Софийски Университет Св.Климент Охридски
Факултет по математика и информатика
Вероятности, Операционни изследвания и Статистика

доц. ДИМИТЪР Л. ВЪНДЕВ

Записки
по
Приложна статистика 1

СОФИЯ, юни, 2003

Съдържание

Предговор

Тези няколко лекции са предназначени за студентите от специалности информатика и приложна математика към ФМИ. Тъй като тези студенти не са имали възможност да слушат лекциите ми по (математическа) статистика, много от темите там ще бъдат повторени.

Записките имат за задача да дадат известна представа за модерните методи за анализ на данни. В голяма степен те са базирани на книжката (Вындев и Матеев 1988), която обаче трудно може да се намери.

Всяка глава (лекция) е снабдена и с вспомагателна литература за допълнително изучаване при наличие на необходимост и допълнителен интерес. Много от примерите са изработвани с пакета Statistica, който в момента е най-разпространен.

Авторът се извинява за многото непълноти и грешки и ще бъде много признателен на всеки, който си направи труда да му ги укаже.

Тема 1

Данни и тяхното представяне

1.1 Въведение

Ще се опитаме да определим проблемите, с които се занимава приложната статистика. Отчасти това са проблеми на всяко приложение на научните постижения в практиката.

Науката е призвана да помага на хората да решават по-добре задачите, които възникват пред тях.

Под тези хубави думи за съжаление стои едно голямо недоразумение - кое е по-хубаво се определя по един начин в сферата на научните оценки, и по друг в сферата на приложенията.

Тъй като тези лекции са предназначени за студенти - математици, а и аз имам донякъде такова образование и мислене, ще се опитам да поставя проблемите в рамките на математиката. Това означава, че ще игнорираме по-голямата част от тях - тези, на които няма логичен отговор.

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 Националният статистически институт (НСИ), в САЩ - Central Statistical Office.

1.2 Произход на данните

Основно понятие в статистиката е понятието генерална съвкупност или популация.

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

За едно изследване на НСИ 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.

1.2.1 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.

1.2.2 Извадки

Основна цел е по даден непълен обем данни да се направи някакво правдоподобно заключение за генералната съвкупност като цяло.

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.

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

Извадките биват систематични, случайни или подходящи за целите на изследването комбинации от двата метода.

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.

Когато обемът на извадката е сравним с този на генералната съвкупност, се налага да различаваме и извадки с връщане и без връщане.

1.2.3 Планиране на експеримента

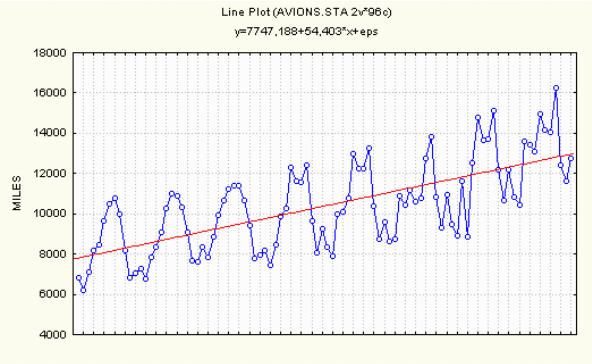
В селското стопанство и техниката често възниква задачата да максимизираме добива или оптимизираме даден производствен процес.

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.

1.2.4 Времени 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.



Фиг. 1.1: Временен ред

В случая това са разстоянията в 1000 мили, изминати от самолетите на Обединеното Кралство за един месец в периода от 1963 до 1970 г.

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.

1.3 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.

1.3.1 Скали на измерване

Данните (наблюденията), с които разполагаме, касаят определени признаци на обектите, които изследваме. Например ако искаме да направим изследване за средната височина на мъжете и жените в България, ще искаме да разполагаме с измервания на височината на представители от двата пола. Конкретните измервания представляват

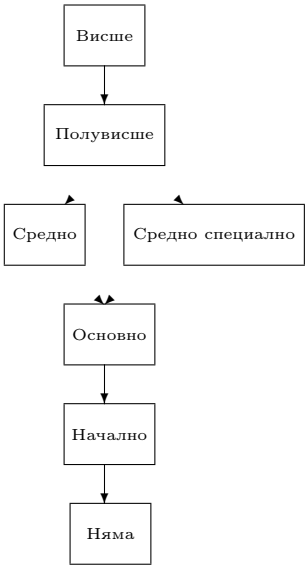
наблюденията, признакът е височината, а обектите с които работим са представителите от двата пола.

Съществуват различни типове признаци. В горния пример, признаците са от числов тип – височината се измерва с число. Има и нечислови признаци, които се наричат още категории.

Обикновено числовите признаци са податливи на анализ. Изследването на нечислови признаци има редица специфични особености.

Категориите могат да бъдат ненаредени – например, цвят или наредени – например, оценка - слаб, среден, добър...

Има и по-сложни ситуации, Ако, например, разгледаме нечисловия признак "образование", то връзката между различните видове образование може да бъде визуализирана както на Фиг. 1.2.



Фиг. 1.2: Образование

В разглеждания пример ситуацията е сравнително добра, понеже върху множеството от различните варианти за образование почти имаме пълна наредба. Често обаче имаме работа със случаи, в които частичната наредба обхваща много по-малка част от наблюдаваните признаци на обектите. Но дори и за ситуации като показаната, където стойностите на признаците са почти наредени, практически не разполагаме със статистически методи специализирани за работа с подобни обекти.

1.4 Кондензиране на данните

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. Ще разгледаме накратко следните методи за кондензиране.

- Таблици и групиране;
- Числови характеристики;
- Графично представяне.

Главна роля в това кондензиране на информация има графичното представяне. То е ефектно и в минимална степен при него се губи информация.

1.4.1 Таблици и групиране

		ПРИЗНАЦИ		
		пр.1	...	*пр.k*
НАБЛЮДЕНИЯ	*№1*	1.32	...	546
	№2	5.46	...	564
	...	...	...	...
	№n	4.89	...	489

Таблица 1.1: n наблюдения върху k признака

Обикновено данните, с които разполагаме могат да се онагледят в таблица по този начин.
Когато наблюденията са много, това става с окрупняване - сумиране и групиране.

Прекрасни примери за това могат да се видят в книгата (**?**) или във всеки статистически годишник издаван от НСИ.

1.4.2 Описателни статистики

1.4.2.1 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.

Така агрегираната числова информация за наблюденията на нечислов признак с k стойности е наборът от честоти (пропорции) на поява:

$$\hat{p}_1 = \frac{n_1}{n}, \hat{p}_2 = \frac{n_2}{n}, \ldots, \hat{p}_k = \frac{n_k}{n}.$$

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 $B(n, p_i)$.

Това е и първият намек за необходимостта от изучаване и прилагане на стохастични (вероятностни) методи в статистиката

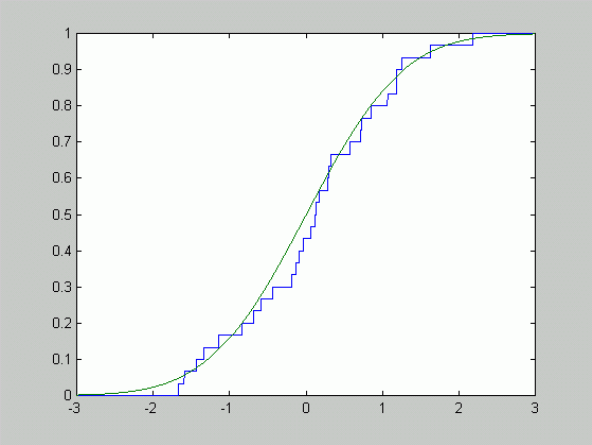
1.4.2.2 Количествени данни

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:*



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

Фиг. 1.3: Извадъчна ф.р.

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. От теоремата на Гливенко-Кантели следва, че когато са изпълнени предположенията на стохастичния модел извадъчните моменти μ_k са "състоятелни" оценки на моментите на сл.в. $\mathbf{E}\,\xi^k$, т.е. те клонят към тях при увеличаване на обема на извадката.

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.

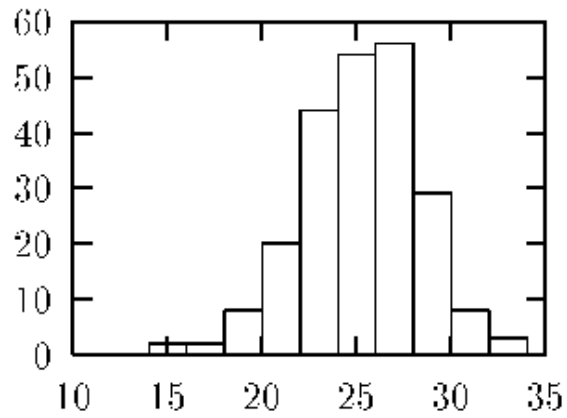
Определение 1.6 *Извадъчен квантил q_α с ниво α на дадена извадка с ф.р. F_n се определя като приближено решение на уравнението: $F_n(q_\alpha) = \alpha$.*

Така медианата $\mu = q_{1/2}$.

1.4.3 Графични методи

Хистограмата е основният вид за представяне на информацията за наблюдения върху числов признак.

Тя се строи по просто правило. Избират се обикновено не много на брой (5 - 20) еднакво големи прилежащи интервала покриващи множеството от стойности на наблюдавания признак.

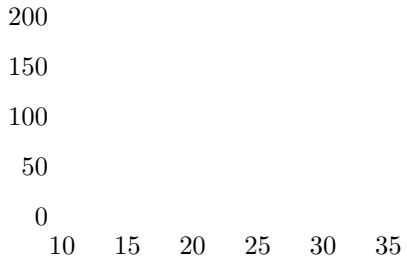


Фиг. 1.4: Хистограма

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 всеки интервал се преброят попаданията на стойностите, то полученото число n_i 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 номера на интервала.

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.1.4).

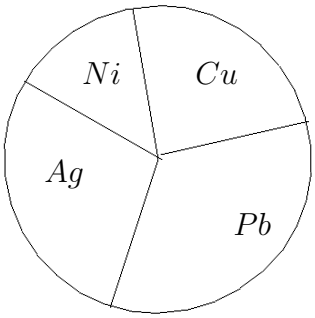


Фиг. 1.5: Кумулативна

Също така много удобна е така наречената кумулативна хистограма (фиг. 1.5). 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 хистограма, 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*, баници или торти (piechart).

Фиг. 1.6: Баница

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

От тези базисни картинки се строят множество по-сложни агрегирани картинки, които могат да се видят в учебника по приложна статистика (**?**) на фирмата, създала пакета STATISTICA. Приета е следната (спорна) класификация:

- 2D (обектите може да се маркират)
  - Бар, Баница, Бокс (за честоти или стойности)
  - Хистограми, Линии (с или без модел)
  - Вероятностни - $P-P$, $Q-Q$,
  - Диаграми на разсейване, Вороной
- 3D
  - XYZ - разсейване на три променливи
  - Последователно натрупани или наложени
  - Тернарни за смеси с отклик (числов или категорен)
- Многомерни и комбинирани
  - nD (Икони за всяко наблюдение)
  - Категорни (2D, 3D за различни категории)
  - Матрични (за много променливи)

1.5 Математически модел

Нека сега предположим, че нашите данни са получени от случайна извадка с краен обем от огромна (безкрайна) генерална съвкупност. Тогава на отделното наблюдение с номер i ние можем да гледаме като на стойност на случайната величина ξ_i, а за случайните величини да предполагаме, че са независими и еднакво разпределени.

Ако те имат разпределение изразявано с някаква хипотетична функция на разпределение $F(x)$, то можем да запишем

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(\xi_i < x), \quad I(\xi_i < x) = \begin{cases} 1 & \xi_i < x \\ 0 & \xi_i \geq x \end{cases}$$

Така лесно получаваме че

$$\mathbf{E}\, F_n(x) = F(x), \qquad \mathbf{D}F_n(x) = F(x)(1-F(x))/n \underset{n\to\infty}{\longrightarrow} 0.$$

Следователно, $F_n(x) \xrightarrow{p} F(x)$ за всяко фиксирано x (вж.фиг.1.3). От закона на големите числа следва повече, а именно че $F_n(x) \xrightarrow{\text{п.с.}} F(x)$.

Предположението и твърдението остават в сила и за крайна генерална съвкупност, стига извадката да е с връщане на премерения обект в генералната съвкупност.

1.5.1 Гливенко – Кантели

Верно е обаче още по - силната

Теорема 1.1 *(Гливенко – Кантели)*

$$\mathsf{P}(\lim_{n\to\infty} \sup_x |F_n(x) - F(x)| = 0) = 1 \tag{1.1}$$

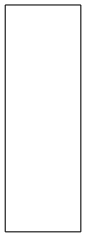
Доказателство: Първо да отбележим, че без намаление на общността можем да считаме $F(x)$ строго растяща функция.

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:

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

където $G_n(u) = F_n(F^{-1}(u))$.

Да фиксираме $\epsilon = 1/r$ и разделим интервала (0,1) на r равни части (подинтервали). Границите на интервалите да означим с $u_i = i/r$.



Фиг. 1.7: Подинтервал

На фигура 1.7 е показан един подинтервал. С удебелена линия е отбелязана "целевата" функция u.
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 $D_n \xrightarrow{\text{п.с.}} 0$. $\square$

Тема 2

Прости статистически 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 използуването на чисто вероятностни методи на разсъждение.

Статистическите изводи са строят обикновено по един от следните 3 различни начина:

- Проверка на статистически хипотези;
- Статистически оценки на параметри;
- Доверителни интервали за неизвестните параметри на генералната съвкупност.

В тази тема ще разгледаме няколко примера на възможно най- прости вероятностни разсъждения в статистиката. Тези примери не се нуждаят от особено силни предположения и, съответно, не

2.1 Данни

Данните в тази тема бяха генерирани в МАТЛАБ с следните команди:

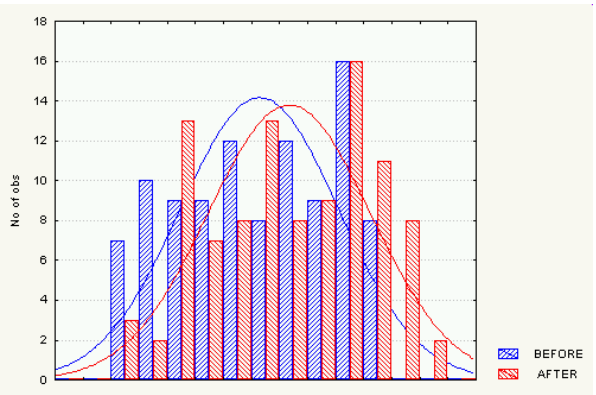
```
x=rand(100,1); y=0.1+0.1*randn(100,1);
x=sort(x); XX=[x,x+y,y]; XX
```

Получените данни от първата колона на матрицата XX са дадени в следната матрица:

Променлива Before

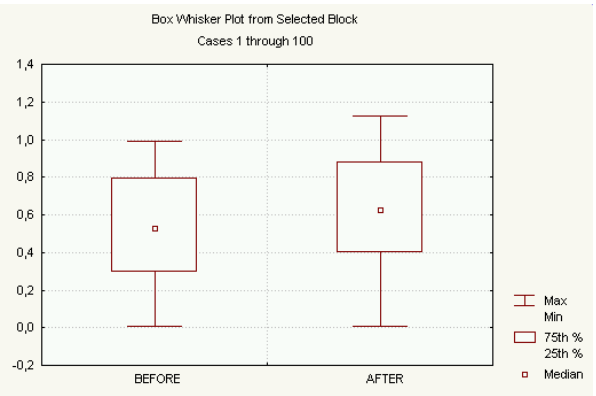
	1	2	3	4	5	6	7	8	9	10
0	,00990	,01180	,01530	,01850	,01960	,05790	,06480	,13650	,13890	,15090
1	,17300	,17630	,18970	,19340	,19870	,19880	,19910	,20260	,20280	,23110
2	,25230	,27140	,27220	,28440	,28970	,29870	,30280	,30460	,30930	,34120
3	,34200	,35290	,37040	,37840	,37950	,40570	,41030	,41860	,42890	,44470
4	,44490	,44510	,45650	,46600	,46920	,48600	,49660	,50280	,52260	,52520
5	,53410	,54170	,54660	,56810	,59360	,60380	,60680	,61540	,62130	,64490
6	,66020	,66140	,67210	,68130	,68220	,69460	,69790	,70270	,70950	,72710
7	,73730	,73820	,74680	,76210	,79190	,79480	,81320	,81800	,82140	,82160
8	,83180	,83810	,83850	,84620	,85370	,86000	,87570	,88010	,89130	,89360
9	,89390	,89980	,91690	,92180	,93180	,93550	,95010	,95680	,97970	,98830

2.2 Тест на знаците



Фиг. 2.1: x и y

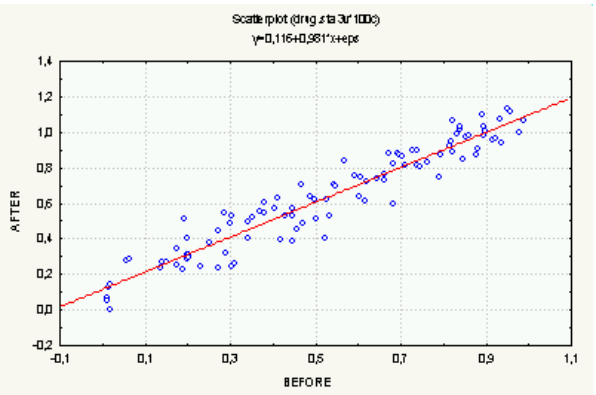
Нека е дадена една извадка от генерална съвкупност и са премерени по два еднотипни параметъра за всяко наблюдение: $x_1, x_2, \ldots, x_n$ и $y_1, y_2, \ldots, y_n$. Такива наблюдения се наричат *сдвоени или повторни*, т.е. на всяко x_i съответствува y_i.



Фиг. 2.2: Преди и след

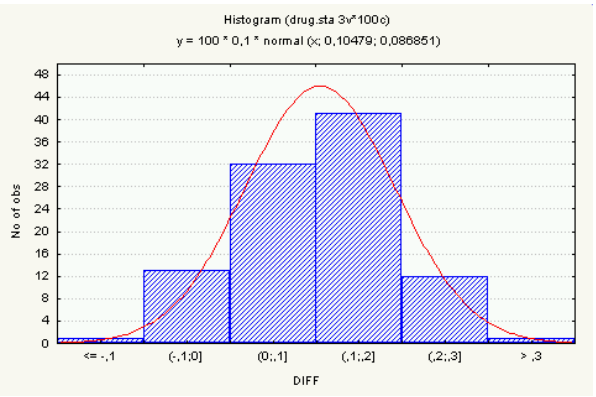
Такава ситуация възниква често в практиката. Например, когато мерим някаква характеристика върху едни и същи обекти преди и след въздействието с някакъв химикал или състоянието на болни преди и след лечението с определено лекарство.

На фигурите 2.1 и 2.2 не се вижда отчетлива разлика на измерванията преди и след.



Фиг. 2.3: $y = f(x)$

Фигура 2.3 показва изразена зависимост между влиянията преди и после. От тези фигури обаче не можем да направим заключение за наличието на влияние на лекарството.



Фиг. 2.4: $y - x$

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). За целта да разгледаме разликата $y - x$.

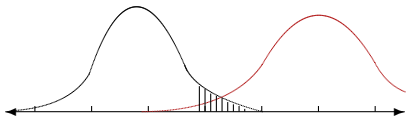
На фигура 2.4 се забелязва изместване на разпределението надясно от 0. Изглежда, че лекарството влияе положително - увеличава стойността на измерването.

За да проверим тази наша хипотеза да построим математически модел. Да разгледаме статистиката (функция от наблюденията) $Z = \#\{i : y_i > x_i\}$ - броят на положителните разлики между наблюденията "след" и "преди".

Нека 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.

Задаваме си малка вероятност α равна, например, на 0.05. Ще определим *критична* за нашата хипотеза област W, така че $\mathsf{P}(W) = \alpha$. Критична - това означава, че ако статистиката попадне в нея, ще отхвърлим хипотезата.

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.



Фиг. 2.5: Определяне на W

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

$$W = \{Z_n : Z_n \geq i\},$$

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

Остава да намерим i. При големи стойности на n може се използува интегралната теорема на Муавър - Лаплас. Това ни дава лесна възможност да намерим необходимото i. Имаме:

$$\Phi(\frac{i-.5n}{.5\sqrt{n}}) = 1-\alpha$$

$$\frac{i-.5n}{.5\sqrt{n}} = \Phi^{-1}(1-\alpha)$$

Така, ако броят на наблюденията с положителен знак $Z > 0.5(n+1.68)]$, би трябвало да отхвърлим хипотезата, че в двете измервания няма разлика. Вероятността да сбъркаме при такова твърдение е малка - $\alpha = 0.05$.

За нашия случай имаме $Z = 86$. Критичната стойност за ниво на грешката 0.05 е $0.5(100 + 1.68\sqrt{100} = 58.4$. Спокойно отхвърляме хипотезата.

Ако извикаме МАТЛАБ да сметнем вероятностната стойност (P-value) на статистиката даже ще получим единица.

```
m=50;s=5; p=normcdf(86,m,s)
   p = 1.0000
```

2.3 Статистически оценки

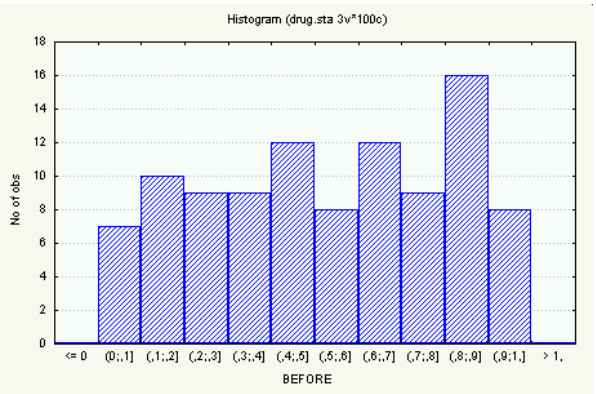
Както видяхме в предната тема най-естествено е да подменим във формулата изразяваща стойността на някакъв параметър чрез функцията на разпределение тази функция с извадъчната.

$$\theta = G(F) \Longrightarrow \hat{\theta} = G(F_n)$$

Този метод се нарича метод на моментите. За данните от таблица 2.1 получаваме следните оценки:

Descriptive Statistics (drug.sta)			Variable: BEFORE	
	Confid.	Confid.		
Valid N	Mean	-95,000%	+95,000%	Median
100	,528567	,472660	,584474	,529650
		Lower	Upper	Quartile
Minimum	Maximum	Quartile	Quartile	Range
,009900	,988300	,294200	,793350	,499150
			Standard	
		Std.Dev.	Error	
		,281756	,028176	
	Std.Err.		Std.Err.	
Skewness	Skewness	Kurtosis	Kurtosis	
-,137807	,241380	-1,15289	,478331	

2.4 Доверителен интервал за медиана



Фиг. 2.6:

Нека си поставим за цел по n наблюдавани стойности да кажем нещо за неизвестната медиана μ на това разпределение. 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.

$$x_{(i)} \leq \mu \leq x_{(n-i+1)}$$

Вероятността в дясно на формула (2.1) 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.

Ако приложим този тест към числата от таблица 2.1 ще получим интервала: $0.44470 \leq \mu \leq 0.66020$

2.5 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:

$$W = \{U_{1-\alpha} \leq U_x\}$$

Стойностите на $U_{1-\alpha}$ се подбират така, че вероятността за грешка да бъде малка:

$$P(W) = \alpha.$$

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

Теорема 2.2 *Ако е изпълнена хипотезата за еднаквост на разпределенията и те са непрекъснати, то са верни следните формули*

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

Доказателство: Да разгледаме сл.в. δ_{ij}. Те са еднакво разпределени, но не са независими. Имаме

$$\mathbf{E}\, \delta_{ij} = 1/2.$$

Значи $\mathbf{E}\, U_x = \frac{n_x n_y}{2}$. За да пресметнем дисперсията да означим с $\gamma_{ij} = \delta_{ij} - \mathbf{E}\, \delta_{ij}$ центрираните величини. Забелязваме, че

$$\mathbf{D}\,(U_x) = \mathbf{E}\,(U_x - \mathbf{E}\, U_x)^2 = \mathbf{E}\,(\sum_{ij} \mathbf{E}\, \gamma_{ij})^2 =$$

$$= \sum_{ij} \mathbf{E}\, \gamma_{ij}^2 + 2 \sum_{ij} \sum_{kl} \mathbf{E}\, \gamma_{ij} \gamma_{kl}.$$

Втората сума ще разделим на 4 части. Като отчетем еднаквата разпределеност на сл.в. γ_{ij}, получаваме:

- (съвпадат и двата индекса) $\mathbf{E}\, \gamma_{ij} \gamma_{ij} = (1/2 + 1/2)1/4 = 1/4$ защото $\mathsf{P}(\gamma_{ij} = 0) = 0$ ($1/4$ след скобите е стойността на сл.в.);
- (не съвпадат и двата индекса) $\mathbf{E}\, \gamma_{ij} \gamma_{kl} = 0$ защото са независими;

- (съвпадат индексите от първата извадка) $\mathbf{E}\,\gamma_{ij}\gamma_{il} = 1/3(1/4 - 1/4 + 1/4) = 1/12$. защото в обединената извадка елемент от X е или по-малък от или между или по-голям от
  два елемента на Y с вероятност $1/3$ ($1/4$ в скобите е стойността на сл.в.);;
- (съвпадат индексите от втората извадка) $\mathbf{E}\,\gamma_{ij}\gamma_{kj} = 1/3(1/4 - 1/4 + 1/4) = 1/12$. по същата причина за елемент от Y,

Така като преброим елементите на всяка от сумите получаваме:

$$\mathbf{D}\,(U_x) = \frac{n_x n_y}{4} + \frac{n_x n_y(n_y - 1) + n_y n_x(n_x - 1)}{12} = \frac{n_x n_y(n_x + n_y + 1)}{12}.\square$$

2.6 Изводи

Научихме, че

- понятието статистическа хипотеза всъщност е предположение за някакъв конкретен вероятностен модел;
- трябва да можем да определим разпределението на статистиката, която ни интересува;
- построяване на доверителен интервал и критична за хипотезата област си приличат;
- вероятностната стойност (P-value) на статистиката замества критичната стойност;
- симулацията може да бъде полезна.

Тема 3

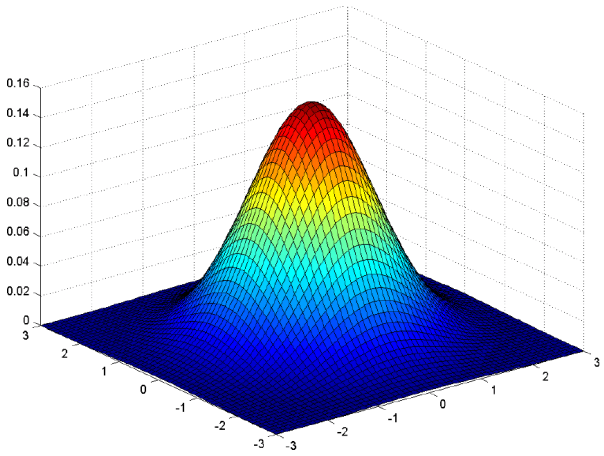
Нормално разпределение в R^n

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.

3.1 Нормално Разпределение

Определение 3.1 *Плътността на стандартното нормално разпределение в R^n има вида:*



Фиг. 3.1: N(0,I) в R^2

$$\varphi(x) = \frac{1}{(2\pi)^{n/2}} e^{-x'x/2}, \qquad (3.1)$$

От определението се вижда, че тази плътност зависи само от $\|x\|^2 = x'x = \sum_{i=1}^{n} x_i^2$. и, следователно,

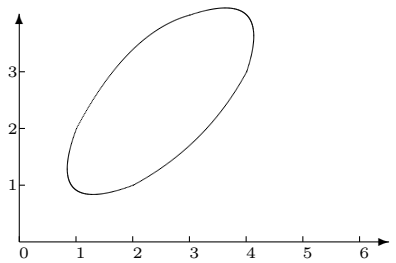
- е инвариантна относно всякакви ортогонални трансформации — те запазват нормата и имат якобиян равен на 1;
- тя може да се представи като произведение на n едномерни стандартни нормални плътности:

$$\varphi(x) = \prod_{i=1}^{n} \frac{1}{2\pi} e^{-x_i^2/2}.$$

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:

$$\phi(x, m, C) = \frac{1}{(2\pi)^{\frac{n}{2}}(det(C))^{\frac{1}{2}}} e^{-(x-m)'C^{-1}(x-m)/2}, \tag{3.2}$$



Фиг. 3.2: Линия на ниво

където $x \in R^n$, $m \in R^n$ е средната стойност, а C – ковариационната матрица.

На фиг.3.2 е показана линия на постоянно ниво на двумерна гаусова плътност $(x-m)'C^{-1}(x-m) = const$ – тя е елипса.

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

Теорема 3.1 *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 3.2 *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.
- произволна линейна функция от (зависими или независисми) нормални сл.в. е нормална сл.в.;
- 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 $m = n - k$ ортогонални помежду

си и на S единични вектори и означим така получената матрица с $\tilde{T}$.

$$\tilde{T} = \left\| \begin{array}{c} T \\ E \end{array} \right\| = \left\| \begin{array}{cccc} t_{11} & t_{12} & \dots & t_{1n} \\ t_{21} & t_{22} & \dots & t_{2n} \\ \dots & \dots & \dots & \dots \\ t_{k1} & t_{k2} & \dots & t_{kn} \\ e_{11} & e_{12} & \dots & e_{1n} \\ \dots & \dots & \dots & \dots \\ e_{m1} & e_{m2} & \dots & e_{mn} \end{array} \right\|.$$

Имаме $TE' = ET' = 0$ и $EE' = I$ в R^k.

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:

$$\tilde{T}\tilde{T}' = \left\| \begin{array}{cc} TT' & TE' \\ ET' & EE' \end{array} \right\| = \left\| \begin{array}{cc} TT' & 0 \\ 0 & I \end{array} \right\|.$$

Такава е и нейната обратна.

Следователно, разпределението се разпада в произведение на две плътности:

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

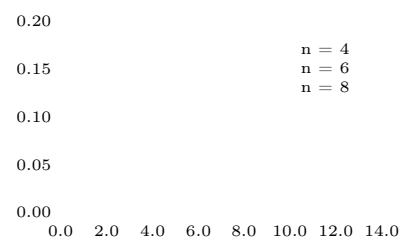
Тук разлагането $x = \{x_1, x_2\} (= x_1 + x_2)$ представя вектора в неговите проекции (координати) в подпространството S и неговото ортогонално допълнение. От тук лесно следва твърдението на теоремата. □

3.2 χ^2 разпределение

Определение 3.2 *Случайната величина $\chi_n^2 = \xi'\xi$ има разпределение $\chi^2(n)$ с n степени на свобода и плътност:*

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

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



Фиг. 3.3: χ^2

Веднага се вижда, че това е Гама-разпределение $\Gamma(\frac{n}{2}, \frac{1}{2})$ и изводът на тази формула може да се направи по индукция от $n = 1$ и възпроизводящите свойства на Гама-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 3.1 *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 3.2 *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 3.1. 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 $d_i = 1, i = 1, 2, ..., n$. □

3.3 Теорема на Кокрън

Теорема 3.3 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 3.2 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 3.2 за операторите R и S и определението на Хи-квадрат разпределение.

Необходимост. Равенството $q = r + s$ следва директно от лема 3.2. □

3.4 Примери

Пример 3.1 *Независимост на* $\bar{x} = 1/n \sum x_i$ *и* $S^2 = \sum (x_i - \bar{x})^2$

Доказателство: Наистина, това е частен случай от теоремата на Кокрън

$$\xi'\xi = \xi' B\xi + \xi'(I - B)\xi = n\bar{\xi}^2 + S^2, \quad B = \left\| \begin{array}{cccc} 1/n & 1/n & \dots & 1/n \\ 1/n & 1/n & \dots & 1/n \\ \dots & \dots & \dots & \dots \\ 1/n & 1/n & \dots & 1/n \end{array} \right\|.$$

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

Пример 3.2 *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:

$$\mathbf{E}\,(\xi_2 \| \xi_1) = a\xi_1 + b, \quad r(\xi_1, \xi_2) = a/S_{22},$$

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 4

Математическа 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 дадем идея 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.

4.1 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.

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 $f(x)$.

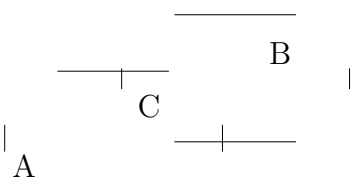
4.1.1 Лема на Нейман-Пирсън

За основен инструмент ни служи следната знаменита лема.

Лема 4.1 *Нека са дадени две плътности $f_0(x), f_1(x)$. Тогава решението на разпределителната задача:*

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

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



Фиг. 4.1: Нейман-Пирсън

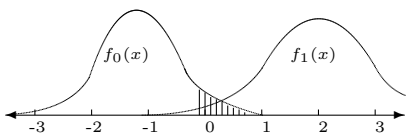
Доказателство: Нека $W = \{x : f_1(x) \geq cf_0(x)\}$ и $\alpha = \int_W f_0(x)dx$. Нека W' е такова, че $\alpha = \int_{W'} f_0(x)dx$.
$A = W \setminus W'$, $B = W' \setminus W$, $C = WW'$
Да разгледаме разликата:

$$\begin{aligned} \int_W f_1(x)dx - \int_{W'} f_1(x)dx &= \textstyle\int_A f_1(x)dx - \int_B f_1(x)dx \quad \geq \\ \int_A cf_0(x)dx - \int_B cf_0(x)dx &= c(\textstyle\int_W f_0(x)dx - \int_{W'} f_0(x)dx) \quad = 0.\square \end{aligned}$$

4.1.2 Критерий за проверка на хипотеза

Лемата на Нейман - Пирсън се използува по следния начин. Искаме да проверим хипотезата H_0, че наблюдението има плътност $f_0(x)$ срещу контра хипотезата 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. Тъй като взимаме решението въз основа на наблюдение, то областтта, в която попада при отхвърляне на хипотезата наричаме критична и означаваме с W.

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.

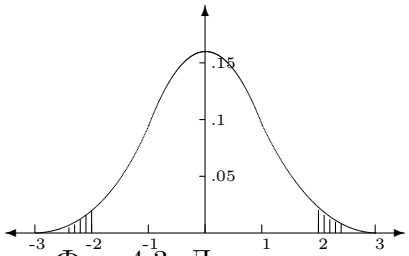


Фиг. 4.2: Едностранен критерий

Числото α наричаме грешка от първи род, а числото $1 - \alpha$ - *ниво на доверие*. На фиг. 4.2 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.

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 д а 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.



Фиг. 4.3: Двустранен критерий

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".

Пример 4.1 *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 4.1 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 c 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 4.2 *Нека $H_0 : \xi \in N(0,1)$, а $H_1 : \xi \in N(0,\sigma^2)$. Нека сме направили n наблюдения.*

Намерете оптималния критерий за всяка от алтернативите: $\sigma > 1, \sigma < 1$.

4.1.3 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 4.1 се вижда, че за всички алтернативи (със средна стойност по - висока от 0) решението ще бъде същото.

Определение 4.1 *Казваме, че критерият е равномерно най-мощен за дадено множество алтернативи, ако той е оптимален за всяка алтернатива поотделно.*

Така на фигура 4.2 е показан критерий, който е равномерно най - мощен за всички "десни" алтернативи.

Пример 4.3 *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.

4.2 Тест и критерий

Тук ще направим някои езикови бележки за да изясним и нашата терминология.

4.2.1 Предпоставки

В англоезичната статистическа литература изобщо не се среща понятието **критерий (criterion)**. Навсякъде се използува тест (test - проверка).

Например:

- (вж. Lehman *Testing statistical hypothesis*)
- Student and Fisher's tests

В руската литература по (математическа) статистика (оригинална и преводна) не се среща понятието **тест** - всичко е или критерий или проверка.

Например:

- (вж. Lehman *Проверка статистических гипотез*)
- Критерий Стьюдента, Фишера.

В българската литература по статистика и двете понятия се използуват равноправно без ясно разграничаване.

4.2.2 Семантика

Думата тест в Българския език е придобила своя отделен смисъл като **процедура за проверка на нещо.**

Думата критерий в Българския език означава определено **условие, изпълнението на което е необходимо или достатъчно за нещо:**

- критерий за влизане в Европейската общност;
- критерий за добро познаване на статистическата литература.

4.2.3 Предложение

Предлагаме да се запази думата **тест** за всички статистически процедури и проверки:

- Тест на Стюдент, тест на Фишер, хи-квадрат тест, тест за нормалност на разпределението.
- Да се използува **критерий само за понятието критична област**, когато можем да определим статистиката, разпределението, мощността на критерия, областта. Така се запазва и семантиката на тази дума.

4.3 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 или интервал. Наричаме такава област *доверителна*, а вероятността на твърдението – ниво на доверие.

Тъй като областта е функция на наблюденията, тя става случайна. Ясно е, че колкото по - широка е тя, толкова по - вероятно е да покрие неизвестния параметър, т.е. нашето твърдение за него да се окаже вярно.

Естествено би било да поискаме и тук някаква оптималност — например, областта да има минимален обем при фиксирано ниво на доверие. Когато говорим за едномерен параметър, се интересуваме от доверителни интервали с минимална дължина.

4.3.1 Толерантен интервал

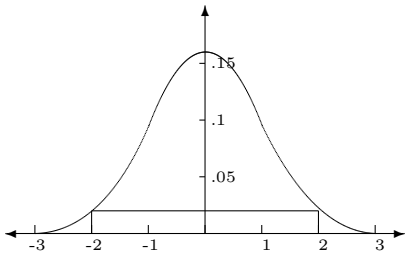
В такава постановка задачата много прилича на лемата на Нейман - Пирсън. Първоначално ще построим доверителна област за наблюдението, така че тя да има минимален обем. Прието е такава област да се нарича толерантен интервал. В последствие (при подходящи условия) тя ще се превърне в доверителна област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

Лема 4.2 *Нека е дадена плътността $f(x)$. Тогава решението на разпределителната задача:*

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

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

Доказателство: Абсолютно същото като на оригиналната лема.□



Фиг. 4.4: Толерантен интервал

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 интервала така, че плътността да бъде равна в двата края.

Обикновено това е достатъчно за проверка на оптималността (минималната дължина) на така построения толерантен интервал.

Определение 4.2 *Функция от данните и параметъра, чието разпределение не зависи от стойността на параметъра се нарича централна (pivotal).*

Плътността на фиг.4.4 е такава.

4.3.2 Доверителен интервал

Сега ще използуваме толерантни интервали за конструкция на доверителни. Естествено е, че за да направим това трябва да познаваме разпределението на съответните статистики.

Пример 4.4 *Нека $\xi \in N(\theta, 1)$. Нека сме направили n наблюдения. Намерете оптимална доверителна област за θ.*

Решение 1. Векторното наблюдение x ще има за плътности многомерната нормална плътност $c_n e^{-\frac{1}{2}\|x-\theta e\|^2}$. От лемата 4.2 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 и е централна. 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. Функцията $\sum_{i=1}^{n}(x_i - \bar{x})^2$ също е централна.

4.4 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 4.4.

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 4.2 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 е разгледана подробно в лекциите по статистика. Да разгледаме още един пример.

Пример 4.5 *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:

$$f(x) = \frac{\lambda^n x^n e^{-\lambda x}}{(n-1)!}, \tag{4.1}$$

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

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

□

4.5 Асимптотика

Много от статистиките представляват суми от прости функции на наблюденията - такива са, например, извадъчните моменти. Тогава по ЦГТ те имат асимптотично нормално разпределение. За да използуваме това разпределение е достатъчно да познаваме само два параметъра м.о. и стандартно отклонение.

В случая, когато статистиката се разглежда като оценка на неизвестен параметър на генералната съвкупност, обикновено тя се подбира така, че да е неизместена:

$$\mathbf{E}\,\hat{\theta}(\xi) = \theta$$

От друга страна за много от статистиките сме в състояние да оценим дисперсията $\sigma^2(\hat{\theta})$. За това могат да се използуват както теоретични, така и извадъчни методи.Да разгледаме функцията:

$$g(\hat{\theta}, \theta) = \frac{\hat{\theta} - \theta}{\sigma(\hat{\theta})}.$$

Тя е централна и нейното разпределение е стандартно нормално (асимптотично). Значи при достатъчен брой наблюдения ние лесно можем да строим доверителни интервали и проверяваме хипотези с използуване на тази асимптотика. Да разгледаме следната таблица произведена от STATISTICA.

Descriptive Statistics (drug.sta)			Variable: BEFORE
	Std.Err.		Std.Err.
Skewness	Skewness	Kurtosis	Kurtosis
-,137807	,241380	-1,15289	,478331

Таблица 4.1: Описателни статистики

Пример 4.6 *Проверете хипотезата за равенство на 0 на асиметрията и ексцеса с използуване на данните от таблицата.*

Решение:
За асиметрията:
$x = 0.137807/0.241380 = 0.5709, \Phi(x) - \Phi(-x) = 0.4319.$
Не можем да отхвърлим хипотезата.
За ексцеса:
$x = 1,15289/0.478331 = 2.4102, \Phi(x) - \Phi(-x) = 0.9841.$
Можем да отхвърлим хипотезата с вероятност за грешка 0.02.

Пример 4.7 *Постройте двустранни асимптотични 95 процентни доверителни интервали за асиметрията и ексцеса по същите данни.*

Решение:
Тъй като 95 процентния двустранен квантил на стандартното нормално разпределение е $z = 1.96$, получаваме:
За асиметрията:

$$As \in (-.137807 - 1.96 * 0.241380, -.137807 + 1.96 * 0.241380)$$

$$\boxed{\text{-0.6109 <As<0.3353}}.$$

Съдържа 0 и не можем да отхвърлим хипотезата с вероятност за грешка 0.05.
За ексцеса:

$$Ex \in (-1.15289 - 1.96 * 0.478331, -1.15289 + 1.96 * 0.478331)$$

$$\boxed{\text{-2.0904 <Ex<-0.2154}}$$

Не съдържа 0 и можем да отхвърлим хипотезата с вероятност за грешка 0.05.

Тема 5

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

5.1 Доверителен интервал за дисперсия

Ще започнем с дисперсията.

Пример 5.1 *Нека $\xi \in N(\mu, \sigma^2)$. Нека сме направили n наблюдения. Намерете оптимална доверителна област за σ^2.*

Решение:
Статистиката $S^2 = \sum_{i=1}^{n}(x_i - \mu)^2$ има разпределение $\sigma^2\chi^2_n$.От лемата 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.

$$\frac{S^2}{q_u} \leq \sigma^2 \leq \frac{S^2}{q_l}. \tag{5.1}$$

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

$$S^2(x) = \sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{5.2}$$

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 13.1.

Така, ако си зададем ниво на доверие .95 и броят на нашите наблюдения е 15, ще използуваме числата $q_l = 5.63$ и $q_u = 26.12$ от реда за 14 степени на свобода.

5.1.1 Най-къс доверителен интервал

Тук предлагаме малка програма на МАТЛАБ за пресмятане на най-къс доверителен интервал за дисперсията. Тя е основана на числената минимизация:

$$\min_{q_l,q_u}(\frac{1}{q_l}-\frac{1}{q_u}), \qquad F(q_u)-F(q_l)\geq 1-\alpha$$

```
function [sd,sdlo,sdup]=sd_bounds(x,alf)
%  [sd,sdlo,sdup]=sd_bounds(x[,alf])
%  calculates shortest (1-alf) confidence
%  interval for variance of vector x
if nargin<2, alf=0.05; end;
[v,dum]=size(x);v=v-1;
y=[chi2inv(alf/2,v);chi2inv(1-alf/2,v)];
y=constr('ch2dv',y,[],[],[],[],v,alf);
sd=std(x);y=sqrt(y/v);
sdlo=sd/y(2);sdup=sd/y(1);
```

Използува се помощната функция `ch2dv`, която пресмята дължината на интервала f и ограниченията $g\leq 0$:

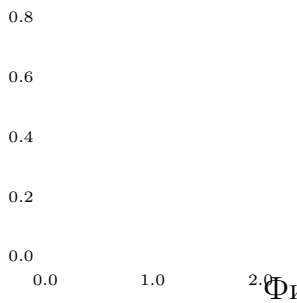
```
function [f,g]=ch2dv(y,v,alf)
f=(1/y(1)-1/y(2));
g=[1-alf-chi2cdf(y(2),v)+chi2cdf(y(1),v);y(1)-y(2)];
```

5.2 Разпределение на Фишер

Определение 5.1 *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. 5.1: Фишер

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

$$C(r,s)x^{r/2-1}(1+rx/s)^{-(r+s)/2}, \quad x>0,$$

където $C(r,s)$ е нормираща константа.

Докажете го като използувате връзката между Гама, Бета и F - разпределение: $b(2a,2b)=\frac{\chi^2_a}{\chi^2_a+\chi^2_b}$.

5.2.1 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:

$$S^2(x) = \sum_{i=1}^{n_1}(x_i - \bar{x})^2, \qquad S^2(y) = \sum_{i=1}^{n_2}(y_i - \bar{y})^2. \tag{5.3}$$

Съгласно теоремата 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 5.1 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

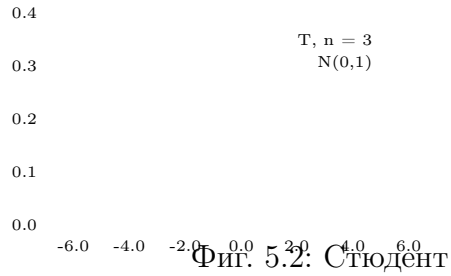
Теорема 5.1 *Нека P, Q са два проектора в R^n такива, че $PQ = QP = 0$ и ξ е стандартна нормална в R^n. Тогава частното $f = (\xi' Q\xi)/(\xi' P\xi)$ има разпределение на Фишер с $dim(Q)$ и $dim(P)$ степени на свобода.*

Като следствие от тази теорема получаваме

Теорема 5.2 *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.*

5.3 Разпределение на Стюдент

Определение 5.2 *Разпределение на Стюдент $T(n)$ с n степени на свобода има частното:*



Фиг. 5.2: Стюдент

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

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

$$C_n(1 + x^2/n)^{-(n+1)/2}.$$

Докажете го чрез 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 $N(0,1)$.

5.3.1 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

$$t = n^{1/2}(\bar{x} - \mu)/s \tag{5.4}$$

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 лемата на Нейман-Пирсън 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 е с минимална дължина.

5.3.2 Критерий на Стюдент

Нека пак предположим, че наблюденията са от $N(\mu, \sigma^2)$. Статистиката t 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.

5.3.3 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 $\sigma_1 = \sigma_2$. Решението правим в две стъпки:

А. Равенство на дисперсиите

1. От начало се прилага теста на Фишер за проверка на равенството на двете дисперсии (виж.5.2.1). Разпределението на Фишер не зависи от стойностите на неизвестните м.о. на двете популации.

2. При отхвърляне на нулевата хипотеза $H_0 : \sigma_1 = \sigma_2$ задачата е решена - двете популации се приемат за различни.

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

Б. Равенство на математическите очаквания

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 оценка на дисперсията σ^2:

$$s^2 = \frac{1}{n+m-2}(S^2(x) + S^2(y)) \tag{5.5}$$

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:

$$t = (\frac{nm}{n+m})^{1/2}\frac{\bar{x} - \bar{y}}{s} \tag{5.6}$$

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 6

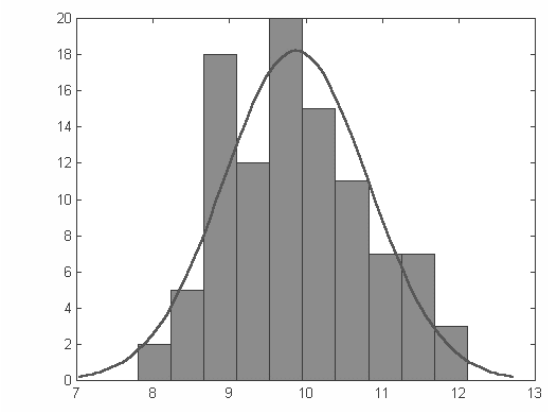
χ^2 и честотни таблици

В много случаи данните се сумират в така наречените честотни таблици. Нека за всеки обект (наблюдение) си отбелязваме проявата на някакви признаци. Ако признаците са два можем да сумираме наблюденията си в таблица. Когато човек отвори един статистически годишник, този начин за представяне на обекти е първото с което ще се запознае. Критерият χ^2, всъщност е съвкупност от статистически процедури основани на свойствата на χ^2 разпределението. В тази лекция ще се запознаем с двата най-популярни в практиката χ^2-критерии

- за съгласуваност на разпределения;
- за проверка на независимостта в честотни таблици.

6.1 Съгласуваност на разпределения

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. 6.1:

Нека ни е зададена някаква теоретична плътност $p(x)$ и имаме за задача да проверим съгласуваността ѝ 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, 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

$$h = \sum_{i=1}^{k} \frac{(np_i - n_i)^2}{np_i}. \tag{6.1}$$

Теорема 6.1 *(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 (6.1) представлява квадрата на центрирана асимптотично нормална сл.в. Действително, $np_i = \mathbf{E}\, n_i$. За съжаление, тези величини са зависими – $\sum_{i=1}^{k} n_i = n$. Оказва се, че условното разпределение на случаен гаусов вектор $\xi \in N(0, I)$ в R^k при условие $(\xi, 1) = 0$ е същото като асимптотичното съвместно разпределение на сл.в. n_i, съответно центрирани и нормирани. □

Същият критерий може с успех да се използува и при сравняването на две независими извадки.

6.2 Двумерни честотни таблици

Анализът на честотните таблици е една от най-популярните процедури в анализа данни. Най доброто в тази област е представено в книгата (Agresti 1990). В този случай има няколко статистики, за които могат да се пресмятат точни разпределения при изпълнена хипотеза за независимост на двата признака. Най-популярна, обаче е

6.2.1 χ^2-статистика

Нека ни е зададена двумерна честотна таблица $n_{i,j}$. За всяка клетка (i, j) предполагаме, че обектът попада в нея с вероятност $p_{i,j}$. Тъй като имаме всичко n обекта, получаваме че разпределението на всяко от сл.в. $n_{i,j}$ е биномно, но те са зависими.

Да си формулираме задачата в термини на проверка на статистическата хипотеза:

$$H_0 : \text{двата наблюдавани признака са независими}$$

срещу алтернативата:

$$H_1 : \text{признаците са зависими.}$$

Прието е проверката на такава хипотеза да се прави с χ^2-статистика.

Ако признаците бяха независими, би трябвало вероятност $p_{i,j}$ да е произведение от двете маргинални вероятности: $p'_i = (\sum_k p_{i,k})$ и $p''_j = (\sum_k p_{k,j})$.

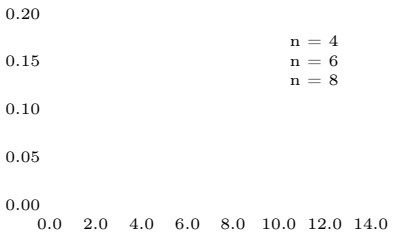
Следователно, χ^2-разстоянието между актуално наблюдаваните честоти $n_{i,j}/n$ и "теоретичните" вероятности $p'_i.p''_j$ би трябвало да отразява отклонението от независимост.

Така получаваме две извадъчни разпределения. За проверка на съгласуваността между тях можем да използуваме това разстояние.

Определение 6.1 *Нека първият признак има k, а вторият – m категории. χ^2-статистиката се определя така:*

$$h = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(np'_i p''_j - n_{i,j})^2}{np'_i p''_j} = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(n'_i n''_j/n - n_{i,j})^2}{n'_i n''_j/n}. \tag{6.2}$$

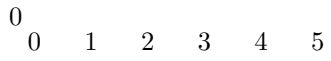
Тук сме означили: $n'_i = (\sum_j n_{i,j})$ и $n''_j = (\sum_i n_{i,j})$.



Фиг. 6.2: χ^2

При изпълнена хипотеза H_0, тя има асимптотично (при голямо n) разпределение χ^2 с $(k-1)(m-1)$ степени на свобода. Когато хипотезата е нарушена би трябвало нейните стойности да нарастнат. Така критерият е с дясна критична област.

В частния случай с таблица 2x2, получаваме χ^2 с 1 степен на свобода.



Фиг. 6.3: χ^2_1

6.2.2 The Case of Luddersby Hall

Ще илюстрираме концепцията със следния пример.

Пример 6.1 *Един ден студентите от едно общежитие в Англия масово започнали да повръщат. Медицинските анализи на фекалиите им показали у много от тях наличието на бактерии салмонела. Възникнало подозрение за яденето от предната вечеря. Сумарните данни на всичките 104 студента били записани в следната таблица.*

	Свинско	Зелен фасул	Торта с лимон
Повръща	39	33	63
Носител	14	4	17
Здрави	9	5	7

Смисълът й е следният. Всеки студент може да попадне в само един ред: повръща, носител – не повръща, но има бактерии, и здрав – нито е повръщал, нито има бактерии. По стълбове ситуацията е по-сложна. Всеки студент е ял или Свинско или Зелен фасул, но повечето са хапнали и десерт. Пита се, можем ли по тези данни да открием причината за инфекцията.

Ще разделим задачата на следните подзадачи, като разгледаме поотделно трите възможни блюда:

1. свинско, 2. зелен фасул, 3. десерт.

За всяко от тях ние можем да извлечем от горната таблица под-таблица, в която да запишем информацията за 4-те различни категории студенти $\{\text{ял}, \text{не ял}\} \times \{\text{болен}, \text{здрав}\}$. За болни ще смятаме студентите, които повръщат и тези които имат бактерии.

Така получаваме следните три под-таблички:

Свинско

	Ял	Не ял	Всичко
болни	53	37	90
здрави	9	5	14
Всичко	62	42	104

Фасул

	Ял	Не ял	Всичко
болни	37	53	90
здрави	5	9	14
Всичко	42	62	104

Торта

	Ял	Не ял	Всичко
болни	80	10	90
здрави	7	7	14
Всичко	87	17	104

Първо ще забележим, че първата и втората таблици ще произведат една и съща статистика и, следователно, не е необходимо да пресмятаме и двете поотделно.

Ще трябва да работим с χ^2 разпределение с 1 степен на свобода. Неговата плътност е дадена на фигура 6.3. А 0.95 квантила му е равен 3.84 (Виж. таблица 13.1). Така, когато статистиката h, пресметната по формула (6.2), надхвърли критичната стойност 3.84 би трябвало да отхвърлим нулевата хипотеза.

За първата (и втората) таблички стойността на статистиката h е равна на 0.1466. Следователно не можем да отхвърлим нулевата хипотеза. Това значи, че вида на консумираното първо блюдо не влияе на заразеността.

За третата табличка обаче, стойността на статистиката h е равна на 13.3994. Следователно трябва да отхвърлим хипотезата за независимост. Тъй като пропорцията на заболелите е по-висока при ялите торта, следва да заключим, че тортата е била източник на инфекцията.

Тема 7

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". Едва ли има по неподходящо название за този метод.

7.1 Основни задачи

Регресионният анализ намира най - често приложение за изследване на причинно - следствени връзки между количествени променливи. 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)?

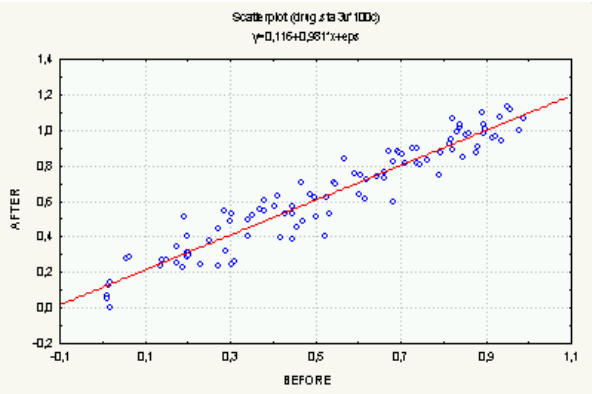
4. Какво можем да очакваме от отклика при зададени нови стойности на предикторите (задача за прогноз)?

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.

7.2 Проста линейна регресия

Нека отначало за илюстрация да въведем само един предиктор и един отклик. С други думи наблюдавали сме само две променливи и предполагаме, че стойностите на едната (отклика) в голяма степен се определят от другата (предиктора):



Фиг. 7.1: Проста регресия

Първо да построим оценка по метода на максималното правдоподобие. За модел на наблюденията имаме:

$$y_i = ax_i + b + \varepsilon_i, \quad \varepsilon_i \in N(0, \sigma^2) \quad (7.1)$$

$i = 1, ..., n$. Нека погледнем как се решават отделните задачи.

Трябва да максимизираме логаритъма на правдоподобието:

$$LL(y, a, b, \sigma) = -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - ax_i - b)^2$$

Забелязваме, че максимума по σ се получава при

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - ax_i - b)^2, \text{ и } \max_{\sigma} LL(y, a, b, \hat{\sigma}) = -(\hat{\sigma})^{\frac{2}{n}} \quad (7.2)$$

Значи максимизирането на правдоподобието е еквивалентно на минимизиране на сумата от квадратите на остатъците:

$$\min_{a,b} SSr = \sum_{i=1}^{n}(y_i - ax_i - b)^2,$$

което е довело до знаменитото название метод на най-малките квадрати, дадено от К.Ф.Гаус.

Така получаваме оценките:

$$\hat{a} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \text{ и } \hat{b} = \bar{y} - \hat{a}\bar{x}. \quad (7.3)$$

Да изразим оценките като функции от сл.в. на модела (да заместим в тях $y_i = ax_i + b + \varepsilon_i$):

$$\hat{a} = a + \frac{\sum_{i=1}^n \varepsilon_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \text{ и } \hat{b} = b + \bar{\varepsilon} + (a - \hat{a})\bar{x}. \tag{7.4}$$

Така получихме, че оценките са

- неизместени - $\mathbf{E}\,\hat{a} = a$, $\mathbf{E}\,\hat{b} = b$;
- линейни функции от гаусовия вектор ε и, следователно, са също гаусово разпределени.
- при определени условия върху предикторите ($\bar{x} = 0$) те стават независими.

Това ни дава възможност да строим лесно доверителни интервали и проверяваме хипотези за параметрите на модела:

$$\hat{a} - z_\alpha \sigma(\hat{a}) < a < \hat{a} + z_\alpha \sigma(\hat{a}), \text{ и } \hat{b} - z_\alpha \sigma(\hat{b}) < b < \hat{a} + z_\alpha \sigma(\hat{b}). \tag{7.5}$$

Достатъчно е да познаваме дисперсиите на оценките $\sigma(\hat{a}), \sigma(\hat{b})$. От израза (7.4), като предположим за простота, че $\bar{x} = 0$, получаваме

$$\sigma^2(\hat{a}) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \text{ и } \sigma^2(\hat{b}) = \frac{\sigma^2}{n}. \tag{7.6}$$

7.3 Линейни модели с гаусова грешка

В цялата лекция нататък ще предполагаме, че $\varepsilon \in N(0, \sigma^2 I)$, 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:

$$y \;\; = \;\; z + \varepsilon. \tag{7.7}$$

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 7.1 *За модела (7.7) 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, $\varepsilon \in N(0, \sigma^2 W)$, 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 7.2 *(За хипотезите в регресията) 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:

$$f_{m-k,n-m} = \frac{\|y_1 - y_0\|^2/(m-k)}{\|y - y_1\|^2/(n-m)} > F_{1-\alpha}, \tag{7.8}$$

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:

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

Проверката на неравенството $\lambda(y) > c$ е еквивалентна на критичната област определена от неравенството (7.8). Твърдението за разпределението е пряко следствие от теоремата на Кокрън. □

Когато към модела (7.7) 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.

7.4 Многомерна линейна регресия

Нека изследваният модел е от вида

$$y = Xa + e, \tag{7.9}$$

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

$$\hat{a} = (X'X)^{-1}X'y \tag{7.10}$$

$$cov(\hat{a}) = \sigma^2(X'X)^{-1} \tag{7.11}$$

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 7.1 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 (7.10). Това решение съществува и е единствено поради пълния ранг на X.

Като заместим y в (7.10) получаваме

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

което влече неизместеността на $\hat{a}$. От същото представяне следва и представянето на $cov(\hat{a})$ в (7.11). □

От теорема 7.1 веднага получаваме, че неизместена оценка на σ^2 ще получим по формулата:

$$\hat{\sigma^2} = \frac{1}{n-k}\|y - X\hat{a}\|^2. \tag{7.12}$$

Тази оценка, обаче, не е максимално правдоподобна.

Тема 8

Хипотези в регресията

В тази лекция ще експлоатираме безпощадно теоремата за хипотезите в регресията (теорема 7.2) за сравняване на различни модели 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.

8.1 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:

$$y = Xa + b\vec{1} + e, \tag{8.1}$$

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 теоремата 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:

- Sum of Squares of Residuals:

$$SSR = \|y - y_1\|^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Sum of Squares due to the Model:

$$SSM = \|y_1 - y_0\|^2 = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$$

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.

8.2 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 теоремата за хипотезите в регресията (теорема 7.2) 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

$$t_i = \frac{\sqrt{(n-m)}\hat{a}_i}{\hat{\sigma}((X'X)^{-1}_{ii})^{1/2}} \tag{8.2}$$

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 (8.2).

8.3 Доверителен интервал за прогноза

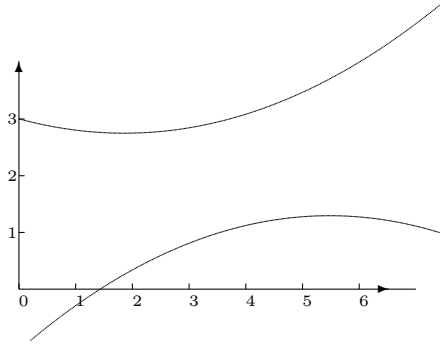
За произволни стойности x на предикторите от областта, за която е верен модела (8.1), случайната величина $\hat{y} = x'\hat{a} + \hat{b}$ е неизместена оценка за $E_x(y)$ и

$$D_x(\hat{y}) = \sigma^2(\frac{1}{n} + (x - \bar{X})'(\tilde{X}'\tilde{X})^{-1}(x - \bar{X})). \tag{8.3}$$

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

$$\sigma_y^2(x) = \sigma^2(1 + \frac{1}{n} + (x - \bar{X})'(\tilde{X}'\tilde{X})^{-1}(x - \bar{X})). \tag{8.4}$$

Проверете уравнения (8.3) и (8.4).



На фигурата е нарисувана апроксимиращата права при простия линеен модел $y = ax + b + \varepsilon$. С двете параболи са отбелязани доверителните граници за наблюдаваната стойност съгласно формула (8.4). С аналогична форма, но значително по - тесен е коридорът за модела – формула (8.3).

Фиг. 8.1: Проста линейна регресия

Така се вижда колко опасни (и понякога безсмислени) могат да бъдат прогнози за далечното бъдеще, основани на тенденция, наблюдавана в краен интервал от време.

8.4 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:

$$SSR = \sum_x (y_i(x) - \overline{y}(x))^2 + \sum_x n_x(\overline{y}(x) - \hat{y}(x))^2 \tag{8.5}$$

Тоест получаваме:

$$SSR = SSI + SSM.$$

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 (8.5). Постройте критичната област.

Тема 9

Стъпкова регресия

В тази тема ще се спрем на един от най-популярните методи за избор на модел. Ще разгледаме

- интерактивни методи на примера на СТАТЛАБ и MCTAT16;
- автоматични методи на примера на Statistica;
- математически основи (sweep operator);

9.1 Интерактивни процедури

Стъпкова регресия е една от основните процедури на регресионния анализ предназначена за избор на модела. Ще я илюстрираме с старият български пакет СТАТЛАБ (Вындев и Матеев 1988), предназначен за 8-битови компютри. За съжаление не познавам друг пакет позволяващ подобна степен на интерактивност.

9.1.1 Стъпкова регресия в СТАТЛАБ

Регресионната процедура в СТАТЛАБ е проектирана така, че максимално да се облекчи преминаването от модел в модел и избора на най-добрия. В пресмятанията са реализирани известните алгоритми за стъпкова регресия (вж.(Jennrich 1977)). Използуването на вътрешно-груповата ковариационна матрица дава възможност за включване в регресионния модел на групиращи променливи по естествен начин (ковариационен анализ).

Ще илюстрираме проблемите по избор на регресионен модел със следния

Пример 9.1 *Данните за престъпността 13.1 с Правец.*

Изборът на подходящ модел става върху следният екран:

LINEAR REG.	MODEL	
R.S.E.	=206.77789	F(8,38) = 1
> ■	R =	
	AGE	***.32760618** T= .99175068
	S	R =.50391799
	ED	***.50391799** T= .99596637
	EX0	***.75184247** T= .99999534
	EX1	R =.99424225
	LF	R =.76231375
<	M	***.14129216** T= .78962600
>	N	R =.73688289
	NW	R =.81953325
<	U1	***-.2469626** T= .85055129
	U2	***.37251087** T= .96414917
<	W	***.31667651** T= .80849620
	X	***.75474519** T= .99957971

Фиг. 9.1: Екранът на Правец

В първата колонка се намира показалец. Той се придвижва нагоре-надолу, както обикновено с A и Z. С клавиша <CR> променяме състоянието в модела на променливата, срещу която е показалеца:

- ако променливата е включена в модела, то тя ще бъде извадена;
- ако променливата е извън модела (и не е отклик), то тя ще бъде включена като предиктор;
- ако това е отклик, то квадратчето се премества надолу на следващата променлива, не включена в модела и тя става зависима (отклик).

Отгоре на екрана е отделен един ред, в който е дадена оценка за остатъчната грешка на модела R.S.E и съответната вероятност на критерия F. Когато тази вероятност е по-ниска от, например, 0.95 е съмнително, че избраните променливи изобщо влияят на отклика, т.е. няма основание да се отхвърли хипотезата, че всичките коефициенти в модела са едновременно 0.

Първоначално, когато в модела няма включени променливи, R.S.E. е оценената стандартна грешка на зависимата променлива. Тази оценка се прави по данните само на тази променлива с отчитане на груповата им принадлежност. Изчислително тя е квадратен корен от съответния диагонален елемент на използваната ковариационна матрица.

По долу за всяка променлива е отделен един ред, като реда на отклика е маркиран с бяло квадратче и върху него няма друга информация освен името на променливата.

Променливите, влезли в модела, се виждат лесно, тъй като числата до имената им са дадени в бяло поле (форма INVERSE), а освен това тези редове съдържат по две числа. Първото (маркирано) число е стандартизираният регресионен коефициент (той съвпада с регресионния коефициент, ако отклика и предикторите са стандартизирани). Тези коефициенти имат предимството, че по тяхната величина можем да съдим за степента на влияние на съответния предиктор върху отклика и те са безразмерни.

Второто число е вероятността, една разпределена по Стюдент случайна величина да не надхвърля t-статистиката на съответния коефициент. Тази вероятност, доколкото тя е свързана с критерия на Стюдент и T - разпределението, ще наричаме T-вероятност на съответния коефициент. Ако тя е по-малка от .95, участието

на предиктора в регресионното уравнение е съмнително и програмата ще подскаже във втората колонка на реда със знака "<" да бъде изключен. Ако обратно, тя е по-голяма от .95, то с ниво на доверие .95 може да се отхвърли хипотезата, че коефициентът е равен на нула. Това означава, че участието на съответният предиктор в модела е съществено.

Макар че в СТАТЛАБ е фиксирано това най-популярно ниво на доверие 0.95 и то се използува за изработване на "подсказванията", нищо не пречи да се зафиксира отнапред някое друго ниво на доверие и работи постоянно с него. Съществено е, да не се променя в процеса на работа с данните.

Друга съществена забележка е, че далеч не винаги отхвърлянето на предиктор с незначима T-вероятност води до съществено подобряване на модела. По-обоснована би била проверката на частния коефициент на корелация на този предиктор с отклика, но СТАТЛАБ не притежава тази възможност.

След имената на променливите, които не са предиктори, е показан квадрата на множествените им коефициенти на корелация R с предикторите - променливите, влезли в модела. Естествено е, да се опитваме да "добавим" към модела онези променливи, които притежават минимален коефициент на множествена корелация R. Това осигурява ниска степен на зависимост между предикторите и, като следствие, малка грешка на оценените коефициенти на регресионния модел. Ако не се придържаме към това правило, може да се окаже, че предикторите са толкова силно корелирани, че тяхната извадъчна ковариационна матрица е почти изродена. Това явление се нарича мулти-колинеарност и на борбата с него е посветена голяма част от литературата по приложен регресионен анализ. В този случай грешките на оценките нарастват в такава степен, че ги обесмислят (става дума както за изчислителните, така и за статистическите - оценената ковариационна матрица на коефициентите).

По този начин можем да се включват и изключват променливи в модела и да се променя отклика. При избора на подходящ модел се ръководим преди всичко от R.S.E - по-добър модел съответствува на по-малък R.S.E.

Целесъобразно е в модела да се включват променливи с минимален R, т.е. най-слабо зависещи от вече включените. Предиктори с относително малък коефициент или с ниска T-вероятност могат да бъдат извадени от модела. Като правило тези две действия водят до намаляване на R.S.E. и се подсказват от СТАТЛАБ във втора колонка съответно с "<" за изваждане и ">" за включване в модела.

LINEAR REG.		MODEL
>R	=-7147.326	+10.08191 * AGE
		+17.42171 * ED
		+9.784477 * EX0
R.S.E.	=206.77789	+1.854476 * M
D.F.	=38	-5.296190 * U1
		+17.05928 * U2
		+1.269328 * W
		+7.316694 * X

Показаният по-горе екран отразява състояние на модела, което не се подобрява повече с включване и изключване на променливите. Ето защо завършваме - натискаме ОСВ и на екрана се показва уравнението на модела.

Фиг. 9.2: Окончателен модел

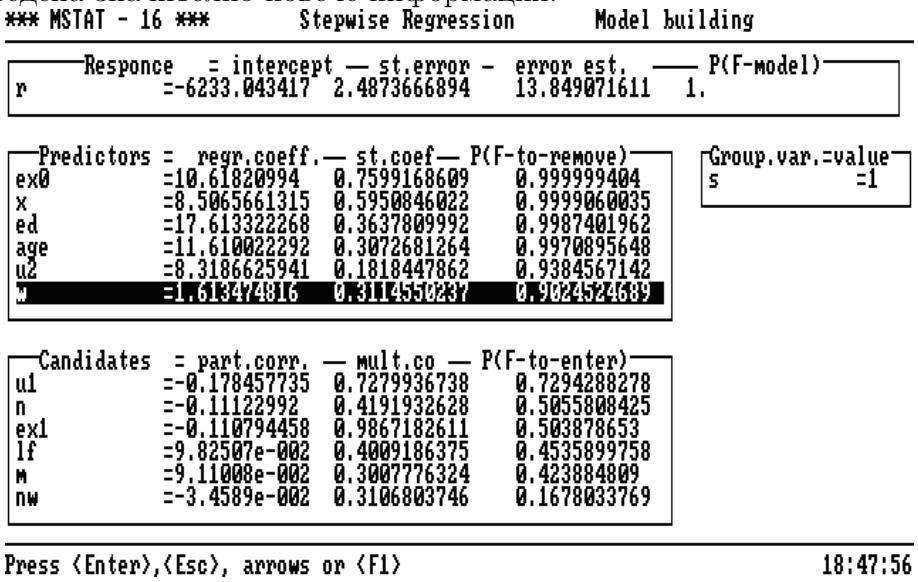
Вече имаме възможност да видим и свободния член на регресионното уравнение

и истинските коефициенти.

С възможностите на Правец повече не беше възможно.

9.1.2 MCTAT-16

В края на 80-те години от същия колектив (Vandev, Dimitrov, Isa, Kaishev, Kovachev, Mateev, and Petrov 1989) бе реализирана версия на пакета за 16-битов компютър. Тук екранът е вече по-голям (80 символа на ред) и беше възможно на него да бъде изведена значително повече информация.



Фиг. 9.3: Регресия в MCTAT

За променливата отклик се показват:

- свободния член и стандартната му грешка;
- оценката на σ за модела;
- P-стойността на хипотезата, че всички коефициенти (освен свободния член) са 0.

За предикторите (променливи в модела)са показани:

- коефициента в модела уравнение;
- стандартизирания коефициент;
- $P(F-to-remove)$ - P-стойността на хипотезата, че коефициентът е 0 - обикновено се изключват от модела променливите с минимална вероятност ($< .95$). Това точно същата T-вероятност, която видяхме в СТАТЛАБ.

За кандидатите за предиктори са показани:

- коефициентът на частична корелация с отклика;
- коефициентът на множествена корелация с предикторите;
- $P(F-to-enter)$ P-стойността на хипотезата, че коефициентът е 0 (T-вероятността), ако променливата стане предиктор (обикновено се включват в модела кандидатите с максимална вероятност $> .95$).

Променливите в тези два прозореца са наредени по вероятностите си.
При наличие на групираща променлива, в най-десния прозорец се показват стойностите ѝ.
Промяната им довежда до преизчисляване на свободния член и неговата стандартна грешка.

Управлението се извършва с клавиши - (премества се осветеното поле) по екрана и действия се извършват с натискане на клавиши

- <Enter> за предприемане на действие (както в СТАТЛАБ),
- <F1> за помощ и препоръка,
- <Esc> за напускане и запазване на модела.

9.2 Автоматична процедура

В стандартните статистически програми Statistica, SPSS, BMDP, Minitab и др. стъпковата регресия е автоматична, неинтерактивна процедура. Например, в Statistica (след избора на отклик, независими променливи и начин за пресмятане на ковариационната матрица) се избира един от методите:

- стандартна регресия,
- напред (forward stepwise),
- назад (backward stepwise).

И двата стъпкови метода се нуждаят от задаване на граници на $P(F-to-enter)$ и $P(F-to-remove)$. Тъй като смисълът на проверките е един и същ, е естествено, границите P_r, P_e да бъдат наредени:

$$P(F-to-remove) < P_r < P_e < P(F-to-enter). \tag{9.1}$$

Променливата се въвежда, когато е изпълнено дясното неравенство, и изважда от модела, когато е изпълнено лявото.

От съображения за "икономия", или не знам защо, вместо неравенствата 9.1 се използуват "еквивалентните":

$$F-to-remove < F_r < F_e < F-to-enter \tag{9.2}$$

и вместо критичните вероятности потребителят трябва да въведе границите (критичните стойности) F_r, F_e. Помислете, защо неравенствата 9.1 и 9.2 не са еквивалентни.

В пакета BMDP стойностите по подразбиране са $F_r = 3.9, F_e = 4.0$.

В Statistica - $F_r = 1, F_e = 2$ и се запомнят от пакета до следващото му изпълнение независимо с какви данни работите.

9.3 SWEEP оператор

Този метод е описан най-пълно в статията (Jennrich 1977), посветена на стъпкова регресия в сборника от статии (Einslein, Ralston, and Wilf 1977). В всички пакети, описани по-горе, изчисленията се извършват на основата на този оператор.

Определение 9.1 *Нека A е квадратна матрица с елементи $a_{i,j}$. Означаваме с S_k следното преобразование:*

$$S_k = \begin{cases} \tilde{a}_{k,k} & := -1/a_{k,k} \\ \tilde{a}_{i,k} & := a_{i,k}/a_{k,k} \\ \tilde{a}_{k,j} & := a_{k,j}/a_{k,k} \\ \tilde{a}_{i,j} & := a_{i,j} - a_{i,k}a_{k,j}/a_{k,k} \end{cases} \tag{9.3}$$

Обратното преобразование S_k^{-1} се задава с формулите:

$$S_k^{-1} = \begin{cases} \tilde{a}_{k,k} & := -1/a_{k,k} \\ \tilde{a}_{i,k} & := -a_{i,k}/a_{k,k} \\ \tilde{a}_{k,j} & := -a_{k,j}/a_{k,k} \\ \tilde{a}_{i,j} & := -a_{i,j} - a_{i,k}a_{k,j}/a_{k,k} \end{cases} \tag{9.4}$$

Както се вижда от определението и операторът S_k и неговият обратен са приложими само когато съответният диагонален елемент a_{kk} на матрицата A е ненулев.

9.3.1 Теорема

Нека първо сумираме по-очевидните свойства на операторите S_k.

- $S_k S_k^{-1} = S_k^{-1} S_k = I$,
- $S_k S_j = S_j S_k$.

Сега ще докажем една лема за свойството на "размяна", което тези оператори притежават.

Лема 9.1 *Нека $U = (u_1, u_2, \ldots, u_m)$ и $V = (v_1, v_2, \ldots, v_m)$ са матрици с еднакви размерности и A е квадратна матрица, такава че $V = UA$. Да означим*

- $\tilde{A} = S_k A$ (предполагаме, че това е възможно),
- $\tilde{V}$ се получава от V със замяна на k-тата колона: $\tilde{v}_k = -u_k$;
- $\tilde{U}$ се получава от U със замяна на k-тата колона: $\tilde{u}_k = v_k$;

Тогава $\tilde{V} = \tilde{U}\tilde{A}$.

Доказателство: За всяко j е изпълнено равенството: $v_j = \sum_{i=1}^{n} u_i a_{ij}$.
Следователно, при $j = k$ имаме $u_k = a_{kk}^{-1}(v_k - \sum_{i \neq k} u_i a_{ik})$ и получаваме

$$\tilde{v}_k = -u_k = -a_{kk}^{-1}(v_k - \sum_{i \neq k} u_i a_{ik}) =$$

$$= \tilde{a}_{kk}\tilde{u}_k + \sum_{i \neq k} \tilde{u}_i \tilde{a}_{ik} = \sum_{i=1}^{n} \tilde{u}_i \tilde{a}_{ik}.$$

Аналогично, при $j \neq k$

$$\tilde{v}_j = v_j = \sum_{i \neq k} u_i a_{ij} + u_k a_{kj} = \sum_{i \neq k} u_i a_{ij} + a_{kk}^{-1}(v_k - \sum_{i \neq k} u_i a_{ik}) a_{kj}$$

$$= \tilde{u}_k \tilde{a}_{kj} + \sum_{i \neq k} \tilde{u}_i (a_{i,j} - a_{i,k} a_{k,j}/a_{k,k}) = \sum_{i=1}^{n} \tilde{u}_i \tilde{a}_{ij}.$$

□

Теорема 9.1 *Нека матрицата A е зададена блочно*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

и операторът S_k е приложен към първите k номера. Тогава резултатът може да бъде записан във вида:

$$\tilde{A} = \begin{pmatrix} -A_{11}^{-1} & A_{11}^{-1} A_{12} \\ A_{21} A_{11}^{-1} & A_{22} - A_{21} A_{11}^{-1} A_{12} \end{pmatrix}$$

Доказателство: Нека разгледаме тривиалното равенство: $A = IA$ и го препишем в блочен вид (аналогично на лемата $V = UA$):

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} I_{11} & 0 \\ 0 & I_{22} \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}.$$

Като приложим сега операторите към първите k диагонални елементи (естествено ако това е възможно) на A_{11}, получаваме $\tilde{A}$. Благодарение на лема 9.1 можем да запишем останалите елементи на равенството $\tilde{V} = \tilde{U}\tilde{A}$:

$$\begin{pmatrix} -I_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} = \begin{pmatrix} A_{11} & 0 \\ A_{21} & I_{22} \end{pmatrix} \begin{pmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ \tilde{A}_{21} & \tilde{A}_{22} \end{pmatrix}.$$

Като извършим блочното умножение лесно получаваме равенствата:

$$\begin{array}{ll} -I_{11} = A_{11}\tilde{A}_{11}, & A_{12} = A_{11}\tilde{A}_{12}, \\ 0 = A_{21}\tilde{A}_{11} + \tilde{A}_{21}, & A_{22} = A_{21}\tilde{A}_{12} + \tilde{A}_{22}. \end{array}$$

Сега остава само да ги решим, което е тривиално. □

9.3.2 Тълкуване

Нека ни е зададена матрицата Z от n наблюдения над m параметъра и ние сме я центрирали, така че $e'Z = 0$. Нека разделим матрицата $Z = (X, Y)$ на две: X - матрица с първите k колони и с Y - матрица с останалите $m - k$ и се интересуваме от решението на регресионната задача: $Y = XB + E$.

Нека сме приложили SWEEP към първите k диагонални елемента на крос-продукт матрицата $R = Z'Z$. Естествено е резултатът $S(R)$ пак да запишем в блочно диагонален вид:

$$R = \begin{pmatrix} X'X & X'Y \\ Y'X & Y'Y \end{pmatrix}, \qquad S(R) = \begin{pmatrix} -(X'X)^{-1} & B \\ B' & C \end{pmatrix}.$$

Нека изтълкуваме смисъла на елементите на матрицата $S(R)$.

$$B = (X'X)^{-1}X'Y, \qquad C = Y'Y - Y'X(X'X)^{-1}X'Y$$

Значи в B са регресионните коефициенти, а в C остатъчните суми от квадрати (или ковариации). С една дума в матрицата $S(R)$ се съдържа цялата необходима информация за модела.

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.

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;
- да разгледаме един популярен пример с данни.

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.

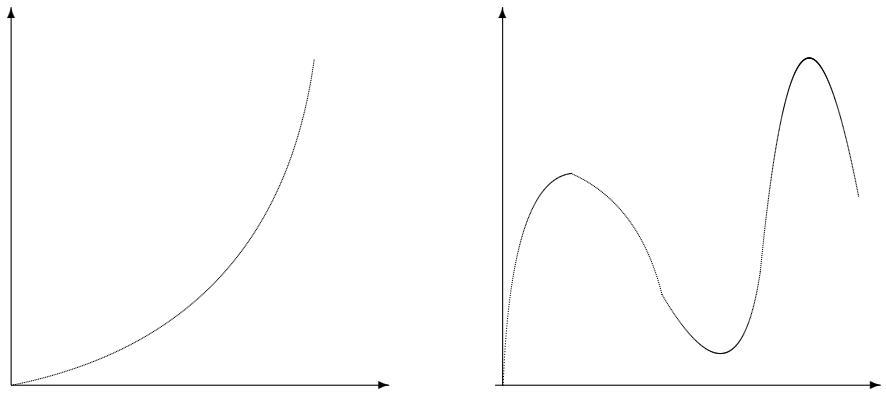
Полиномният модел може да се запише във формата:

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

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

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

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



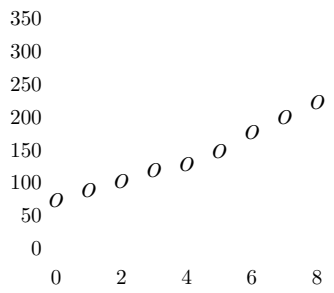
Фиг. 10.1: Криволинейни данни

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

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

Приготвянето на такава матрица изисква пресмятането на твърде голям брой суми. Още по-ужасно 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 - съществоното 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 ($j < n-2$):

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

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:

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

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.

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

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

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

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

10.3 Оптимална степен

Нека се опитаме да намерим "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:

$$y = \sum_{k=0}^{N-1} \tilde{b}_k \tilde{P}_k, \quad \tilde{b}_k = (y, \tilde{P}_k) \tag{10.6}$$

$$\|y\|^2 = \sum_{i=1}^{N} y_i^2 = \sum_{k=0}^{N-1} \tilde{b}_k^2. \tag{10.7}$$

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:

$$\tilde{b}_k = \|P_k\|\hat{b}_k. \tag{10.8}$$

От равенства (10.6-10.8) следват търсените формули:

$$\hat{b}_k = \frac{\sum_{i=1}^{N} y_i P_k(x_i)}{\sum_{i=1}^{N} P_k^2(x_i)}, \qquad \sigma^2(\hat{b}_k) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{N} P_k^2(x_i)}, \tag{10.9}$$

$$\hat{\sigma}^2 = \frac{1}{N-n-1}\sum_{k=n+1}^{N-1} \tilde{b}_k^2 = \frac{1}{N-n-1}\sum_{k=n+1}^{N-1} \hat{b}_k^2 \sum_{i=1}^{N} P_k^2(x_i). \tag{10.10}$$

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 за проекторите 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.

	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

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 (даже и с оптималния 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 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, а f_n - стойността на статистиката).

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:

$$n = \min\{k : \alpha_k \leq \alpha\} - 1.$$

Тема 11

Анализ на остатъците

В тази тема ще се спрем на един от най-популярните методи за проверка на модела в многомерната регресия. Ще разгледаме

- методите за проверка на хипотезата за разпределението на грешката;
- методи за откриване на несъстоятелни наблюдения;
- проверка за наличие на различни модели в данните.

Най-добрата книга по въпроса е на (Attkinson 1990), много полезна информация и примери могат да бъдат намерени в (Кокс и Снелл 1984) и (Дрейпер и Смит 1973).

11.1 Разпределение на остатъците

Тук ще се опитаме да сумираме свойствата на модела, които следват от представянето ($X = (x_1, x_2, \ldots, x_n)'$):

$$Y = Xa + E, \quad \hat{Y} = HY, \quad H = X(X'X)^{-1}X'. \tag{11.1}$$

При това за простота ще смятаме, че матрицата от данни е центрирана: $e'X = 0, e'Y = 0$.

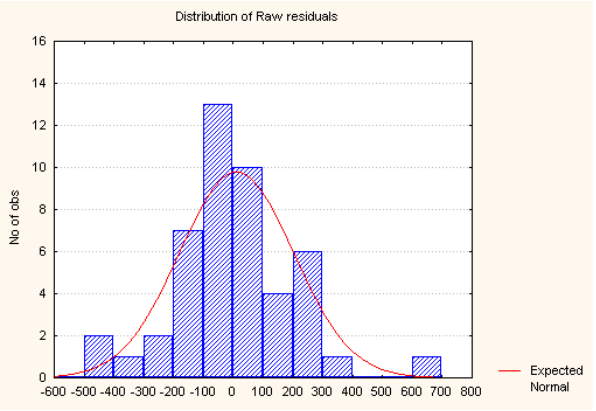
Тъй като остатъците се определят като: $\hat{R} = Y - \hat{Y}$ получаваме:

$$\hat{Y} = HY = H(Xa + E) = Xa + HE$$

$$\hat{R} = Y - \hat{Y} = (I - H)Y = (I - H)E$$

Тъй като матрицата H е проектор, то и $I - H$ е такъв. Значи при верен модел $\hat{R}$ имат изродено гаусово разпределение в подпространство на R^n с размерност $n - p$, където $p = rang(X)$.

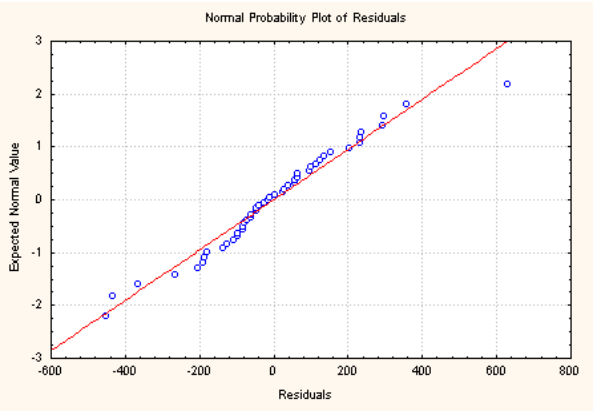
Проверка за нормалност



Фиг. 11.1: Хистограма на остатъците

На хистограмата, при значителен брой наблюдения лесно се забелязват както отклонения от нормалността, така и наличие на големи остатъци.

Може да се приложи и χ^2-тест за нормалност.



Фиг. 11.2: Нормална хартия

Вероятно това е най-възприетата и лесна проверка. Нормалната хартия има и вариант с отстранен тренд, който е още по-показателен. Тежки опашки се проявяват в S образно изменение, а обратното S се получава при разпределение близко до равномерното.

Статистика на Дърбин - Уотсън

Статистиката на Дърбин - Уотсън е просто пресметната автокорелация на остатъците и служи за проверка за наличие на друга, не обусловена от модела зависимост,

Durbin-Watson d (coxsnell.sta) serial correlation of residuals		
	Durbin-Watson d	Serial Corr.
Estimate	1.807608	.086171

Таблица 11.1: Durbin-Watson

11.2 Видове остатъци

Анализирането на остатъците е сред най-важните и използувани методи за откриване на несъстоятелни наблюдения. Обикновено това са наблюдения с рязко отклоняващи се стойности било в някои от предикторите, било в отклика.

Причините за поява на такива наблюдения може да са различни:

- грешка при записване на стойността,
- наблюдението е от друга група.

Във всеки случай те не се съгласуват с модела. За да разгледаме влиянието на отделните наблюдения да означим с $h_{ii}, i = 1, 2, \ldots, n$ диагоналните елементи на H. Тогава:

$$h_{ii} = x_i'(X'X)^{-1}x_i.$$

При нарастване на броя на наблюденията естествено $h_{ii} \to 0$.

От друга страна да си спомним как се оценява дисперсията на конкретно наблюдение y в точката x и я приложим за оригиналните наблюдения:

$$\sigma_i^2 = \hat{\sigma}^2(1 + \frac{1}{n} + h_{ii}). \tag{11.2}$$

където

$$\hat{\sigma}^2 = \boxed{\text{RMS}} = \frac{1}{n-p}\|Y - \hat{Y}\|^2 =$$

$$= \frac{1}{n-p}\|(I - H)E\|^2 \stackrel{d}{=} \frac{\sigma^2}{n-p}\sum_{i=1}^{n-p}\xi_i^2.$$

Да си спомним, че $e_i \in N(0, \sigma^2)$ и $\xi_i \in N(0, 1)$.

Прието е, освен обикновените остатъци r_i, да се изучават следните диагностични величини:

- стандартизирани остатъци - $r_i' = r_i/\sigma_i$
- jack-knive остатъци r_i^* - пресметнати по модел от който е изключено конкретното i-то наблюдение;
- разстояния на Махалонобис на предикторите от центъра -

$$m_i^2 = nx_i'(X'X)^{-1}x_i = nh_{ii};$$

- разстояния на Cook - друга функция от h_{ii}, отразяваща влиянието на наблюдението върху коефициентите на регресията.

$$D_i = r_i'^2\frac{h_{ii}}{p(1 - h_{ii})}$$

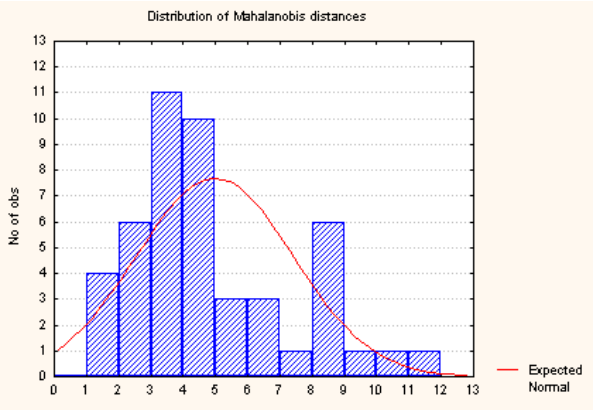
- модифицирани разстояния на Cook:

$$C_i = |r_i^*|\sqrt{\frac{n-p}{n}\frac{h_{ii}}{1 - h_{ii}}}$$

Така, ако всичко е наред с модела и нашите данни, трябва

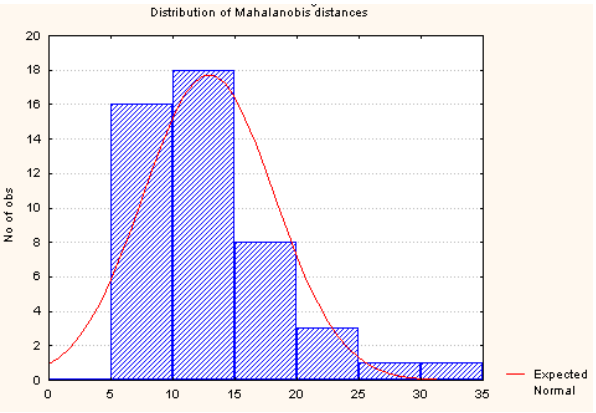
- остатъците от всички видове трябва да изглеждат добре на хистограмата;
- разстоянията на Махалонобис m_i не трябва да стават прекалено големи.
- разстоянията на Cook също не трябва да надхвърлят определени стойности.

За данните с престъпността, обаче нещата не изглеждат така добре. Остатъците от всички категории са добре разпределени, но



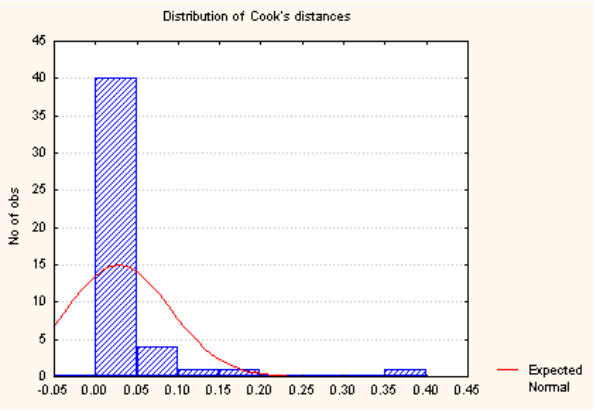
Фиг. 11.3: S не е включена

Разстоянията на Махалонобис m_i са прекалено големи. При това ясно се вижда двугърбо разпределение, което се дължи вероятно на южните щати - 17 на брой.



Фиг. 11.4: S включена

Ако се погрижим S да участвува сред предикторите, получаваме съвсем друга картинка.



Фиг. 11.5: Cook's

Разстоянията на Cook показват наличието на 1-2 наблюдения прекалено силно влияещи на модела

11.3 Групиране

Рисунката 11.3 ни навежда на мисълта, че ще трябва да разглеждаме по отделно северни и южни щати. Ако се опитаме да строим независими регресионни модели за двете групи наблюдения ще се окаже, че не разполагаме с достатъчно данни.

Прието е, модели в които се включват едновременно категорни и количестве-

ни предиктори да се групират под названието ковариационен анализ. С тях ще се запознаем в темата за Дисперсионен анализ. За съжаление, обаче в регресионната програма на пакета Statistica това не е възможно.

Тема 12

Дисперсионен и ковариационен анализи

Дисперсионният анализ (ANOVA) е част от статистиката, изучаваща влиянието на една или няколко групиращи променливи върху една количествена. В основата на дисперсионния анализ лежи възможността сумата от квадрати на отклонения на отклика SSY да бъде разложена на няколко независими суми от квадрати.

12.1 Понятия

Както и в регресията, е прието зависимата променлива се нарича отклик.

В дисперсионния анализ е възприето групиращата променлива да се нарича "фактор", стойностите й - "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.

12.1.1 Задачи и модели

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.

12.1.2 Планиране на експеримента

Най-голяма популярност ANOVA 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 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 ANOVA 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.

12.2 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:

$$\begin{aligned} y &= Z\mu + e \\ y_{ij} &= m + a_i + \epsilon_{ij}. \end{aligned} \tag{12.1}$$

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

$$\sum_i a_i = 0. \tag{12.2}$$

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

$$\sum_i \sum_j (y_{ij} - y_{..})^2 = \sum_i \sum_j (y_{i.} - y_{..})^2 + \sum_i \sum_j (y_{ij} - y_{i.})^2, \tag{12.3}$$

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

$$F = (SSm/(M-1))/(SSr/(N-M)) \tag{12.4}$$

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.

12.3 Множествени сравнения

В много случаи ни е необходимо да направим едновременно заключение за много от параметрите наведнаж. Естествено и тук проверката на различни хипотези се оказва еквивалентна на построяването на доверителни интервали.

12.3.1 Неравенство на Бонферони

Можем да използуваме следното знаменито неравенство на Бонферони:

$$\mathsf{P}(\overline{\cap I_i}) \geq \prod \mathsf{P}(\overline{I_i}). \tag{12.5}$$

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.

12.3.2 Метод на Тюки

Да разгледаме, например модела (12.1). Да си поставим следните задачи:

1. Да намерим доверителни интервали I_i за параметрите $\beta_i = m + a_i$, такива, че

$$\mathsf{P}(\cap_i\{\beta_i \in I_i\}) \geq 1-\alpha, \tag{12.6}$$

2. Да намерим доверителни интервали $I_{i,j}$ за параметрите $a_i - a_j$, такива, че

$$\mathsf{P}(\cap_{i<j}\{a_i - a_j \in I_{i,j}\}) \geq 1-\alpha. \tag{12.7}$$

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:

$$\mathsf{P}(\cap_i\{|\beta_i - y_{i,.}| < Cs\}) \geq 1-\alpha, \tag{12.8}$$

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:

$$I_{i,j} = \{y_{i,.} - y_{j,.} - Cs, y_{i,.} - y_{j,.} + Cs\}. \tag{12.9}$$

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

12.3.3 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, че $k||a - \hat{a}||^2/\sigma^2$ е Хи-квадрат с $M - 1$ степени на свобода. Следователно

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

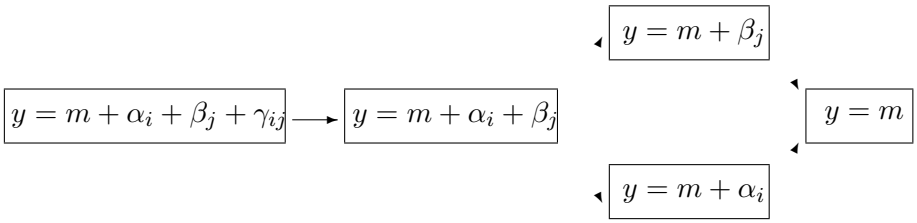
има F разпределение с $M-1$ и $M(k-1)$ степени на свобода. Така за всички контрасти y получаваме съвместни доверителни интервали:

$$I_a = \{\hat{y} - C^{1/2}s||c||, \hat{y} + C^{1/2}s||c||\}, \tag{12.10}$$

където $C = (M-1)f_{1-\alpha}k$.

12.4 Двуфакторен анализ

Тук вече можем да избираме измежду няколко възможни модела:



Фиг. 12.1: Връзки между моделите

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:

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

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 теоремата за сравняване на моделите (вж. модел (7.7 и теорема 7.1) трябва да е изпълнено неравенството: $km > k + m - 1$. Това е винаги така, стига да е изпълнено неравенството: $2 \leq k, m$.

12.5 Примери

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

Пример 12.1 *Пример за еднофакторен дисперсионен анализ. (пример 13.2)*

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:

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 12.1: Данни за торенето - ANOVA1

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

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 12.2 *Двуфакторен дисперсионен анализ*

Ще разгледаме още един пример от (Dunn and Clark 1974) показан на таблица 13.3 . 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
14.3	1	1
18.1	1	2
17.6	1	3
...	...	...

Получаваме дисперсионна таблица на двуфакторния дисперсионен анализ - таблица 12.2.

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 12.2: ANOVA2 -таблица

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 12.3 *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.

12.6 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:

$$y = Z\mu + Xa + e \tag{12.11}$$

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 частен случай от така наречения ковариационен анализ (виж (Шеффе 1963)), където може и векторът a да зависи 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 на регресионния анализ матрицата $X'X/n$ трябва да се замести с вътрешно-груповата ко-

вариационна матрица

$$V = \frac{1}{n-G} \sum_{k=1}^{G} \sum_{i=1}^{m_k} (x_i^k - \bar{x}^k)(x_i^k - \bar{x}^k)',$$

а числото n да се замести с $n - G$.

Тема 13

Приложение

13.1 Примерни данни

В това приложение ще приведем данните и техните описания, използувани за илюстрация на статистическите методи.

Пример 13.1 *П1.1. Данни за престъпността в 47 щата на САЩ (Кокс и Снелл 1984), стр. 183 - 185*

Данните са от отчет за престъпността, съставен от ФБР и се отнасят за 1960 календарна година. Те съдържат следните 14 променливи за 47 щата:

R	-	брой регистрирани от полицията правонарушения на 1 млн. жители;
AGE	-	мъже (на 1000) на възраст от 14 до 24 години;
S	-	1 за южен щат и 0 за северен;
ED	-	10 пъти средния брой години на обучение за лицата на възраст над 25 години;
EX0	-	разходи за полицията за 1960 г. на глава от населението;
EX1	-	разходи за полицията за 1959 г. на глава от населението;
LF	-	работещи мъже (на 1000) на възраст от 14 до 24 г.;
M	-	брой мъже на 1000 жени;
N	-	население на щата в 100 хил.;
NW	-	брой цветнокожи на 1000 д.;
U1	-	безработни мъже (на 1000) в градовете от 14 до 24 г.;
U2	-	безработни мъже (на 1000) в градовете от 35 до 39 г.;
W	-	благосъстояние измерено като медиана на средния доход (в 10 дол.);
X	-	брой на семействата (на 1000) с доход под половината на медианния.

R	AGE	S	ED	EX0	EX1	LF	M	N	NW	U1	U2	W	X
791	151	1	91	58	56	510	950	33	301	108	41	394	261
1635	143	0	113	103	95	583	1012	13	102	96	36	557	194
578	142	1	89	45	44	533	969	18	219	94	33	318	250
1969	136	0	121	149	141	577	994	157	80	102	39	673	167
1234	141	0	121	109	101	591	985	18	30	91	20	578	174
682	121	0	110	118	115	547	964	25	44	84	29	689	126
963	127	1	111	82	79	519	982	4	139	97	38	620	168
1555	131	1	109	115	109	542	969	50	179	79	35	472	206
856	157	1	90	65	62	553	955	39	286	81	28	421	239
705	140	0	118	71	68	632	1029	7	15	100	24	526	174
1674	124	0	105	121	116	580	966	101	106	77	35	657	170
849	134	0	108	75	71	595	972	47	59	83	31	580	172
511	128	0	113	67	60	624	972	28	10	77	25	507	206
664	135	0	117	62	61	595	986	22	46	77	27	529	190
798	152	1	87	57	53	530	986	30	72	92	43	405	264
946	142	1	88	81	77	497	956	33	321	116	47	427	247
539	143	0	110	66	63	537	977	10	6	114	35	487	166
929	135	1	104	123	115	537	978	31	170	89	34	631	165
750	130	0	116	128	128	536	934	51	24	78	34	627	135
1225	125	0	108	113	105	567	985	78	94	130	58	626	166
742	126	0	108	74	67	602	984	34	12	102	33	557	195
439	157	1	89	47	44	512	962	22	423	97	34	288	276
1216	132	0	96	87	83	564	953	43	92	83	32	513	227
968	131	0	116	78	73	574	1038	7	36	142	42	540	176
523	130	0	116	63	57	641	984	14	26	70	21	486	196
1996	131	0	121	160	143	631	1071	3	77	102	41	674	152
342	135	0	109	69	71	540	965	6	4	80	22	564	139
1216	152	0	112	82	76	571	1018	10	79	103	28	537	215
1043	119	0	107	166	157	521	938	168	89	92	36	637	154
696	166	1	89	58	54	521	973	46	254	72	26	396	237
373	140	0	93	55	54	535	1045	6	20	135	40	453	200
754	125	0	109	90	81	586	964	97	82	105	43	617	163
1072	147	1	104	63	64	560	972	23	95	76	24	462	233
923	126	0	118	97	97	542	990	18	21	102	35	589	166
653	123	0	102	97	87	526	948	113	76	124	50	572	158
1272	150	0	100	109	98	531	964	9	24	87	38	559	153
831	177	1	87	58	56	638	974	24	349	76	28	382	254
566	133	0	104	51	47	599	1024	7	40	99	27	425	225
826	149	1	88	61	54	515	953	36	165	86	35	395	251
1151	145	1	104	82	74	560	981	96	126	88	31	488	228
880	148	0	122	72	66	601	998	9	19	84	20	590	144
542	141	0	109	56	54	523	968	4	2	107	37	489	170
823	162	1	99	75	70	522	996	40	208	73	27	496	224
1030	136	0	121	95	96	574	1012	29	36	111	37	622	162
455	139	1	88	46	41	480	968	19	49	135	53	457	249
508	126	0	104	106	97	599	989	40	24	78	25	593	171
849	130	0	121	90	91	623	1049	3	22	113	40	588	160

Пример 13.2 *Еднофакторен дисперсионен анализ (вж.(Dunn and Clark 1974))*

Един експериментатор желае да изучи влиянието на различни торове върху добива. За целта той избира 24 еднакви по форма и площ полета и ги засява със своята култура. Върху полетата той ще изпита 4 вида торене:

1.	2.	3.	4.
Никакво торене,	$K_2O + N$	$K_2O + P_2O_5$	$N + P_2O_5$

Получените данни за добива изглеждат така:

тор	1	2	3	4	5	6
1.	99	40	61	72	76	84
2.	96	84	82	104	99	570
3.	63	57	81	59	64	396
4.	79	92	91	87	78	498

Пример 13.3 *Двуфакторен дисперсионен анализ (вж. (Dunn and Clark 1974))*

Тук 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. Полученият добив е записан в следната таблица:

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

Пример 13.4 *Двуфакторен дисперсионен анализ (вж.(Готтсданкер 1982), стр.46 и стр.265)*

Изследва се скоростта на реакцията на човек при подаване на светлинен (A,C) и звуков (B,D) сигнали. Откликът (скоростта на реакцията) се мери в милисекунди. При A и B избор няма и реакцията е "проста". B и D описват реакция с "избор" - условията на експеримента са такива, че изпитваният трябва да избере между два типа светлинен и звуков сигнали.

X	A	B	C	D
1	223	181	304	272
2	184	194	268	264
3	209	173	272	256
4	183	153	262	269
5	180	168	283	285
6	168	176	265	247
7	215	163	286	250
8	172	152	257	245
9	200	155	279	251
10	191	156	275	261
11	197	178	268	250
12	188	160	254	228
13	174	164	245	257
14	176	169	253	214
15	155	155	235	242
16	115	122	260	222
17	163	144	246	234

Пример 13.5 *Спектрален анализ вж.(Кендалл 1981), стр. 18.*

Разстоянията в 1000 мили, изминати от самолетите на Обединеното Кралство за един месец в периода от 1963 до 1970 г.

	1963	1964	1965	1966	1967	1968	1969	1970
яну	6827	7269	8350	8186	8334	8639	9491	10840
фев	6178	6775	7829	7444	7899	8772	8919	10436
март	7084	7819	8829	8484	9994	10894	11607	13589
апр	8162	8371	9948	9864	10078	10455	8852	13402
май	8462	9069	10638	10252	10801	11179	12537	13103
юни	9644	10248	11253	12282	12950	10588	14759	14933
юли	10466	11030	11424	11637	12222	10794	13667	14147
авг	10748	10882	11391	11578	12246	12770	13731	14057
септ	9963	10333	10665	12417	13281	13812	15110	16234
окт	8194	9109	9396	9637	10366	10857	12185	12389
ноем	6848	7685	7775	8094	8730	9290	10645	11595
дек	7027	7602	7933	9280	9614	10925	12161	12772

Пример 13.6 *Крос-спектрален анализ*

Австралийският ентомолог А.Николсон е провел обширни изследвания (от 1950 г.) върху популации на мухата Lucilia cuprina при различни условия. Тук привеждаме част от негови наблюдения (експеримент L97), взети от (D.R.Brillinger, J.Guckenkeimer, P.Guttorp, and G.Oster 1980), pp.65–90.

Данните са за популация, която е била при едни и същи условия (контролни наблюдения). Първият временен ред са наличните яйца, вторият - особите достигнали зряла възраст, а третият - брой на умрелите насекоми. Общият брой насекоми може да получим, ако към общия брой от предното наблюдение добавим достигналите зряла възраст и извадим умрелите. В началния момент общият брой е 948. Наблюденията са правени през ден.

Продължителността на отделните фази на развитие на насекомото са: яйце - от 12 до 24 часа, ларва - 5 до 10 дни, пашкул - 6 до 8 дни, развитие на насекомото - 4 дни, живот на възрастно насекомо - 1 до 35 дни.

	EGGS	EMERGING	DEATHS		EGGS	EMERGING	DEATHS
1.	0	948	0	65.	3234	326	33
2.	0	4	10	66.	1574	729	24
3.	0	0	31	67.	7445	1072	60
4.	0	0	53	68.	2019	1747	165
5.	2149	0	57	69.	672	3794	1404
6.	4627	0	125	70.	0	2576	2788
7.	4523	0	172	71.	0	2220	2427
8.	6030	0	107	72.	0	3248	2385
9.	2684	0	149	73.	0	4808	3521
10.	3373	0	102	74.	33	1052	4123
11.	446	1763	108	75.	26	1270	2186
12.	133	4487	53	76.	32	95	1335
13.	17	1830	2091	77.	0	3	981
14.	56	7952	5005	78.	321	0	510
15.	58	1953	4264	79.	230	0	521
16.	0	2419	3056	80.	392	18	213
17.	6	1966	2266	81.	253	10	152
18.	25	132	1930	82.	324	19	106
19	30	36	1550	83.	1369	27	63
20.	0	16	1025	84.	1252	120	20
21.	0	85	211	85.	977	435	17
22.	548	17	331	86.	1336	298	34
23.	461	0	391	87.	1450	247	51
24.	1638	5	163	88.	2520	456	77
25.	1524	41	175	89.	3057	897	230
26.	2338	10	97	90.	302	1459	436
27.	1473	0	60	91.	166	721	540
28.	3287	494	34	92.	20	1258	1092
29.	1367	424	27	93.	27	1045	979
30.	617	1541	49	94.	70	2300	1127
31.	936	1317	61	95.	4	1782	1294
32.	112	2016	1160	96.	38	1252	1726
33.	91	1496	1504	97.	5	232	2093
34.	0	2898	1992	98.	48	53	1132
35.	0	855	1884	99.	115	23	748
36.	0	469	1859	100.	364	50	199
37.	1	925	1291	101.	1034	30	201
38.	47	522	1211	102.	1295	39	172
39.	8	55	727	103.	2383	0	144
40.	0	45	425	104.	2435	33	188
41.	77	0	364	105.	2535	113	47
42.	598	0	238	106.	1701	383	94
43.	6814	0	146	107.	1601	928	51
44.	1537	5	77	108.	580	1201	35
45.	1296	45	47	109.	460	2175	570
46.	451	2	35	110.	363	1668	920
47.	1863	47	39	111.	0	2980	1609
48.	2408	5205	14	112.	106	1872	2064
49.	358	1496	82	113.	80	1228	2876
50.	847	1957	3189	114.	0	461	1692
51.	14	855	2309	115.	303	660	1462
52.	0	501	1536	116.	755	165	953
53.	62	1771	1075	117.	536	16	447
54.	4	1607	1033	118.	602	106	524
55.	39	674	1007	119.	688	78	311
56.	0	197	1791	120.	1244	0	75
57.	0	700	1486	121.	1662	247	157
58.	107	0	581	122.	5050	623	179
59.	172	16	205	123.	926	479	163
60.	828	43	216	124.	3011	775	175
61.	1216	0	183	125.	1461	678	479
62.	2109	37	146	126.	526	1078	524
63.	3460	0	82	127.	107	1504	916
64.	2438	105	71	128.	136	4771	1081

Пример 13.7 *Нелинейна регресия вж.(Химмелъблау 1973), стр. 432.*

За реакцията на хидриране в проточен реактор са получени следните данни:

x	20	30	35	40	50	55	60
y	.0680	.0858	.0939	.0999	.1130	.1162	.1190

Тук x е пълното налягане в г/см^2, а y е скоростта на реакцията в гмол/ч.

13.2 Таблици

Таблица 13.1: Хи-квадрат разпределение (квантили)

df/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 13.2: Нормално разпределение

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 13.3: Разпределение на Стюдент (двустранни квантили)

	.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 13.4: Квантили на разпределение на Фишер ($p = .95$)

df1	df2=1	2	3	4	5	6	7	8	9
1	161.4476	18.5128	10.1280	7.7086	6.6079	5.9874	5.5914	5.3177	5.1174
2	199.5000	19.0000	9.5521	6.9443	5.7861	5.1433	4.7374	4.4590	4.2565
3	215.7073	19.1643	9.2766	6.5914	5.4095	4.7571	4.3468	4.0662	3.8625
4	224.5832	19.2468	9.1172	6.3882	5.1922	4.5337	4.1203	3.8379	3.6331
5	230.1619	19.2964	9.0135	6.2561	5.0503	4.3874	3.9715	3.6875	3.4817
6	233.9860	19.3295	8.9406	6.1631	4.9503	4.2839	3.8660	3.5806	3.3738
7	236.7684	19.3532	8.8867	6.0942	4.8759	4.2067	3.7870	3.5005	3.2927
8	238.8827	19.3710	8.8452	6.0410	4.8183	4.1468	3.7257	3.4381	3.2296
9	240.5433	19.3848	8.8123	5.9988	4.7725	4.0990	3.6767	3.3881	3.1789
10	241.8817	19.3959	8.7855	5.9644	4.7351	4.0600	3.6365	3.3472	3.1373
11	242.9835	19.4050	8.7633	5.9358	4.7040	4.0274	3.6030	3.3130	3.1025
12	243.9060	19.4125	8.7446	5.9117	4.6777	3.9999	3.5747	3.2839	3.0729
13	244.6898	19.4189	8.7287	5.8911	4.6552	3.9764	3.5503	3.2590	3.0475
14	245.3640	19.4244	8.7149	5.8733	4.6358	3.9559	3.5292	3.2374	3.0255
15	245.9499	19.4291	8.7029	5.8578	4.6188	3.9381	3.5107	3.2184	3.0061
20	248.0131	19.4458	8.6602	5.8025	4.5581	3.8742	3.4445	3.1503	2.9365
40	251.1432	19.4707	8.5944	5.7170	4.4638	3.7743	3.3404	3.0428	2.8259
80	252.7237	19.4832	8.5607	5.6730	4.4150	3.7223	3.2860	2.9862	2.7675

df1	df2=10	11	12	13	14	15	20	40	80
1	4.9646	4.8443	4.7472	4.6672	4.6001	4.5431	4.3512	4.0847	3.9604
2	4.1028	3.9823	3.8853	3.8056	3.7389	3.6823	3.4928	3.2317	3.1108
3	3.7083	3.5874	3.4903	3.4105	3.3439	3.2874	3.0984	2.8387	2.7188
4	3.4780	3.3567	3.2592	3.1791	3.1122	3.0556	2.8661	2.6060	2.4859
5	3.3258	3.2039	3.1059	3.0254	2.9582	2.9013	2.7109	2.4495	2.3287
6	3.2172	3.0946	2.9961	2.9153	2.8477	2.7905	2.5990	2.3359	2.2142
7	3.1355	3.0123	2.9134	2.8321	2.7642	2.7066	2.5140	2.2490	2.1263
8	3.0717	2.9480	2.8486	2.7669	2.6987	2.6408	2.4471	2.1802	2.0564
9	3.0204	2.8962	2.7964	2.7144	2.6458	2.5876	2.3928	2.1240	1.9991
10	2.9782	2.8536	2.7534	2.6710	2.6022	2.5437	2.3479	2.0772	1.9512
11	2.9430	2.8179	2.7173	2.6347	2.5655	2.5068	2.3100	2.0376	1.9105
12	2.9130	2.7876	2.6866	2.6037	2.5342	2.4753	2.2776	2.0035	1.8753
13	2.8872	2.7614	2.6602	2.5769	2.5073	2.4481	2.2495	1.9738	1.8445
14	2.8647	2.7386	2.6371	2.5536	2.4837	2.4244	2.2250	1.9476	1.8174
15	2.8450	2.7186	2.6169	2.5331	2.4630	2.4034	2.2033	1.9245	1.7932
20	2.7740	2.6464	2.5436	2.4589	2.3879	2.3275	2.1242	1.8389	1.7032
40	2.6609	2.5309	2.4259	2.3392	2.2664	2.2043	1.9938	1.6928	1.5449
80	2.6008	2.4692	2.3628	2.2747	2.2006	2.1373	1.9217	1.6077	1.4477

Литература

Agresti, A. (1990). *Categorical Data Analysis*. Wiley.

Attkinson (1990). *Regression Analysis*. UK: Oxford.

D.R.Brillinger, J.Guckenkeimer, P.Guttorp, and G.Oster (1980). Empirical modelling of population time series data: The case of age and density dependent vital rates. In G. Oster (Ed.), *Some Mathematical Questions in Biology*, Providence, Rhode Island. AMS.

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

Einslein, K., A. Ralston, and H. S. Wilf (Eds.) (1977). *Statistical Methods for Digital Computers*. New York: John Wiley & Sons.

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

Jennrich, R. I. (1977). Stepwise regression. See Einslein, Ralston, and Wilf (1977), pp. 58–75.

StatSoft, I. (1998). *STATISTICA for Windows [Computer program manual]*. StatSoft, Inc., 2300 East 14th Street, Tulsa, OK 74104: Tulsa, OK.

Vandev, D., E. Dimitrov, S. Isa, V. Kaishev, B. Kovachev, P. Mateev, and P. Petrov (1989). *A program package for multivarate statistical analysis. STATLAB. MSTAT - 16.* Sofia: SPS.

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

Въндев, Д. и П. Матеев (1988). *Статистика с Правец 82*. София: Наука и Искуство.

Готтсданкер, Р. (1982). *Основы психологического эксперимента*. Москва: МГУ.

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

Кендалл, М. (1981). *Временные ряды*. Москва: Финансы и статистика.

Кокс, Д. и Э. Снелл (1984). *Прикладная статистика. Принципы и примеры*. Москва: Мир.

Поповъ, К. (1916). *Стопанска България през 1911 г. Статистически изследвания*. София: Държавна печатница.

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

Химмельблау, Д. (1973). *Анализ процессов статистическими методами*. Москва: Мир.

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