобода съответно $N - M$ и $M - 1$ (с M сме означили броя на непразните нива на фактора, а с N – общия брой наблюдения). F статистиката строим по естествената формула

$$F = (SSm/(M-1))/(SSr/(N-M)) \tag{11.4}$$

и отхвърляме хипотезата, ако тя надхвърли критичната стойност на съответното разпределение на Фишер.

Естествено и тук могат да бъдат избрани по-сложни алтернативи от тривиалната – пълен модел. Такава може да бъде например хипотезата: $H_1 : a_1 = -a_2$. При такава проверка ролите на SSm и SSr се заемат от други суми от квадрати. Такива помощни алтернативи се наричат *контрасти*.

11.2 Множествени сравнения

В много случаи ни е необходимо да направим едновременно заключение за много от параметрите наведнаж. Можем да използуваме следното знаменито неравенство на Бонферони:

$$\mathsf{P}(\overline{\cap I_i}) \geq \prod \mathsf{P}(\overline{I_i}). \tag{11.5}$$

Така, ако I_i са доверителни интервали за M параметъра a_i с ниво на доверие $1 - \alpha/M$, то $\cap I_i$ е съвместен доверителен интервал за всичките параметри с гарантирано ниво на доверие $1 - \alpha$. Тези доверителни интервали обаче са твърде неточни (големи). Затова в тази секция ще разгледаме два метода за построяване на съвместни доверителни интервали особено подходящи за линейни модели.

11.2.1 Метод на Тюки

Да разгледаме, например модела (11.1). Да си поставим следните задачи:

1. Да намерим доверителни интервали I_i за параметрите $\beta_i = m + a_i$, такива, че
$$\mathsf{P}(\cap_i \{\beta_i \in I_i\}) \geq 1 - \alpha, \tag{11.6}$$
2. Да намерим доверителни интервали $I_{i,j}$ за параметрите $a_i - a_j$, такива, че
$$\mathsf{P}(\cap_{i<j} \{a_i - a_j \in I_{i,j}\}) \geq 1 - \alpha. \tag{11.7}$$

Ще започнем решението на задача 1 със следната постановка. Нека броят на нивата на фактора е фиксиран M и броят на наблюдения за всяко ниво – еднакъв k. Търси се константа C такава, че да е изпълнено следното равенство:

$$\mathsf{P}(\cap_i\{|\beta_i - y_{i,.}| < Cs\}) \geq 1 - \alpha, \tag{11.8}$$

където $s^2 = SSr/(n-1)k$ е естествената неизместена оценка на дисперсията σ^2. Имаме $SSr/\sigma \in \chi^2_{M(k-1)}$. Оценките на β_i са независими и независими в съвкупност от s. Следователно

$$\mathsf{P}(\cap_i\{|\beta_i - y_{i,.}| < Cs\}) = \mathsf{P}(\max_i(|\beta_i - y_{i,.}|) < Cs)$$
$$= \mathsf{P}(\max_i(\frac{k^{-1/2}|\beta_i - y_{i,.}|}{\sigma}) < Ck^{-1/2}\frac{s}{\sigma}) = \mathsf{P}(\frac{\max_i|\xi_i|}{\eta} < Ck^{-1/2}),$$

Това разпределение зависи само от два параметъра (k, M) и несложно може да се табулира. Пресмятаме от там търсената стойност на C за зададеното ниво на доверие α и така получаваме точен съвместен доверителен интервал:

$$\cap I_i = \cap\{y_{i,.} - Cs, y_{i,.} + Cs\}.$$

Втората задача решаваме аналогично:

$$\mathsf{P}(\cap_{i<j}\{|a_i - y_{i,.} - a_j + y_{j,.}| < Cs\}) = \mathsf{P}(\max_{i<j}(|\beta_i - y_{i,.} - \beta_j| + y_{j,.}) < Cs)$$
$$= \mathsf{P}(\frac{\max_{i<j}|\xi_i - \xi_j|}{\eta} < Ck^{-1/2}).$$

Сега интервалите за проверка имат вида:

$$I_{i,j} = \{y_{i,.} - y_{j,.} - Cs, y_{i,.} - y_{j,.} + Cs\}. \tag{11.9}$$

Двете разпределения, които се използуват в метода на Тюки са табулирани и могат да се намерят, например в (Hartley and Pearson 1966).

11.2.2 Метод на Шефе

При сравненията по двойки използувахме разликите $\beta_i - \beta_j$. Понякога се налага да се сравняват групи параметри. Например, при обработката на почва по 4 различни начина при 2 от тях внасяме азотен тор, а при другите 2 не внасяме. Ясно е, че бихме могли да оценим например контраста (функцията):

$$\phi = 1/2(\beta_1 + \beta_2) - 1/2(\beta_3 + \beta_4).$$

Когато обаче отнапред не знаем къде да търсим разликата, трябва да разполагаме със средство за оценка на значимостта на всички линейни функции от параметрите. Такова средство ни дава метода на Шефе. Той е основан на следното геометрично тъждество:

$$||x|| = \sup_c \frac{c'x}{||c||}.$$

Нека разгледаме един контраст $c = \{c_1, c_2, ..., c_M\}'$ за параметрите a. Да напомним, че $\sum c_i = 0$ и означим $y = c'a = c'\beta$. Тогава

$$\sup_c \frac{y - \hat{y}}{||c||} = \sup_c \frac{c'(a - \hat{a})}{||c||} = ||a - \hat{a}||.$$

Оценките на β_i в разглеждания модел са независими и независими в съвкупност от s. Поради линейното условие върху a имаме

$$k||a - \hat{a}||^2/\sigma^2$$

е Хи-квадрат с $M - 1$ степени на свобода. Следователно

$$f = k||a - \hat{a}||^2/(M - 1)s^2$$

има F разпределение с $M - 1$ и $M(k - 1)$ степени на свобода. Така за всички контрасти y получаваме съвместни доверителни интервали:

$$I_a = \{\hat{y} - C^{1/2}s||c||, \hat{y} + C^{1/2}s||c||\}, \tag{11.10}$$

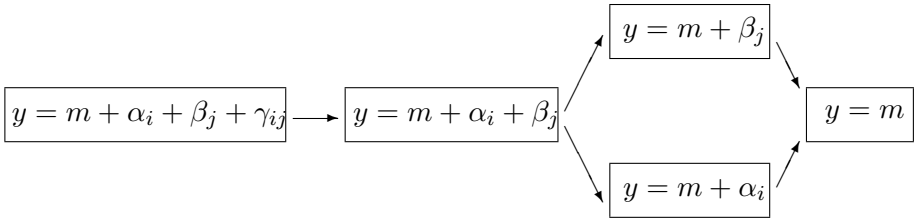
където $C = (M - 1)f_{1-\alpha}k$.

11.3 Двуфакторен анализ

Тук вече можем да избираме измежду няколко възможни модела:

Стрелките показват естествените връзки между моделите, а също и пътя, по който строим и сравняваме нашите хипотези. Прието е, както при полиномната регресия, да започваме от най - сложния модел. Нека разгледаме за пример два такива модела свързани със стрелка:

$$Z_1 : y = m + \alpha_i + \beta_j + \gamma_{ij},$$
$$Z_2 : y = m + \alpha_i + \beta_j;$$



Фиг. 11.1: Връзки между моделите

и за яснота да означим с k и m броя на нивата на факторите: ($i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, m$). Броят на оценяваните параметри в модела Z_1 е равен фактически на броя на клетките определени от всевъзможните комбинации от нива на двата фактора: km. За втория модел този брой е $k+m-1$. За да можем да използуваме модела 8.1 и теоремата 8.1 трябва да е изпълнено неравенството: $km > k+m-1$. Това е винаги така, стига да е изпълнено $2 \leq k, m$.

11.4 Примери

Тук ще разгледаме няколко примера с реални данни заимствани от книгата (Dunn and Clark 1974).

Пример 11.1 *Пример за еднофакторен дисперсионен анализ.*

Целта е да се изучи влиянието на четири типа тор върху добива. За целта 24 еднакви по форма и площ полета са засети с една и съща култура. В дисперсионния анализ се казва, че факторът тор има 4 нива. По случаен начин експериментаторът избрал типът торене върху всяко от полетата, така всеки тип торене се среща 6 пъти. Тези данни трябва да бъдат въведени като две променливи - първата количествена - ДОБИВ и втората - групираща ТОР. Матрицата от данни трябва да изглежда така:

Така въведени данните могат вече да бъдат подложени на дисперсионен анализ. Получаваме следната таблица на дисперсионен анализ:

ДОБИВ	ТОР	ДОБИВ	ТОР	ДОБИВ	ТОР	ДОБИВ	ТОР
99	1	96	2	63	3	79	4
40	1	84	2	57	3	92	4
61	1	82	2	81	3	91	4
72	1	104	2	59	3	87	4
76	1	99	2	64	3	78	4
84	1	570	2	396	3	498	4

Таблица 11.1: Данни за торенето - ANOVA1

Anova 1 Table			
SOURCE OF VARIATION	SUM OF SQUARES	D.F. SQUARE	MEAN
TREATMENT	2940	3	980
RESIDUAL	3272	20	163.6
TOTAL	6212	23	
COMPUTED			
F= 5.99022		P= .995613	

Стойността на F статистиката, както и вероятността P са твърде големи и позволяват с висока степен на доверие да отхвърлим хипотезата, че факторът торене не влияе на добива.

Пример 11.2 *Двуфакторен дисперсионен анализ*

Ще разгледаме още един пример от (Dunn and Clark 1974) показан на таблица 11.2. В него се изучава добива на ръж като функция от типа на семената и торенето. В този случай торенето се избира по три възможни начина: ниско, средно и високо, и се използуват два типа семена. Експериментаторът и в този случай е разполагал с 24 полета и за всяка от шестте възможни комбинации тор - семе е избрал случайно по 4 полета. Естествено е да разглеждаме фиксирани ефекти.

Тези данни трябва да се представят в следната форма. Като променливи се определят: откликът ДОБИВ, и фактори (или групиращи променливи) СЕМЕ и ТОР, като последните съответно се кодират. Началото на получената матрица данни ще изглежда така:

ДОБИВ	СЕМЕ	ТОР	ДОБИВ	СЕМЕ	ТОР
14.3	1	1			
18.1	1	2			
17.6	1	3			

ТИП НА СЕМЕНАТА	НИВО НА ТОРЕНЕ НИСКО	СРЕДНО	ВИСОКО
1	14.3	18.1	17.6
	14.5	17.6	18.2
	11.5	17.1	18.9
	13.6	17.6	18.2
2	12.6	16.5	15.7
	11.2	12.8	17.6
	11	8.3	16.7
	12.1	9.1	16.6

Таблица 11.2: Данни за добива при различно торене и семена

Anova 2 Table			
SOURCE OF VARIATION	SUM OF SQUARES	D.F.	MEAN SQUARE
A	77.4004	1	77.4004
B	99.8725	2	49.9362
A B	44.1058	2	22.0529
RESIDUAL	21.9975	18	1.22208
TOTAL	243.376	23	
Fixed			
	FA	FB	FAB
	63.3348	40.8615	18.0453
	.999999	.999999	.999949
Random			
	FA	FB	
	3.50975	2.26438	
	.798127	.693663	

Таблица 11.3: ANOVA2 -таблица

Получаваме дисперсионна таблица на двуфакторния дисперсионен анализ - таблица 11.3.

От тази таблица заключаваме, че съществува изразено взаимодействие между торенето и типа на семената при влиянието им върху добива -$FAB = 18.0453$, а вероятността .999949 говори, че хипотезата за незначимост на смесените ефекти се отхвърля. След като смесените ефекти на двата фактора са значими, не бива да проверяваме поотделно хипотезите за простите ефекти. Може веднага да се приеме, че влиянието на типа на семената и торенето като цяло върху добива е съществено.

Тъй като този пример не е особено поучителен, не илюстрира пълните възможности на процедурата, ще разгледаме още един пример от областта на психологията.

Пример 11.3 *Данните за скоростта на реакцията на човек при подаване на светлинен* (A, C) *и звуков* (B, D) *сигнали.*

Изучават се два типа реакция: при A и B - реакцията е проста, а при C и D - с избор. Естествено е, да разглеждаме две групиращи променливи. Първата описва типа на сигнала (светлинен или звуков), а втората - условията на експеримента (с или без избор). За да въведем данните в паметта, трябва да ги прекодираме аналогично на предния пример. За тези данни таблицата на двуфакторния анализ изглежда иначе:

Anova 2 Table			
SOURCE OF VARIATION	SUM OF SQUARES	D.F.	MEAN SQUARE
A	123932.	1	123932.
B	5206.24	1	5206.24
A B	62.1323	1	62.1323
RESIDUAL	24495.7	64	382.746
TOTAL	153696.	67	
Fixed			
	FA	FB	FAB
	323.797	13.6023	.162332
	1	.999531	.311639
Random			
	FA	FB	
	1994.65	83.7929	
	.985748	.930728	

Тук вече взаимодействието между факторите отсъствува - статистиката FAB е незначима. По-отделно обаче, влиянието и на двата фактора

е значимо и не може да бъде пренебрегнато. При желание може да се пресметнат оценените вътрешно групови средни стойности при адитивното влияние на двата фактора.

11.5 Ковариационен анализ

Нека разгледаме сега пак регресионния модел със свободен член. Ще включим в модела групираща променлива и нека тя да е една. Ще представим наблюденията върху нея в матрицата Z. Сега моделът приема следната форма:

$$y = Z\mu + Xa + e \tag{11.11}$$

Групиращата променлива приема стойности от 1 до G. Матрицата Z е с размерност $(n \times G)$, като всеки ред е индикатор (съдържа нули и една единица) за групата, на която принадлежи съответното наблюдение. Сега броят на параметрите е вече $m + G$ и разбира се, трябва да бъде изпълнено неравенството $m + G < N$. С μ сме означили вектора от параметри (с размерност G), отразяващ влиянието на стойността на групиращата променлива за дадено наблюдение. Този модел е обект на така наречения ковариационен анализ (виж (Шеффе 1963)), където се предполага зависимост и на вектора a от стойността на групиращата променлива.

При $G = 1$ моделът се свежда към класическа линейна регресия със свободен член. При $G > 1$ формулите за пресмятане претърпяват незначителни изменения. При $a = 0$ параметрите μ очевидно са "вътрешно-груповите"м.о. с естествени си оценки. За да се запази това им качество и при ненулево a, във всички формули по горе матрицата $X'X/n$ трябва да се замести с вътрешно-груповата ковариационна матрица V_i и числото n да се замести с $n - G$.

Тема 12

Дискриминантен анализ

Тук ще представим една процедура от многомерния анализ на данни базирана на вероятностен модел. Другото название на процедури от този тип е разпознаване на образи.

12.1 Основни понятия

Тази статистическа процедура се използува, когато се нуждаем от "прогнозиране" стойностите на групираща променлива. Понякога това се нарича класификация или разпознаване на образи. Нека нашата извадка е неоднородна или с други думи, се състои от няколко групи наблюдения с различни вероятностни характеристики. Целта ни е да се научим от тази извадка, по зададени параметри на дадено наблюдение, да определим принадлежността му към класа, от който произлиза.

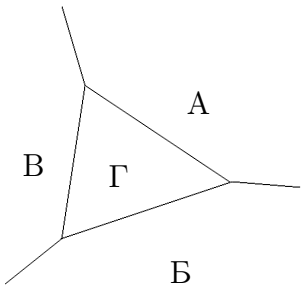
В първата си част, фазата на обучение, процедурата на дискриминантния анализ обработва тази информация с цел да я кондензира в тъй наречените решаващи правила. Когато те са получени, естествено е те да бъдат изпробвани върху обектите от обучаващата извадка или върху други обекти с известен клас. При положение, че тези обекти (или поне достатъчно голям процент от тях) бъдат класифицирани правилно, можем да очакваме, че разпознаващите правила са добри и коректно ще работят и за обекти от неизвестен клас.

Разбира се, конкретното прилагане на различните методики на дискриминантния анализ има ред тънкости. Тук ще се спрем по - подробно на най - разпространената процедура за стъпков линеен дискриминантен анализ. Тя притежава ред недостатъци, но и някои преимущества - в частност дава прости решаващи правила.

В линейния дискриминантен анализ се строят линейни дискрими-

нантни функции от предикторите. За всеки клас има точно една такава функция. Правилото за класификация изглежда така:

Наблюдението се класифицира към класа с максимална дискриминантна функция.



Фиг. 12.1: Класификационни области

Областта от стойности на предикторите, при попадане в която наблюденията се класифицират към даден клас, е изпъкнал многоъгълник (възможно отворен). Тя се нарича класификационно множество на класа. При два предиктора и 4 класа класификационните множества биха могли да изглеждат по указания начин. На фигура 12.1 с големи букви са отбелязани груповите средни.

Когато класовете са ясно разграничени, не е трудно те да бъдат отделени. Когато обаче те се пресичат, въпросът за оптимален (с най - малко грешки) избор на класификационни правила е сложен и изисква допълнителна априорна информация.

Линейният дискриминантен анализ предполага, че разпределенията на количествените променливи (предикторите) вътре в класовете са нормални и се различават само по средните си стойности. Тогава процедурата произвежда оптимални решаващи правила. Разбира се, тя може да се използува и при случайна (по групиращата променлива) извадка, но появата на празен клас е недопустима.

Когато броят на количествените променливи е по - голям, за простотата на решаващите правила е съществено да се отберат за предиктори само тези променливи, които носят важната за разделянето информация. В това помага статистиката на Махалонобис. Тя позволява да се провери хипотезата за съвпадане на груповите средни на предикторите като цяло. За простота и тук, вместо критичната област за статистиката, се използува вероятността съответно разпределената случайна величина да не надхвърли стойността на статистиката. Тази вероятност расте

докато променливите допринасят за по - доброто разделяне на класовете и започва да намалява, когато предикторите станат твърде много. Естествено, добро разделяне може да се очаква, само когато хипотезата се отхвърля с висока вероятност.

Нека вече са избрани най - добрите променливи за предиктори. Това още не означава, че са построени класификационните множества. Да напомним, че основна цел на дискриминантния анализ е да се получи правило за причисляване на едно ново наблюдение към даден клас. За това наблюдение може да съществува априорна информация за неговата възможна принадлежност към класовете. Прието е такава информация да бъде формулирана в термини на априорни вероятности, които са необходими за определяне на оптимални класификационни правила.

Ако такава информация не съществува, естествено е априорните вероятности на класовете да се приемат за равни. Когато пък извадката е случайна и новото наблюдение се избира по същия начин, може те да се приемат за пропорционални на обема на класовете в обучаващата извадка.

Изборът на априорните вероятности фиксира оптимални дискриминантни функции и класификационни правила. Не е удобно обаче, за всяко ново наблюдение да се въвеждат априорни вероятности. Това е свързано и със значителни изчислителни трудности, особено когато броят на класовете е голям. Един възможен начин за заобикаляне на това неудобство е представянето на класовете с помощта на няколко групиращи променливи. Такова представяне съответствува и на редица практически задачи. Ако са известни стойностите на поне една от групиращите променливи, това е съществена априорна информация - фиксирането на тази променлива е еквивалентно на задаването на нулева априорна вероятност за поне половината от класовете.

12.2 Вероятностна формулировка

12.2.1 Бейсов подход

Нека допуснем, че вероятностите $\{p(g)\}$, груповите средни $\{m(g)\}$ и вътрешно - груповата ковариационна матрица $C(g) = C, g = 1, 2, \ldots, G$, са известни. Тогава по формулата на Бейс, апостериорната вероятност за класификация в класа g на наблюдението $(x, .)$ ще бъде

$$q(g) = c.p(g).f(x, m(g), C). \tag{12.1}$$

Тук f е плътността на нормалното разпределение със средна стой-

ност $m(g)$ и ковариационна матрица C, а c е нормираща константа (такава, че $\sum q(g) = 1$).

12.2.2 Класификационните правила

Съгласно принципа за максимално правдоподобие, класифицира се по правилото:

$$\hat{g} = \max h : q(h). \tag{12.2}$$

Класификационните правила могат да бъдат записани във вида:

$$p(\hat{g}).f(x, m(\hat{g}), C) \geq p(h).f(x, m(h), C), h = 1, 2, \ldots, G, \tag{12.3}$$

за които след логаритмуване и съкращаване, получаваме:

$$b(\hat{g})'x + a(\hat{g}) \geq b(h)'x + a(h), h = 1, 2, \ldots, G, \tag{12.4}$$

Векторът $b(g)$ и числото $a(g)$ се получават по формулите:

$$b(h) = m(h)'C^{-1}, \qquad a(h) = logp(h) - m(h)'C^{-1}m(h) \tag{12.5}$$

Оттук се вижда, че в неравенствата (12.4) участвуват линейни функции относно променливите и това обстоятелство е дало името на линейния дискриминантен анализ.

12.2.3 Априорни вероятности. Модели

За оценка на априорните вероятности $\{p(g)\}$ можем да използуваме най - добрите им оценки $\{n(g)/N\}$ при случайна извадка или друга априорна информация. За оценка на $\{m(g)\}$ и C се използуват вътрешно - груповите средни и обединената извадъчна вътрешно - групова ковариация.

Когато групиращите променливи са повече от една, броят на класовете G нараства. Вероятността за поява на празни клетки $(n(g) = 0)$ при случайна извадка с ограничен обем рязко се увеличава. Затруднява се и оценката за $\{m(g)\}$. В такива случаи се препоръчва използуването на оценки, получени от линеен модел, като се направят съответните проверки с методите на дисперсионния анализ. Съответно, ще се промени и оценката за C. Аналогично, за оценяване на честотите $n(g)$ могат да се прилагат тъй наречените логаритмично - линейни (log - linear) модели.

12.3 Стъпков дискриминантен анализ

Аналогично на стъпковия регресионен анализ и тук е възприета концепцията за избор на подходящ набор от количествени променливи, с които да построим модела. Единственото средство, което ни трябва са аналозите на P(F-to-enter) и P(F-to-remove). Те се строят аналогично на регресията, но ролята на сумите от квадрати играят
- вътрешно - груповата ковариационна матрица
- между - групова ковариационна матрица.

Както и в едномерния случай, така и в многомерния е верно следното равенство:

$$\sum_i \sum_j (x_{ij} - \bar{x})(x_{ij} - \bar{x})' =$$
$$\sum_i \sum_j (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)' + \sum_i n_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})'$$
$$SS = SS_{in} + SS_{mod}. \qquad (12.6)$$

Матрицата SS_{im} се тълкува като "сума от квадрати", отговаряща на разсейването на данните около техните локални средни и с нейна помощ се строи оценка за вътрешно - груповата ковариационна матрица C:

$$\hat{C} = \frac{1}{N - G + 1} SS_{in}.$$

Лесно се вижда, че по диагонала на матричното равенство (12.6) стои добре известното ни разлагане на сумата от квадрати в дисперсионния анализ.

Матрицата SS_{in} се тълкува като "сума от квадрати", отговаряща на разсейването на груповите средни и с нейна помощ се строи оценка за между - груповата ковариационна матрица C_{mg}:

$$\hat{C}_{mg} = \frac{1}{G - 1} SS_{mg}.$$

Точно изменението на детерминантата на тази матрица ни служи за критерий при избора на нова променлива за въвеждане в модела или за нейното отстраняване.

Тема 13

Критерии за съгласие

Тези методи са основани изцяло на свойствата на порядковите статистики и извадъчната функция на разпределение. В тази лекция ще се заемем по - подробно със свойствата на тези статистики, по специално:

- ще докажем теоремата на Гливенко - Кантели;
- ще опишем критерия на Колмогоров - Смирнов;
- ще опишем χ^2 критерия за съгласие;

13.1 Теорема на Гливенко-Кантели

Фиг. 13.1: Извадъчна ф.р. на 30 наблюдения

Имаме $\mathbf{E}F_n(x) = F(x), \quad \mathbf{D}F_n(x) = F(x)(1-F(x))/n \to 0$. Следователно, $F_n(x)$ клони към теоретичната $F(x)$ за всяко фиксирано x (вж.фиг.1.4). От закона на големите числа следва и че $F_n(x) \xrightarrow{\text{п.с.}} F(x)$.

(Докажете го). Верно е обаче още по - силното твърдение на Гливенко – Кантели:

Теорема 13.1

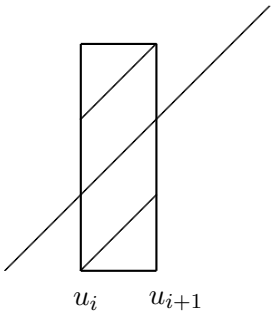
$$\mathsf{P}(\lim_{n\to\infty} \sup_x |F_n(x) - F(x)| = 0) = 1 \tag{13.1}$$

Доказателство: Първо да отбележим, че без намаление на общността можем да считаме $F(x)$ строго растяща функция. Ще докажем теоремата при допълнителното ограничение, че тя няма скокове - т.е. е непрекъсната. Тогава тя притежава обратна функция $F^{-1}(u)$, определена на интервала (0,1).

Сега твърдението на теоремата може да се препише в следната форма:

$$D_n = \sup_{0<u<1} |G_n(u) - u| \xrightarrow{\text{п.с.}} 0,$$

където $G_n(u) = F_n(F^{-1}(u))$. Да фиксираме $\epsilon = 1/r$ и разделим интервала (0,1) на r равни части.



Фиг. 13.2: Един интервал

На фигура 13.2 е показан един такъв интервал. С удебелена линия е отбелязана "целевата" функция u. Границите на интервалите да означим с $u_i = i/r$. Тъй като във всяка от точките u_i $G_n(u_i)$ клони п.с. към u_i, можем да подберем N така, че $\forall n > N$ да е изпълнено неравенството:

$$\mathsf{P}(\sup_i |G_n(u_i) - u_i| > \epsilon) < \epsilon.$$

Сега остава да отбележим, че G_n е монотонна, и ако в краищата на интервала $|G_n(u_i) - u_i| \leq \epsilon$, то вътре в интервала $|G_n(u) - u| \leq 2\epsilon$.

Така получаваме:

$$\mathsf{P}(\sup_u |G_n(u) - u| > 2\epsilon) < \epsilon, \quad \forall n > N.$$

Това означава, че редицата $D_n \xrightarrow{\text{п.с.}} 0$. □Така с помощта на тази теорема получаваме лесно силна състоятелност на много статистически оценки.

13.2 Критерий на Колмогоров - Смирнов

От доказателството на теорема 13.1 е ясна пълната независимост на поведението на редицата D_n от вида на функцията на разпределение $F(x)$, стига да е непрекъсната. Твърдения от този вид се наричат *свободни от разпределение* (distribution free). С тяхна помощ се строят редица статистически критерии.

Една от основните задачи на статистиката е проверката на съответствие на теоретичното разпределение с извадката. Това е нещо като проверка на адекватността на модела – нашите предположения за разпределението. Следната теорема ни дава средства за това:

Теорема 13.2 *(Колмогоров - Смирнов) За всяка непрекъсната функция на разпределение $F(x)$ е в сила съотношението:*

$$\mathsf{P}(\sqrt{n}D_n < x) \longrightarrow K(x),$$

където функцията на разпределение $K(x)$ се представя във вида:

$$K(x) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2x^2}, \quad 0 < x.$$

Доказателство: Виж например (Димитров и Янев 1990).□

Функцията е $K(x)$ табулирана и това прави използуването на критерия сравнително леко. От таблицата се извлича квантила $k_\alpha : K(k_\alpha) = 1 - \alpha$ за необходимото критично ниво α. От данните и теоретичната ф.р. $F(x)$ се пресмята статистиката $\sqrt{n}D_n$ и се сравнява с k_α. Ако я надхвърли, хипотезата за съгласие се отхвърля.

13.3 χ^2-критерий

В много случаи данните са представени във вид на хистограма или са групирани в определени категории – така, например, те се дават в

статистическия годишник на Република България. За да можем и за такива данни да проверяваме съгласие с дадено теоретично разпределение се използува т.н. χ^2-критерий.

Фиг. 13.3: Хистограма на 100 нормални наблюдения

Нека ни е зададена някаква теоретична плътност $p(x)$ и имаме за задача да проверим съгласуваността й с извадката. Нека разполагаме с n наблюдения. Разделяме множеството от стойности на сл.в. (или носителя на плътността) на k интервала H_i, така че $p_i = \int_{H_i} p(x)dx > 0$ и $np_i > 5$. Това изискване е важно. Понякога се налага някои интервали да се обединяват, за да може то да се удовлетвори. Съществува и вариант на хистограмата (и критерия), когато интервалите се избират равновероятни, т.е. $p_i = p_j$. Нека означим с n_i броя на наблюденията попаднали в H_i. Пресмятаме статистиката

$$h = \sum_{i=1}^{k} \frac{(np_i - n_i)^2}{np_i}. \tag{13.2}$$

Теорема 13.3 *(Пирсън) Статистиката h има асимптотично (при $n \to \infty$) разпределение χ^2 с $k-1$ степени на свобода.*

Доказателство: Строгото доказателство е твърде трудоемко. Затова тук ще покажем само идеята. Всяко от събираемите в (13.2) представлява квадрата на центрирана асимптотично нормална сл.в. Действително, $np_i = \mathbf{E}\, n_i$. За съжаление, тези величини са зависими – $\sum n_i = n$. Оказва се, че условното разпределение на сл. гаусов вектор $\xi \in N(0, I)$ в R^n при

условие $(\xi, 1) = 0$ е същото като асимптотичното съвместно разпределение на сл.в. n_i, съответно центрирани и нормирани. $\square$

Тема 14

Оценка на плътности

Както видяхме в предишните лекции много важни за статистическите изводи са качествата на изследваната плътност на разпределение. В тази лекция ще разгледаме накратко най-разпространените методи за непараметрична оценка на плътности. Думата непараметрична използуваме за да подчертаем, че няма да използуваме някое известно семейство разпределения като, например, гаусовото или Гама разпределенията. За такива семейства задачата се свежда до оценка на неизвестните параметри по данните и се решава с методите на точково оценяване.

14.1 Криви на Пирсън

Кривите на Пирсън са всъщност пак семейство от разпределения, но с 4 параметъра. Методът се основава на семейството от плътности удовлетворяващи следното диференциално уравнение:

$$\frac{dp(x)}{dx} = \frac{x-a}{b_0 + b_1 x + b_2 x^2} p(x) \tag{14.1}$$

В зависимост от типа на корените $a_1 \leq a_2$ на полинома в знаменателя $P(x) = b_0 + b_1 x + b_2 x^2$, получаваме 12 различни типа плътности. Всичките са унимодални. В таблица 14.1 ще покажем най-важните 7 типа. Останалите 5 се получават като частни случаи от тях.

Коефициентите в уравнението (14.1) се определят еднозначно от първите 4 момента на разпределението. Това дава възможност, замествайки теоретичните с извадъчните моменти и решавайки уравнението, да получим смислена оценка на плътността, тъй като тези първи четири момента - м.о., дисперсията, асиметрията и ексцеса - доста прилично описват формата на гладко унимодално разпределение.

Тип	Параметри	Плътност	Ограничения	Пример
	$b_1 = b_2 = 0$	$ce^{\frac{1}{2}\frac{(x+a)^2}{b_0}}$	$b_0 < 0$	Нормално
I	$b_2 > 0, a_1 \neq a_2$	$c(1+\frac{x}{a_1})^{p_1}(1-\frac{x}{a_2})^{p_2}$	$-a_1 < x < a_2, -1 < p_1, p_2$	Бета
II	$b_2 > 0, -a_1 = a_2 = \alpha$	$c(1-\frac{x^2}{\alpha^2})^p$	$\|x\| < a, p > -1/2$	Равномерно
III	$b_2 = 0, b_1 \neq 0$	$c(1+\frac{x}{a})^p e^{-\mu x}$	$-a < x < \infty, 0 < \mu, -1 < p$	Гама, χ^2
IV	$b_2 \neq 0, P(x) > 0$	$c(1+\frac{x^2}{a^2})^p e^{-\mu arctg(\frac{x}{a})}$	$0 < a, 0 < \mu, p < -1/2$	
V	$P(x) = c(x-\alpha)^2$	$cx^{-p}e^{\frac{a}{x}}$	$0 < x, 0 < a, 1 < p$	от тип III
VI	$b_2 > 0, a_1 \neq a_2$	$c(1+\frac{x}{a_1})^{p_1}(1-\frac{x}{a_2})^{p_2}$	$a_2 < x, -1 < p_2, p_1 + p_2 < -1$	Фишер
VII	$b_1 = 0, b_0 b_2 > 0$	$c(1+\frac{x}{a})^{-p}$	$p > 1/2$	Стюдент

Таблица 14.1: Криви на Пирсън

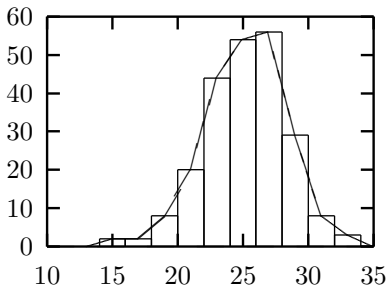
Хубавото на кривите на Пирсън е, че сред тях са и повечето използувани в теорията на статистиката разпределения: гаусовото, Гама, Бета, Фишер, Стюдент, равномерно и др.

Подробно описание на типовете криви на Пирсън и методите за оценка на параметрите им може да се намери у (Поллард 1982), (Митропольский 1964)

14.2 Изглаждане на хистограми

Когато апроксимирането с 4 параметъра не е достатъчно, се прибягва до истински непараметрични методи. Най-лесно това става чрез подходящо изглаждане на хистограмата или извадъчната функция на разпределение.

Най-лесно е простото свързване на средите на стълбчетата на хистограмата. За крайните стълбове се прави отстъп с по половин интервал. Естествено по-гладка крива би се получила при "свързване" с помощта на така наречените *сплайн - функции*. Това са криви, които във всеки интервал са полиноми, но така се слепват в краищата, че обеспечават освен равенство на стойностите си, равенство и на производните си. Най-разпространени са кубичните сплайни.



Фиг. 14.1: Изгладена хистограма

14.3 Ядра на Розенблат - Парзен

Да означим с $\{x_1, x_2, \ldots, x_n\}$ независимите наблюдения на сл.в. с плътност $f(x)$. Непараметричните ядрени оценки се задават във формата:

$$\hat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K(\frac{x_i - x}{h_n}), \tag{14.2}$$

където $K(x)$ е подходящо избрана фиксирана гладка плътност, наричана ядро: $K(x) \geq 0, K(-x) = K(x), \int K(x)dx = 1, \int x^2 K(x)dx = 1. \int K^2(x)dx < \infty$. Често се използува гаусово ядро. Редицата от константи h_n трябва да клони към нула, но така че $nh_n \to \infty$.

Всички анализи на асимптотичното поведение на оценката f_n във фиксирана точка x_0 се основават на развитието в ред на Тейлор на плътността f около тази точка:

$$f(x) = f(x_0) + \sum_{i=1}^{k} \frac{f^{(i)}(x_0)}{i!}(x - x_0)^i + o(|x - x_0|^k) \tag{14.3}$$

Разбира се, то има смисъл, ако съществуват производните на неизвестната плътност f до ред k в точката x_0. Като поставим $x - x_0 = yh_n$ и използуваме (14.3), получаваме, че изместването B_n на оценката е

$$B_n = \mathbf{E}\,\hat{f}_n(x_0) - f(x_0) = \int K(y)(f(x_0 + y * h_n) - f(x_0))dy =$$

$$f'(x_0)h_n \int yK(y)dy + f''(x_0)\frac{h_n^2}{2} \int y^2 K(y)dy + \ldots = O(h_n^2)$$

От друга страна дисперсията на тази оценка (като сума на независими сл.в.) може да се оцени така:

$$D_n = \mathbf{D}\,(\hat{f}_n(x_0)) = \frac{f(x_0)}{nh_n} \int K^2(y)dy + o(\frac{1}{nh_n}) = O(\frac{1}{nh_n})$$

Така като използуваме равенството

$$\mathbf{E}\,(\hat{f}_n(x_0) - f(x_0))^2 = D_n + B_n^2 = O(\frac{1}{nh_n}) + O(h_n^4), \qquad (14.4)$$

получаваме, че асимптотично оптимален избор за константата h_n се получава при $h_n = cn^{-1/5}$.

Приложение A

Таблици

$df \vert p$	.005	.01	.025	.05	.10	.90	.95	.975	.99	.995
1	.00004	.00016	.00098	.0039	.0158	2.71	3.84	5.02	6.63	7.88
2	.0100	.0201	.0506	.1026	.2107	4.61	5.99	7.38	9.21	10.60
3	.0717	.115	.216	.352	.584	6.25	7.81	9.35	11.34	12.84
4	.207	.297	.484	.711	1.064	7.78	9.49	11.14	13.28	14.86
5	.412	.554	.831	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	.676	.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.6	5.23	6.26	7.26	8.55	22.31	25	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
120	83.85	86.92	91.58	95.70	100.62	140.23	146.57	152.21	158.95	163.64

Таблица A.1: Хи-квадрат разпределение (квантили)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5190	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7157	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7969	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8513	.8554	.8577	.8529	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9215	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9492	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Таблица А.2: Нормално разпределение

	.60	.70	.80	.90	.95	.975	.99	.995
1	.325	.727	1.367	3.078	6.314	12.706	31.821	63.657
2	.289	.617	1.061	1.886	2.920	4.303	6.965	9.925
3	.277	.584	.978	1.638	2.353	3.182	4.541	5.841
4	.271	.569	.941	1.533	2.132	2.776	3.747	4.604
5	.267	.559	.920	1.476	2.015	2.571	3.365	4.032
6	.265	.553	.906	1.440	1.943	2.447	3.143	3.707
7	.263	.549	.896	1.415	1.895	2.365	2.998	3.499
8	.262	.546	.889	1.397	1.860	2.306	2.896	3.355
9	.261	.543	.883	1.383	1.833	2.262	2.821	3.250
10	.260	.542	.879	1.372	1.812	2.228	2.764	3.169
11	.260	.540	.876	1.363	1.796	2.201	2.718	3.106
12	.259	.539	.873	1.356	1.782	2.179	2.681	3.055
13	.259	.538	.870	1.350	1.771	2.160	2.650	3.012
14	.258	.537	.868	1.345	1.761	2.145	2.624	2.977
15	.258	.536	.866	1.341	1.753	2.131	2.602	2.947
16	.258	.535	.865	1.337	1.746	2.120	2.583	2.921
17	.257	.534	.863	1.333	1.740	2.110	2.567	2.898
18	.257	.534	.862	1.330	1.734	2.101	2.552	2.878
19	.257	.533	.861	1.328	1.729	2.093	2.539	2.861
20	.257	.533	.860	1.325	1.725	2.086	2.528	2.845
21	.257	.532	.859	1.323	1.721	2.080	2.518	2.831
22	.256	.532	.858	1.321	1.717	2.074	2.508	2.819
23	.256	.532	.858	1.319	1.714	2.069	2.500	2.807
24	.256	.531	.857	1.316	1.708	2.060	2.485	2.787
25	.256	.531	.856	1.316	1.708	2.060	2.485	2.787
26	.256	.531	.856	1.315	1.706	2.056	2.479	2.779
27	.256	.531	.855	1.314	1.703	2.052	2.473	2.771
28	.256	.530	.855	1.313	1.701	2.048	2.467	2.763
29	.256	.530	.854	1.310	1.697	2.042	2.457	2.750
30	.256	.530	.854	1.310	1.697	2.042	2.457	2.750
40	.255	.529	.851	1.303	1.684	2.021	2.423	2.704
60	.254	.527	.848	1.296	1.671	2.000	2.390	2.660
120	.254	.526	.845	1.289	1.658	1.980	2.358	2.617
∞	.253	.524	.842	1.282	1.645	1.960	2.326	2.576

Таблица А.3: Разпределение на Стюдент (квантили)

Литература

Dunn, O. and V. Clark (1974). *Applied Statistics. Analysis of variance and regression.* John Wiley & S.Inc. 11.4, 11.4

Hartley, H. and E. Pearson (1966). *Biometricca Tables for Statisticians. vol.I, 3rd Edition, 1966 vol.II, 1973.* Cambridge: Cambridge University Press. 11.2.1

Афифи, А. и С. Айзен (1982). *Статистический анализ. Подход с использованием ЭВМ.* Москва: Мир. 11

Въндев, Д. и П. Матеев (1988). *Статистика с Правец.* София: Наука и изкуство. (document)

Димитров, Б. и Н. Янев (1990). *Теория на вероятностите и математическа статистика.* София: Наука и изкуство. (document), 13.2

Дрейпер, Н. и Г. Смит (1973). *Прикладной регресионный анализ.* Москва: Статистика. (document), 8

Митропольский (1964). *Техника статистических вычислений.* Москва: Наука. 14.1

Н. Янев, М. Т. (1989). *Ръководство за упражнения по математическа статистика.* София: Софийски Университет "Кл.Охридски". (document)

Поллард, Д. (1982). *Справочник по вычислительным методам статистики.* Москва: Финансы и статистика. 14.1

Проданова, К. (1998). *Въведение в статистическите методи, Част Първа - Количествени методи.* София: Сиела. (document)

Себер, Д. (1976). *Линейный регресионный анализ.* Москва: Мир. 8

Уилкс, С. (1967). *Математическая статистика.* Москва: Наука. (document)

Шеффе, Г. (1963). *Дисперсионный анализ.* Москва: ГИЗ Физ.Мат.Лит. 11, 11.1, 11.5

$(\Omega, \mathfrak{A}, \mathsf{P})$	-	Вероятностно пространство;
Ω	-	Множество от *елементарни събития*; достоверно събитие;
$\mathfrak{A}$	-	σ-алгебра от подмножества на Ω;
$\mathsf{P}(.)$	-	Вероятност определена на $\mathfrak{A}$;
$\mathsf{P}(.)$	-	Вероятност определена на $\mathfrak{A}$;
$A, B, \ldots, Z$	-	Множества,*събития* (елементи на $\mathfrak{A}$) или матрици;
$\overline{A}$	-	Допълнение на множеството; противоположно събитие;
$\xi, \eta, \ldots, \zeta$	-	*случайни величини (сл.в.)*;
γ, ν	-	*измерими разделяния*; *пълни групи от събития*;
$\emptyset$	-	Празно множество; невъзможно събитие;
$A \bigcap B$	-	сечение на множествата A и B; сбъдват се и двете събития;
$A \bigcup B$	-	обединение на множествата A и B; сбъдва се поне едното събитие;
$A + B$	-	обединение на несъвместими събития; сума на матрици;
AB	-	сечение на множествата A и B; произведение на матрици;
$A \perp B, \xi \perp \eta$	-	независими събития и сл.в.;
$\mathbf{E}, \mathbf{D}$	-	Математическо очакване и дисперсия;
$f(\cdot) : A \longrightarrow B$	-	Функция, дефинирана в множеството A със стойности в множеството B;
$R = R^1$	-	Реалната числова права;
R_+	-	Неотрицателните реални числа;
$A \times B$	-	Декартово произведение на множества;
$x = (x^1, \ldots, x^n)'$	-	n-мерен вектор (точка в) R^n;
$\|\|x\|\|$	-	Норма на $x \in R^n$;
$x'y$	-	Скаларно произведение на вектори;
$\exists$	-	Знак означаващ "съществува";
$\forall$	-	Знак означаващ "за всяко".
$\bar{x}$	-	$(1/n)\sum_{i=1}^{n} x_i$
$s^2(x)$	-	$(1/(n-1))\sum_{i=1}^{n}(x_i - \bar{x})^2$
п.с.	-	почти сигурно, с вероятност 1
$f(x, \theta)$	-	функция на правдоподобие, плътност на наблюдавана сл.в.
$LL(x, \theta)$	-	$\log f(x, \theta)$.

Списък на таблиците

Списък на илюстрациите