СОФИЙСКИ УНИВЕРСИТЕТ “КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
КАТЕДРА “ВЕРОЯТНОСТИ И СТАТИСТИКА“

доц. ДИМИТЪР Л. ВЪНДЕВ

Записки
по
ВЕРОЯТНОСТИ и СТАТИСТИКА

за физици

СОФИЯ, 1998

Съдържание

Увод

Названието Вероятности и Статистика цели да обхване най - простите идеи и методи на теорията на вероятностите като математически модел на статистиката.

Цел на тези кратки записки е да се даде едно допълнително пособие на студентите по физика, което да ги снабди с минимум сведения, присъствуващи във всички стандартни учебници по теория на вероятностите и някои книги по статистика.

Авторът би искал да се надява, че тези бегли записки ще събудят интереса към стохастичните методи поне у някои студенти. Затова и списъкът от литература е значително по-широк от необходимия за взимане на изпита.

Това е първи вариант на записките. Той съдържа много непълноти и грешки. Авторът ще бъде много благодарен на всеки, който си направи труда да му ги посочи.

януари, 1998

КОНСПЕКТ

1. Предмет на статистиката. Видове статистически данни. Крайна и безкрайна генерални съвкупности. Случайни извадки. Представяне на данните. Хистограми. Извадъчни разпределения.

2. Вероятност - емпирични основи. Алгебра на събитията. Аксиоматика.

3. Независимост и условна вероятност. Теорема за пълната вероятност. Теорема на Бейс.

4. Случайни величини и разпределения. Математическо очакване.

5. Дискретни сл.в. Пораждащи функции. Биномно, геометрично, хипергеометрично и поасоново разпределения. Теорема на Поасон

6. Количествени характеристики на разпределенията. Средно, мода и медиана. Квантили и квартили. Дисперсия и размах.

7. Схема на Бернули. Нормална апроксимация. Доверителни интервали за вероятност Доверителен интервал за неизвестна медиана.

8. Непрекъснати разпределения. Плътност и интеграл. Моменти. Характеристични функции.

9. Преобразования на сл.в. Нормално разпределение. Гама и Бета - разпределения

10. Проверка на хипотези. Грешки от първи и втори род. Мощност. Лема на Нейман - Пирсън. Хипотези за средната на популацията при известна дисперсия.

11. Точкови оценки и доверителни интервали. Разпределение на средното при известна дисперсия. Доверителни интервали за м.о. Размер на извадката.

12. Метод на максималното правдоподбие. Неизместеност. Метод на моментите

13. Многомерно нормално разпределение. Хи-квадрат разпределение. Доверителен интервал за неизвестна дисперсия.

14. Разпределение на Фишер и проверка на различие между дисперсии. Разпределение на средното и доверителен интервал за м.о. при неизвестна дисперсия. Разлика между средни - тест на Стюдент.

15. Непараметрични методи. Проверка на извадката. Двуизвадкови методи. Тест на сериите. Знаков и рангов тестове.

16. Хи-квадрат за проверка на съотвествие. Хи-квадрат тест за независимост на честотни таблици.

17. Линейна регресия. Предположения. Теорема на Гаус - Марков

18. Коефициент на детерминация. Проверка на хипотези за коефициентите. Грешки на оценките и предсказанието. Изследване на остатъците.

19. Апроксимация на плътности. Криви на Пирсън. Ядра на Розенблат - Парзен.

Тема 1

Основни задачи на статистиката

Основната задача на статистиката е да се интерпретират данните. Големите масиви от числа да се представят в обозрима форма. За това служат преди всичко описателните и графични методи. Когато е необходимо от малка част на изследваната съвкупност да се направи някакво заключение за цялото явление, се използува така наречената извадъчна статистика, основана изцяло на математическите модели на теория на вероятностите.

С времето тя се е оформила като отделна математическа дисциплина, наречена математическа статистика. В тези няколко лекции ще дадем бегла представа за математическите модели, които тя използува, и които са широко разпространени в практиката на цялата статистика.

1.1 Видове статистически данни. Случайни извадки.

В тази глава се описват най - простите (традиционните) средства за обработка на повечето от събираните в практиката данни.

1.1.1 Генерална съвкупност

Групата от данни (стойности) се нарича статистическа съвкупност, а всеки неин член е елемент или варианта на тази съвкупност. Множеството от всички възможни (мислими) варианти, които могат да се получат при многократно репродуциране на опробването, се нарича генерална съвкупност.

Наричаме изчерпателни такива данни, които напълно описват дадено явление. Такива са например данните от преброяването на населението. 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, ние ги наричаме генерална съвкупност. Така че генералната съвкупност е абстрактно множество от обекти представляващо цел на нашето изследване. В случая с преброяването, например, това

са жителите на страната.

1.1.2 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.

1.1.3 Случайни и неслучайни извадки

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

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

Например, една систематична извадка на дадено рудно находище предполага сондажи разположени равномерно по площта му. Изчерпателното преброяване на населението гарантира пълна информация за генералната съвкупност - населението на тази страна.

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.

1.1.4 Крайни и безкрайни генерални съвкупмости

Случайните извадки могат да се строят по следното правило:

Към генералната съвкупност прилагаме механизъм, който отбелязва за взимане в извадката точно един от обектите ѝ, с определен, еднакъв за всички обекти шанс.

Прилагайки този механизъм определен брой пъти, получаваме набора отбелязани обекти за включване в извадката. Може да се окаже при това, че някой обект е отбелязан два пъти.

Такива извадки наричаме извадки със връщане. Ако коригираме генералната си съвкупност, отстранявайки отбелязания обект след всяко действие на механизма, извадките наричаме извадки без връщане.

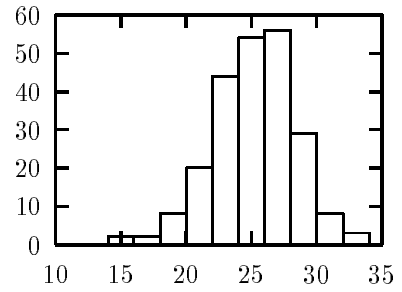
Когато обема на изследваното множество е голям, шансът в една случайна извадка да попадне два пъти същия обект е нищожна. Тогава за модел може да се приеме, че генералната съвкупност е безкрайна и двата типа извадки съвпадат.

1.2 Графично представяне на данни.

Представянето на данни всъщност е основна задача както на изчерпателната така и на извадъчната статистика. Информацията, която се съдържа в милионите числа трябва да бъде представена в обозрима форма, така че всеки да си представи основните качества на множеството обекти. Главна роля в това кондензиране на информация има графичното представяне. То е ефектно и в минимална степен при него се губи информация.

1.2.1 Хистограми

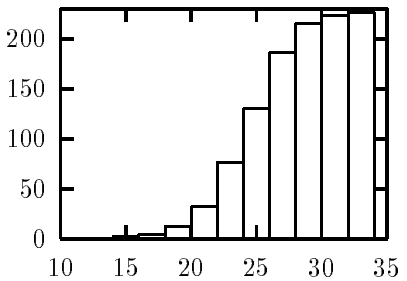
Хистограмата е основният вид за представяне на информацията за наблюдения върху числов признак. Тя се строи по просто правило. Избират се обикновено еднакво големи е не много на брой (5 - 20) еднакво големи прилежащи интервала покриващи множеството от стойности на наблюдавания признак. Те се нанасят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



Фигура 1.1: Съдържания на апатит

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

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



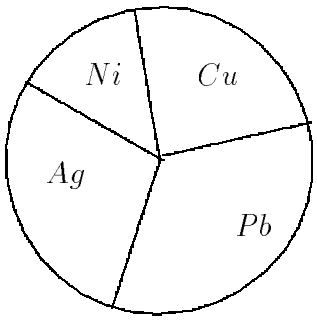
Фигура 1.2: Кумулативно представяне

Също така много удобна е така наречената кумулативна хистограма (фиг. 1.2). Тя се строи по натрупаните данни и позволява лесен отговор на въпроси от вида:

- каква е частта от наблюденията, попаднали под дадена граница;

- кое е числото под което са половината наблюдения – т.н. медиана.

1.2.2 Торта или секторна диаграма



Фигура 1.3: Секторна диаграма

Когато изследваме нечислови признаци, най - подходящото представяне е като процентно съдържание, например на различните минерали в една извадка. 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* или торти (piechart).

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

1.3 Извадъчно разпределение и дескриптивни статистики

1.3.1 Извадъчно разпределение

Определение 1.1 *Наредените по големина стоиности 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 $x_{(1)} = \mathbf{min}_I(\xi_i)$, а последната $x_{(n)} = \mathbf{max}_I(\xi_i)$. От това определение се вижда, че вариационния ред е векторна случайна величина — функция от вектора $\xi_1, \xi_2, \ldots, \xi_n$.

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 1.2 *Извадъчна функция на разпределение наричаме случайната функция:*

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

1.3.2 Дескриптивни статистики

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

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

Те лесно се изразяват чрез извадъчната функция на разпределение:

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

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

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

Ще видим по-нататък какъв е смисълът на дескриптивните статистики. Те описват локацията, мащаба и формата на разпределенията.

Тема 2

Аксиоматика

В тази лекция си поставяме следните цели:

- да разгледаме генезиса на понятието вероятност;
- да въведем събития и действия с тях;
- да определим вероятностно пространство;
- да дадем примери за прости вероятностни пространства.

2.1 Емпирични основи

Историята ни учи, че основите на понятието “шанс“ са твърде стари. Това, което хората първо са забелязали, е устойчивостта на средната аритметична с нарастването на броя наблюдения. В миналото например, мерките за дължина са се определяли с “усреднявяне“. В Англия, една от популярните мерки за дължина се е определяла като средна дължина на ходилото на първите 30 човека излизащи от черквата в неделя сутринта, в древния Египет - като общата дължина на определен брой семена от свещенно растение.

2.1.1 Класическа вероятност

Първите опити да се построи математически модел са свързани с понятието равен “шанс“. Предполага се, че даден опит има краен брой изходи, които са равноправни. При провеждане на опита се случва някой от тези изходи, при това всеки от тях може да се случи с еднакъв “шанс“. Най-простите примери за такава концепция са свързани с хазартните игри, където се хвърлят зарове или използуват добре разбъркани тестета карти.

Пример 2.1 *Хвърляме зар. Каква е вероятността да получим четно число?*

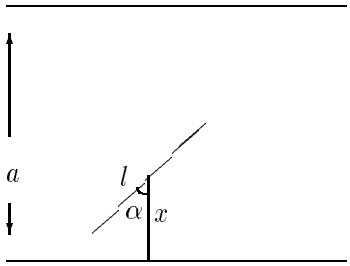
Рецептата е проста. Достатъчно е да преброим благоприятните изходи и разделим това число с броя на всички изходи:

$$\text{Класическа вероятност} = \frac{\text{Брой на благоприятните изходи}}{\text{Брой на всички възможни изходи}} \tag{2.1}$$

Така за нашата задача отговорът трябва да бъде $3/6 = 1/2$.

2.1.2 Геометрична вероятност

Пример 2.2 *(Задача на Бюфон) Хвърляме игла върху раирана покривка. Каква е вероятността иглата да пресече раето?*



Фигура 2.1: Иглата на Бюфон

За да решим задачата трябва да формализираме условията. Нека означим с l дължината на иглата и с a - разстоянието между раетата. За простота ще сметнем, че широчината на едно рае е 0. Да означим с x разстоянието от средата на иглата до по - близкото рае,а с α - острия ъгъл, които иглата сключва с перпендикуляра към същото рае. Тогава имаме $0 \leq x \leq a/2$ и $0 \leq \alpha \leq \pi/2$. Това са всички възможности. Благоприятните (когато иглата пресече раето) се определят от неравенството: $(l/2)cos\alpha > x$.

Рецептата е проста:

$$\text{Геометрична вероятност} = \frac{\text{Площ на благоприятните изходи}}{\text{Обща площ}}$$

Така, ако $l < a$, задачата се свежда до пресмятането на

$$p = \frac{2l}{a\pi}\int_0^{\frac{\pi}{2}} \cos\alpha d\alpha = \frac{2l}{a\pi}.$$

2.1.3 Честотна вероятност

Пример 2.3 *Хвърляме монета многократно. Колко пъти ще получим ези?*

Нека означим общия брой хвърляния с N,а броят на получените ези с M. Тогава честотата M/N на поява на ези би трябвало да клони към едно постоянно число:

$$\text{Честотна вероятност} = \lim_{n\to\infty} \frac{\text{Брой на благоприятните изходи}}{\text{Брой на извършените опити}} \tag{2.2}$$

Така, ако монетата е правилна и хвърляме честно, би трябвало броят на езитата разделен на броя на опитите да клони към половина. Ако монетата не е правилна, граничната вероятност ще се окаже друго число.

2.2 Аксиоматика

Теория на вероятностите става строга математическа теория едва след въвеждането в 1939 г. от А.Н.Колмогоров на следната аксиоматика, основана на теория на мярката (теория на интеграла).

2.2.1 Алгебра на събитията

Елементарно събитие е първично понятие – нещо като точка в геометрията.

Множеството от всички елементарни събития наричаме "достоверно събитие" и означаваме с Ω. Празното множество бележим с $\emptyset$ и наричаме "невъзможно събитие". Всички събития са подмножества на Ω и с тях могат да се правят обичайните в теория на множествата действия. В теория на вероятностите събитието има смисъла на логическото твърдение *сбъднало се е някое от елементарните събития в* A. Със събитията могат да се правят обичайните за множествата действия:*допълнение, обединение, сечение*, които обаче носят други имена.

Допълнението $\Omega \setminus A$ на множеството A в Ω означаваме с $\bar{A}$ и наричаме допълнително събитие (или отрицание) на събитието A.

Сечението на множествата A, B означаваме с $A \bigcap B$ и казваме, че са се сбъднали съвместно събитията A и B.

Обединението на множествата A, B означаваме с $A \bigcup B$ и казваме, че се е сбъднало поне едно от събитията A и B. За краткост това се произнася сбъднало се е A или B.

Когато $A \subset B$ казваме, че събитието A "влече" събитието B.

Операциите със събития удовлетворяват обичайните свойства на операциите с множества. Те лесно се разпространяват и върху безкраен брой събития. Изпълнени са и т.н. закони на де Морган:

$$\overline{\bigcup_k A_k} = \bigcap_k \overline{A_k}, \qquad \overline{\bigcap_k A_k} = \bigcup_k \overline{A_k} \tag{2.3}$$

За удобство са въведени и някои производни определения и операции:

- означаваме с $AB = A \bigcap B$;
- събитията A и B наричаме несъвместими, ако $AB = \emptyset$;
- за несъвместими събития вместо $A \bigcup B$ използуваме знака събиране - пишем $A + B$;
- означаваме с $A \Delta B = \bar{A}B + A\bar{B}$.

За да си осигурим възможността да правим всичките тези операции ще поискаме множеството от събития да го допуска.

Определение 2.1 *Семейство $\mathfrak{A}$ от подмножества на Ω се нарича булова алгебра, ако удовлетворява следните три условия:*

1. $\Omega \in \mathfrak{A}$;
2. *ако* $A \in \mathfrak{A}$, *то* $\bar{A} \in \mathfrak{A}$;
3. *ако* $A, B \in \mathfrak{A}$, *то* $A \bigcup B \in \mathfrak{A}$.

Веднага се вижда от 2.3, че буловата алгебра от множества е затворена и относно операциите $\bigcap, \Delta, +$. Тя обаче не е длъжна да бъде затворена относно операции с безкраен брой множества.

Определение 2.2 *Булова алгебра $\mathfrak{A}$, която е затворена относно изброимите операции обединение и сечение, се нарича булова σ-алгебра – ако $A_k \in \mathfrak{A} (k = 1, 2, \dots)$, то $\bigcup_k A_k, \bigcap_k A_k \in \mathfrak{A}$.*

Определение 2.3 *Двойката $(\Omega, \mathfrak{A})$, където $\mathfrak{A}$ е булова σ-алгебра, се нарича измеримо пространство. Елементите на $\mathfrak{A}$ наричат случайни събития.*

Сега вече сме в състояние да въведем и граница на събития и така, че тя да се окаже събитие. Означаваме:

$$A^* = \limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k, \qquad A_* = \liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k \tag{2.4}$$

Интерпретацията на така определените гранични събития е следната:

- A_* - се състои от тези елементарни събития, които влекат безкраен брой елементи A_n;
- A^* - се състои от тези елементарни събития, които влекат всички елементи A_n от дадено място нататък;

σ-алгебрите притежават някои универсални свойства. Например, сечение на произволен брой σ-алгебри е σ-алгебра. Това ни дава възможност да определим минималната σ-алгебра съдържаща семейството множества $\mathfrak{F}$ като сечение на всички σ-алгебри, съдържащи семейството $\mathfrak{F}$. Ще означаваме тази σ-алгебра $\sigma(\mathfrak{F})$.

2.2.2 Вероятностно пространство

Определение 2.4 *Реалната функция* $\mathbf{P}$ *определена върху елементите на булевата* σ*-алгебра* $\mathfrak{A}$ *се нарича вероятност, ако удовлетворява условията:*

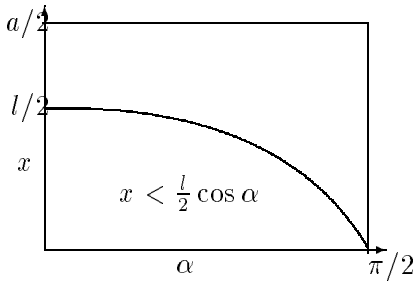
1. *неотрицателност:* $\mathbf{P}(A) \geq 0, \forall A \in \mathfrak{A}$;
2. *нормираност:* $\mathbf{P}(\Omega) = 1$;
3. *адитивност:* $\mathbf{P}(A_1 + A_2 + \ldots) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \ldots$;

Определение 2.5 *Тройката* $(\Omega, \mathfrak{A}, \mathbf{P})$ *наричаме вероятностно пространство.*

От аксиомите 2.4 лесно следват следните свойства на случайните събития.

- $\mathbf{P}(\emptyset) = 0$;
- $\mathbf{P}(\bar{A}) = 1 - \mathbf{P}(A)$;
- *непрекъснатост в* $\emptyset$. Ако $A_i, i = 1, 2, \ldots$ е намаляваща редица от събития, т.е. $A_{i+1} \subset A_i$ и $\bigcap_i A_i = \emptyset$, то $lim_i \mathbf{P}(A_i) = 0$.

Да се върнем към примерите. Във пример 2.1 Ω се състои от 6 елемента, $\mathfrak{A}$ е множеството от всички подмножества на това крайно множество. Вероятността се определя просто – всички елементарни събития са равновероятни.



Фигура 2.2: Иглата на Бюфон

Значително по сложна е ситуацията при примера 2.2. Тук ролята на Ω се поема от множеството от всички точки (α, x) в правоъгълника $0 \leq \alpha \leq \pi/2$ и $0 \leq x \leq a/2$. σ-алгебрата $\mathfrak{A}$ се състои от измеримите по Лебег подмножества на този правоъгълник, т.е. тези на които можем да мерим лице или площ. Вероятността е относителната площ, заемана от тях в правоъгълника. Елементарните събития в това пространство притежават нулева вероятност.

Тема 3

Независимост

В тази лекция си поставяме следните цели:

- да определим понятието условна вероятност;
- да определим понятието независимост;
- да дадем примери и контрапримери.

3.1 Условна вероятност и независимост

Независимостта е най-фундаменталното понятие на теорията на вероятностите. Макар че, тя е някакъв еквивалент на декартовото произведение на множества, или на правото произведение на алгебри, т.е. в математически смисъл едва ли привнася нещо ново, независимостта в действителност е основата на тази теория. Това е понятието, което прави теорията незаменима, когато има нужда от математическо моделиране на явления с непредсказуем изход.

Независимостта, като строго понятие от математиката, се оказва неимоверно близка до нормалните, езикови или човешки представи за същото — кога едно събитие оказва (или не) някакво влияние върху възможността друго събитие да настъпи.

Като всяко математическо понятие независимостта има и редица недостатъци. Основният е, навярно, стриктността — изискванията са толкова строги, че стават непроверяеми. С други думи, когато ние казваме,че две величини или събития са независими, ние влагаме в това твърдение много повече вяра, от колкото бихме могли (със средствата на математиката) да проверим.

Нека $B \in \mathfrak{A}$ и $\mathbf{P}(B) > 0$.

Определение 3.1 *За всяко събитие $A \in \mathfrak{A}$ ще наречем числото*

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

условна вероятност на събитието A при условие събитието B.

Лесно е да се види, че ако зафиксираме условието B, условната вероятност притежава всичките свойства на безусловната. Събитието B (и всички съдържащи го събития) притежава вероятност 1. Събитията влечащи B повишават своята вероятност, а несъвместимите с B стават "невъзможни". Така върху същата σ-алгебра е породена нова вероятност отразяваща факта за настъпването на събитието B.

Регистрацията на настъпване на дадено случайно събитие променя състоянието на вероятностното пространство — вече е невъзможно настъпването на елементарни събития извън (не влечащи) това събитие. Тази ситуация е отразена в изменението на вероятността на другите събития — условната им вероятност не винаги е същата като безусловната.

Теорема 3.1 *(Формула за умножение на вероятности) Вярна е следната формула:*

$$\mathbf{P}(A_1A_2\dots A_n) = \mathbf{P}(A_1)\mathbf{P}(A_2|A_1)\mathbf{P}(A_3|A_1A_2)\dots\mathbf{P}(A_n|A_1A_2\dots A_{n-1}). \tag{3.1}$$

Доказателство: Ще докажем твърдението по индукция. За $n = 2$ то е очевидно следствие от определение 3.1 на условна вероятност. Нека то е изпълнено за някое n. Тогава да приложим същото определение за събитията $B = A_1A_2\dots A_n$ и A_{n+1}:

$$\mathbf{P}(A_1A_2\dots A_{n+1}) = \mathbf{P}(BA_{n+1}) = \mathbf{P}(B)\mathbf{P}(A_{n+1}|B)$$
$$= \mathbf{P}(A_1)\mathbf{P}(A_2|A_1)\mathbf{P}(A_3|A_1A_2)\dots\mathbf{P}(A_n|A_1A_2\dots A_{n-1}).$$

■

3.1.1 Формула на пълната вероятност

Определение 3.2 *Казваме, че събитията $(H_1, H_2, \dots, H_n)$ образуват пълна група, когато $H_iH_j = \emptyset, \forall i \neq j$ и $H_1 + H_2 + \dots + H_n = \Omega$. Прието е събитията от пълната група да се наричат хипотези.*

Нека е зададена пълната група събития $(H_1, H_2, \dots, H_n)$. Изпълнена е следната формула за пълната вероятност:

$$\mathbf{P}(A) = \sum_{i=1}^{n} \mathbf{P}(A|H_i)\mathbf{P}(H_i). \tag{3.2}$$

Доказателство: Следва лесно от очевидното равенство:

$$A = AH_1 + AH_2 + \dots + AH_n$$

и определение 3.1 на условна вероятност . ■

3.1.2 Формула на Бейс

Изпълнена е следната формула на Бейс:

$$\mathbf{P}(H_k|A) = \frac{\mathbf{P}(A|H_k)\mathbf{P}(H_k)}{\sum_{i=1}^{n}\mathbf{P}(A|H_i)\mathbf{P}(H_i)}. \tag{3.3}$$

Доказателство: Следва от определение 3.1 на условна вероятност. ■

3.2 Независимост

В някои, редки случаи, обаче настъпването на някои събития не оказва такова влияние върху шансовете на други събития.

Тук ще дадем формално определение на понятието независимост. Ще се убедим, че в тази си формулировка, то изключва някаква причинно следствена връзка между явленията, които наричаме независими.

Определение 3.3 *Казваме че събитията A, B са независими, ако* $\mathbf{P}(AB) = \mathbf{P}(A)\mathbf{P}(B)$. *Ще бележим независимите събития* $A \perp B$.

От това определение веднага следва, че условната вероятност на всяко от двете събития е равна на неговата безусловна вероятност. С други думи, вероятността да настъпи събитието A не зависи от това, дали е настъпило или не, събитието B.

Определение 3.4 *Казваме че събитията* $\{A_k, \quad k = 1, 2, \ldots, n\}$ *са независими в съвкупност, ако вероятността на всяко от тях не зависи от това дали се е случила някоя комбинация от останалите събития.*

От това определение следва, че когато събитията са независими в съвкупност, имаме:

$$\mathbf{P}(A_1A_2\ldots A_n) = \mathbf{P}(A_1)\mathbf{P}(A_2)\ldots\mathbf{P}(A_n) \tag{3.4}$$

Това условие, обаче, не е достатъчно за да бъдат събитията незвисими две по две, както и обратното.

Пример 3.1 *Да разгледаме следното вероятностното пространство състоящо се от 4 равновероятни елементарни събития:* $\{\omega_i, i = 1, 2, 3, 4\}$. *Тогава събитията* $A = \{\omega_1, \omega_2\}, B = \{\omega_1, \omega_3\}, C = \{\omega_1, \omega_4\}$ *са независими две по две, но не са независими в съвкупност.*

Наистина, $\mathbf{P}(A) = \mathbf{P}(B) = \mathbf{P}(C) = \frac{1}{2}$, $\mathbf{P}(AB) = \mathbf{P}(BC) = \mathbf{P}(AC) = \frac{1}{4}$, но $\mathbf{P}(ABC) = \mathbf{P}(\{\omega_1\}) = \frac{1}{4} \neq \frac{1}{8}$.

Тема 4

Случайни величини

В тази лекция си поставяме следните цели:

- да определим случайна величина (сл.в.);
- да определим разпределение на сл.в.;
- да определим математическо очакване и изведем основните му свойства;
- да дадем примери за сл.в..

Случайните събития представляват най-простия пример за модел на наблюдение със случаен (неопределен отнапред) изход. Често се налага да на практика наблюденията да бъдат измервания – резултатът от експеримента да се записва с число. Модел на такива експерименти са случайните величини.

4.1 Прости случайни величини

Сл.в. са числови функции определени върху множеството от елементарни събития Ω, но тяхното определение силно зависи от това кои са случайните събития Ω.

Определение 4.1 *Нека е зададена пълната група събития* $(H_1, H_2, \ldots, H_n)$. *Ще казваме, че е определена проста сл.в., ако* $\xi(\omega) = x_i, \forall \omega \in H_i, i = 1, 2, \ldots, n$.

Пример 4.1 *Хвърляне на два зара. Нека разгледаме вероятностно пространство състоящо се от 36 равновероятни елементарни събития. Да ги означим с* $w_{i,j}, i, j =$

$1, 2, \ldots, 6$). *Да определим на това вероятностно пространство 2 сл.в.* $\xi(w_{i,j}) = i, \eta(w_{i,j}) = j$).

Намерете пълните групи на двете сл.в.

Теорема 4.1 *Линейна комбинация, произведение и функция на прости сл.в. е проста сл.в.*

Доказателство: Нека пълната група събития $(H_1, H_2, \ldots, H_n)$ съответствува на сл.в. ξ, пълната група събития $(G_1, G_2, \ldots, G_m)$ на сл.в. η, а техните стойности са съответно $\{x_1, x_2, \ldots x_n\}$ и. $\{y_1, y_2, \ldots y_m\}$.

Първо ще докажем, че събитията $\{H_iG_j,\ i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, m\}$ образуват пълна група.

$$H_iG_j \bigcap H_kG_l = H_iH_k \bigcap G_jG_l = \emptyset, \text{ когато } i \neq k \text{ или } j \neq l.$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m} H_i \bigcap G_j = \sum_{i=1}^{n} H_i \bigcap \sum_{j=1}^{m} G_j = \sum_{i=1}^{n} H_i\Omega = \sum_{i=1}^{n} H_i = \Omega.$$

Тогава сл.в. $\alpha\xi + \beta\eta$ и $\xi\eta$ ще приемат стойности $\alpha x_i + \beta y_j$ и x_iy_j за всяко елементарно събитие от фиксирано събитие от пълната група от събития. Функцията $f(\xi, \eta)$ съответно ще приема стойности $f(x_i, y_j)$ върху същото събитие. ■

Определение 4.2 *Казваме, че простите сл.в.* ξ *и* η *са независими, ако е независимо всяко от събитията на едната пълна група с всяко от събитията на другата пълна група. Бележим това с* $\xi \perp \eta$.

Покажете, че в пример 4.1 $\xi \perp \eta$.

Определение 4.3 *Ще казваме, че функцията* $\xi(\omega)$, *определена на* Ω *със стойности в* R, *е сл.в., ако* $\forall x \in R^1$ *множеството* $\{\omega : \xi(\omega) < x\} \in \mathfrak{A}$.

Теорема 4.2 *Линейна комбинация, произведение и измерима функция на сл.в. е сл.в.*

Доказателството на тази теорема изисква известна математическа подготовка.

4.2 Функция на разпределение и плътност

Определение 4.4 *Ще наричаме функция на разпределение на сл.в. ξ функцията $F(x) = \mathbf{P}(w : \xi(w) < x)$.*

Функцията $F(x)$ е монотонно ненамаляваща и непрекъсната от ляво. Освен това $F(-\infty) = 0, \quad F(\infty) = 1$. В термини на разпределението си сл.в. се класифицират лесно.

Определение 4.5 *Случайната величина, която приема само стойностите $x_1, x_2, x_3, \ldots$ с вероятности съответно $p_1, p_2, p_3, \ldots$ се нарича дискретна.*

Естествено $\sum p_i = 1$ и $p_i \geq 0$. Тогава функцията ѝ на разпределение $F(X)$ има само скокове в точките x_i, навсякъде другаде е константа. В точката x_i скокът ѝ е равен точно на числото p_i.

Определение 4.6 *Ако сл.в. е такава, че за всяко x съществува производна на функцията на разпределение $f(x) = F'(x)$, то ще я наричаме непрекъсната. Производната наричаме плътност.*

Естествено е че за да бъде една функция плътност на случайна величина, тя трябва да отговаря на две изисквания:

- неотрицателност $f(x) \geq 0$ и
- нормираност - $\int_{-\infty}^{\infty} f(x)dx = 1$.

Функцията на разпределение на непрекъсната сл.в. се представя като интеграл от плътността:

$$F(x) = \int_{-\infty}^{x} f(y)dy,$$

и естествено е непрекъсната функция, т.е. няма никакви скокове. Всъщност за всяка монотонно ненамаляваща, непрекъсната отляво функция $F(x), \quad F(-\infty) = 0, \quad F(\infty) = 1$ съшествува сл.в. ξ, такава, че $F(x) = \mathbf{P}(w : \xi(w) < x)$. Това означава, че разпределенията на сл.в. могат да бъдат и по-сложни.

Пример 4.2 *Нека разгледаме двете ф.р. – непрекъсната $F_1(x)$ и дискретна $F_2(x)$. Тогава функцията (смес) $F(x) = \alpha F_1(x) + (1 - \alpha)F_2(x)$ е ф.р. за всяко $0 \leq \alpha \leq 1$ е също ф.р. на някаква сл.в.*

4.3 Математическо очакване.

Определение 4.7 *Математическо очакване на простата сл.в. ξ приемаща стойности $x_1, x_2, \ldots, x_n$ върху събитията от пълната група определяме като $\sum_{k=1}^{n} x_k \mathbf{P}(H_k)$.*

От това определение се вижда веднага, че числото $\mathbf{E}\,\xi$ зависи само от стойностите и вероятностите, с които те се приемат, но не зависи от това за кои точно елементарни събития и каква пълна група това става. Т.е. то не зависи от това в какво вероятностно пространство е реализирана сл.в. Наистина, тъй като простата сл.в. е дискретна, то математическото й очакване се пресмята като сумата: $\mathbf{E}\,\xi = \sum_i x_i p_i (p_i = \mathbf{P}(H_i))$.

Теорема 4.3 *За прости сл.в. математическото очакване притежава следните свойства:*

1. *монотонност - Ако $\xi < \eta$, то $\mathbf{E}\,\xi < \mathbf{E}\,\eta$;*
2. *линейност $\mathbf{E}\,(\alpha\xi + \beta\eta) = \alpha\mathbf{E}\,\xi + \beta\mathbf{E}\,\eta$;*
3. *мултипликативност - Ако $\xi \perp \eta$, то $\mathbf{E}\,\xi\eta = \mathbf{E}\,\xi\mathbf{E}\,\eta$;*

Тези свойства на математическото очакване и особенно неговата монотонност позволяват то лесно да се разпространи за произволни неотрицателни сл.в. Може обаче да се окаже, че то е безкрайно.

Математическото очакване на непрекъснатата сл.в. ξ се пресмята като интеграла:

$$\mathbf{E}\,\xi = \int x dF(x) = \int x f(x) dx,$$

а това на дискретна като сумата

$$\mathbf{E}\,\xi = \int x dF(x) = \sum_i x_i p_i,$$

когато това е възможно, т.е. съответния интеграл (или сума) е абсолютно сходящ: $\int |x| dF(x) < \infty$. Възможно е, да се даде и абстрактна дефиниция на интеграла на Стилтес: $\int g(x) dF(x)$, която обединява горните две формули.

Тема 5

Числови характеристики

Всяка от следните характеристики може да се пресмята както за теоретичните разпределения, представени от своята функция на разпределение $F(x)$, така и за извадъчните, представени от $F_n(x)$.

5.1 Локация

М.о. е най - важната характеристика за положението на стойностите на сл.в. върху числовата ос. За съжаление не за всички сл.в. тя е определена. Съответната извадъчна характеристика (тя е определена винаги) е средната на извадката:

$$\bar{x} = \int x dF_n x = \frac{1}{n} \sum_i x_i. \tag{5.1}$$

Определение 5.1 *Медиана се определя като решение на уравнението:* $F(\mu) = \frac{1}{2}$. *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 математическото очакване.

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

Определение 5.2 *Квантил с ниво α на дадено разпределение F се определя като решение на уравнението:*

$$F(q_\alpha) = \alpha.$$

Така медианата $\mu = q_{1/2}$.

Определение 5.3 *Мода е най - вероятното число за дискретни сл.в., а за непрекъснати — максимума на плътността.*

За симетрични разпределения, очевидно трите характерстики: мода, медиана и м.о. съвпадат.

5.2 Мащаб

Определение 5.4 *Дисперсия на сл.в. ξ се определя като числото* $\mathbf{D}\,\xi = \mathbf{E}\,(\xi - \mathbf{E}\,\xi)^2$. *Може да се окаже и безкрайна.*

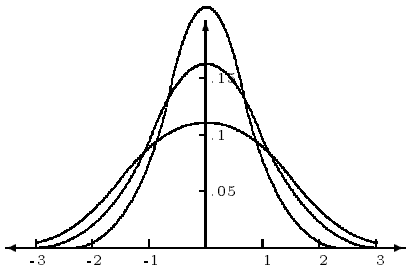
Дисперсията е най - важната характеристика на разсейване на стойностите на сл.в. За дискретни и непрекъснати разпределения тя се пресмята по формулите:

$$\mathbf{D}\,\xi = \int (x - \mathbf{E}\,\xi)^2 f(x)dx, \quad \mathbf{D}\,\xi = \sum_i (x_i - \mathbf{E}\,\xi)^2 p_i. \tag{5.2}$$

Фактически вместо дисперсията, както в числовите, така и в аналитичните сметки, се използува стандартната грешка или *стандартното отклонение* . Това е:

$$\sigma(\xi) = \sqrt{\mathbf{D}\,\xi} = \sqrt{\mathbf{E}\,(\xi - \mathbf{E}\,\xi)^2}. \tag{5.3}$$

Тази характеристика се мери в същите физически единици, като ξ и може да бъде съответно интерпретирана.



Фигура 5.1: $N(0,.75), N(0,1), N(0,1.5)$

Тук са показани плътности от нормалното семейство с различни стандартни отклонения. Колкото по - малка е дисперсията или стандартното отклонение, толкова по - сгъстени са стойностите и по - вероятни са те в центъра на разпределението. За това, когато искаме да се отървем от размерността, например за да сравним разпределенията на две различни сл.в., прилагаме т.н. *центриране и нормиране*. Вместо величината ξ разглеждаме центрираната и нормирана величина

$$\tilde{\xi} = \frac{\xi - \mathbf{E}\,\xi}{\sigma(\xi)}. \tag{5.4}$$

Когато дисперсията е безкрайна за "определяне" на мащаба се използува т.н.

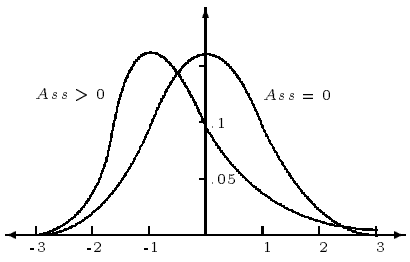
Определение 5.5 *интерквартилен размах* $q_{3/4} - q_{1/4}$.

5.3 Форма

Следните две характеристики на разпределенията не зависят от мерните единици, с които са отчитани съответните сл.в., както и от условните начала на скалите. С други думи, те са безразмерни. Те отразяват различията във формата на разпределенията, но не зависят от мащаба и локацията.

Определение 5.6 *Ще наричаме асиметрия на* ξ *числото (когато съществува):*

$$Ass(\xi) = \frac{\mathbf{E}\,(\xi - \mathbf{E}\,\xi)^3}{\sigma^3(\xi)} = \mathbf{E}\,\tilde{\xi}^3. \tag{5.5}$$



Фигура 5.2: Положителна асиметрия

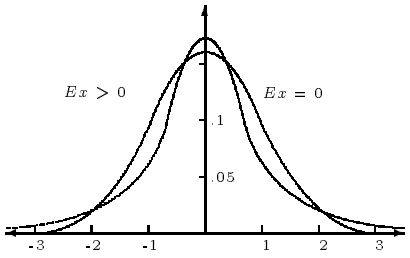
На тази фигура е дадено сравнение на положително асиметрична плътност с плътността на стандартния нормален закон, която е симетрична и има асиметрия 0. Положителната асиметрия се характеризира с "по - тежка" дясна опашка на разпределението.

При асиметричните разпределения се променят обикновено и взаимните положения на модата, медианата и математическото очакване. За разпределния с положителна асиметрия те се нареждат в посочения ред, а за тези с отрицателна — в обратния. Това правило, разбира се, е верно само за унимодални разпределения с проста аналитична форма на плътността.

Определение 5.7 *Ще наричаме ексцес на* ξ *числото (когато съществува):*

$$Ex(\xi) = \frac{\mathbf{E}\,(\xi - \mathbf{E}\,\xi)^4}{\sigma^4(\xi)} - 3 = \mathbf{E}\,\tilde{\xi}^4 - 3. \tag{5.6}$$

Тук е представено разпределение с положителен ексцесс. То има по - дълги и тежки опашки от нормалното (с ексцесс 0). Разпределенията с отрицателен ексцесс може изобщо да нямат опашки — например, такова е равномерното в краен интервал. Изобщо казано, двата параметъра асиметрия и ексцес дават достатъчно пълна картина за формата на разпределението, само когато то е унимодално и гладко. Всъщност такива разпределения обикновено принадлежат на семейство описвано с няколко параметъра.



Фигура 5.3: Положителен ексцес

Тема 6

Дискретни разпределения

В тази лекция си поставяме за цел да обобщим и разширим понятията си за:

- целочислена сл.в. и нейното разпределение;
- да въведем някои най-срещани разпределения;
- ще покажем как се използуват някои от средствата на анализа за облекчаване на пресмятанията на разпределенията и техните количествени характеристики - моментите.

6.1 Целочислени сл.в.и пораждащи функции

Определение 6.1 *Сл.в. приемаща за стойности натуралните числа наричаме целочислена.*

Целочислените сл.в. са особено удобни за моделиране на реални явления като брой успехи или други бройки.

За пресмятане на моментите на целочислени сл.в. особено удобни са така наречените пораждащи функции.

Определение 6.2 *Пораждащата функция на целочислена сл.в. ξ се задава с формулата:*

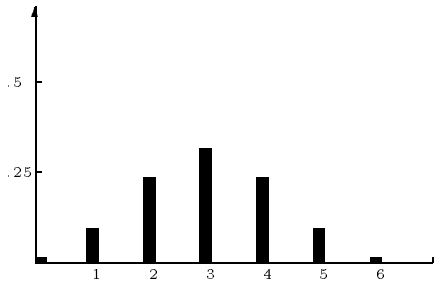
$$p(s) = \mathbf{E}\, s^{\xi} \tag{6.1}$$

Пораждащата функция е удобна защото съществува винаги (при достатъчно малко s, например, когато $s \leq 1$). Тя притежава следните свойства:

- $p(1) = 1$;
- $p(0) = \mathbf{P}(\xi = 0)$;
- $p'(1) = \mathbf{E}\,\xi$, когато съществува;
- $p''(1) = \mathbf{E}\,\xi(\xi - 1) = \mathbf{E}\,\xi^2 - \mathbf{E}\,\xi$, когато съществува.
- Когато $\xi \perp \eta$, $p_{\xi+\eta}(s) = p_\xi(s)p_\eta(s)$.

6.2 Схема на Бернули

Определение 6.3 *Редица от независими еднакво разпределени случайни величини* $\{\xi_i,\ i = 1, 2, ...\}$, *всяка от които приема две стойности: 1 и 0 с вероятности (съответно)* p *и* $q = 1 - p$, *наричаме схема на Бернули.*



Фигура 6.1: Биномно разпределение, $p = .5$.

Да разгледаме сумата η_n на n сл.в. от схемата на Бернули. Това е целочислена случайна величина, приемаща
стойности от 0 до n. Ние я интерпретираме като *Брой успехи от* n *опита с постоянна вероятност* p *за успех във всеки опит.* Разпределението на тази сл.в. наричаме *биномно*. Вероятността тази сл.в. да приеме стоиност k наричаме биномна и означаваме с $b(n, k, p)$.

Теорема 6.1 *Биномните вероятности се пресмятат по формулата:*

$$b(n, k, p) = \binom{n}{k} p^k q^{n-k} \tag{6.2}$$

Доказателство: Първо да пресметнем вероятността на събитието $W_{\epsilon_1,\epsilon_2,...,\epsilon_n} = \bigcap_{i=1}^n \{\xi_i = \epsilon_i\}$, където $\epsilon_j \in \{0, 1\}, \quad j = 1, 2, \ldots, n$. Тъй като сл.в. са независими и $P(\xi = 1) = p$, получаваме

$$\mathbf{P}(W_{\epsilon_1,\epsilon_2,...,\epsilon_n}) = p^{\sum_{i=1}^n \epsilon_i} q^{n - \sum_{i=1}^n \epsilon_i} \tag{6.3}$$

Ако означим $\eta_n = \sum_{i=1}^n \xi_i$ и $k = \sum_{i=1}^n \epsilon_i$, ще получим

$$P(\eta_n = k) = \sum_{\epsilon_1+\epsilon_2+...+\epsilon_n=k} p^k q^{n-k} = p^k q^{n-k} \sum_{\epsilon_1+\epsilon_2+...+\epsilon_n=k} 1.$$

Но от тук следва търсената формула. ■

Моментите на биномното разпределение се пресмятат лесно:

$$\mathbf{E}\,\eta_n = \sum \mathbf{E}\,\xi_i = n\mathbf{E}\,\xi_1 = n(1.p + 0.q) = np$$
$$\mathbf{D}\,\eta_n = \sum D\xi_i = n\mathbf{D}\,\xi_1 = n(\mathbf{E}\,\xi_1^2 - (\mathbf{E}\,\xi_1)^2) = n(p - p^2) = npq$$

Пораждащата функция на биномното разпределение се пресмята лесно, защото биномната сл.в. η е сума на еднакво разпределени независими сл.в.

$$\mathbf{E}\,s^{\eta} = \mathbf{E}\prod_{i=1}^{n} s^{\xi_i} = (\mathbf{E}\,s^{\xi_1})^n = (ps + q)^n.$$

6.2.1 Геометрично разпределение

Нека разгледаме в ситуацията на независими опити (схема на Бернули) сл.в. ξ — брой опити до достигане на успех. Във всеки отделен опит нека вероятността за неуспех да означим с p и нека опитите да са независими.

Определение 6.4 *Казваме, че целочислената сл.в. ξ има геометрично разпределение, ако:*

$$\mathbf{P}(\xi = m) = p^m q, \qquad m = 0, 1, 2, \ldots. \tag{6.4}$$

Математическото очакване и дисперсията на това разпределение се пресмятат лесно:

$$\mathbf{E}\,\xi = q\sum_{k=0}^{\infty} kp^k = qp\sum_{k=0}^{\infty} kp^{k-1} = qp\frac{d}{dp}(\frac{1}{1-p}) = \frac{p}{q},$$
$$\mathbf{D}\,\xi = \mathbf{E}\,\xi(\xi-1) + \mathbf{E}\,\xi - (\mathbf{E}\,\xi)^2 = q\sum_{k=0}^{\infty} k(k-1)p^k + \frac{p}{q} - (\frac{p}{q})^2 =$$
$$qp^2\frac{d^2}{dp^2}(\frac{1}{1-p}) + \frac{p}{q} - (\frac{p}{q})^2 = (\frac{p}{q})^2 + \frac{p}{q}.$$

6.2.2 Хипергеометрично разпределение

Да разгледаме една задача от статистическия качествен контрол. Нека е дадена партида съдържаща N изделия, от които M са дефектни. Правим случайна извадка от $n < N$ изделия. Пита се каква е вероятността точно m от тях да са дефектни.

Оказва се, че разпределението на сл.в. брой дефектни е следното:

Определение 6.5 *Казваме, че целочислената сл.в. ξ има хипергеометрично разпределение, ако:*

$$\mathbf{P}(\xi = m) = \frac{\binom{n}{m}\binom{N-n}{M-m}}{\binom{N}{n}}, \qquad m = 0, 1, \ldots, M. \tag{6.5}$$

Тази формула се извежда лесно. Броят на всички възможни извадки е очевидно $\binom{N}{n}$ (смятаме ги за равновероятни). "Благоприятните", тези които съдържат точно m дефектни детайла, могат да се получат чрез комбиниране на извадка от m дефектни и извадка от $n-m$ изправни. Така този брой става $\binom{M}{m}\binom{N-M}{n-m}$.

Математическото очакване и дисперсията на това разпределение също се пресмятат лесно:

$$\mathbf{E}\,\xi = np, \quad p = \frac{M}{N},$$
$$\mathbf{D}\,\xi = npq\frac{N-1}{N-n}.$$

От тези формули се вижда, че това разпределение клони към биномното при голям брой N на детайлите в партидата.

6.3 Разпределение на Поасон

Поасоновото разпределение се определя лесно кото граница на биномни разпределения, когато $n \to \infty$ така че $np \to \lambda > 0$. Сл.в. може да приема всякакви целочислени стойности:

$$\mathbf{P}(\xi = k) = e^{-\lambda}\frac{\lambda^k}{k!}. \tag{6.6}$$

То е особено подходящо за моделиране на броя на случайни редки събития – брой частици на единица обем, брой радиоактивни разпадания за единица време и т.н. Средното и дисперсията му съвпадат: $\mathbf{E}\,\xi = \mathbf{D}\,\xi = \lambda$. Това най-лесно се вижда от пораждащата функция на поасоновото разпределение, която се пресмята директно:

$$\mathbf{E}\,s^\eta = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda s)^k}{k!} = e^{\lambda(s-1)}.$$

Тук ще разгледаме едно много полезно и старо приближение на биномната вероятност при малки k.

Теорема 6.2 *(Теорема на Поасон) Ако в схемата на Бернули $np_n \to \lambda$, то*

$$b(n,k,p_n) \longrightarrow \frac{\lambda^k}{k!}e^{-\lambda}.$$

Доказателство: Да означим $\lambda = np$. Можем да запишем биномната вероятност във формата:

$$b(n,k,p) = \frac{n(n-1)\ldots(n-k+1)}{k!}p^k(1-p)^{n-k} = \frac{\lambda^k}{k!}e^{-\lambda}\epsilon(k,n,\lambda),$$

където

$$\epsilon(k,n,\lambda) = \prod_{i=0}^{k-1}(1+\frac{i}{n}) \quad (1+\frac{\lambda}{n})^k \quad e^{\lambda}(1-\frac{\lambda}{n})^n. \tag{6.7}$$

Всеки от трите съмножителя на дясната страна клони към 1 при фиксирано k и $np_n \to \lambda$. ■

Още по-лесно се доказва тази теорема с помощта на пораждащи функции. Наистина, достатъчно е да покажем, че

$$(ps+q)^n = (1+\frac{(s-1)\lambda}{n})^n \to e^{\lambda(s-1)}.$$

Тема 7

Схема на Бернули

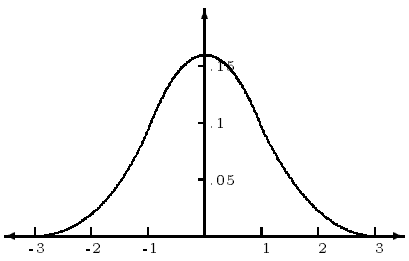
В тази лекция си поставяме следните цели:

- да разгледаме най - простата непрекъсната сл.в.;
- да покажем връзката между непрекъснати и дискретни разпределения;
- на примера на най - простите статистически задачи да илюстрираме начина, по който се строят статистическите изводи.

7.1 Теореми на Муавър-Лаплас

Определение 7.1 *Казваме, че сл.в. с непрекъснато разпределение е нормално разпределена $N(\mu, \sigma)$, ако нейната плътност има вида:*

$$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}. \tag{7.1}$$



Фигура 7.1: Плътност $\phi(x)$

Математическото очакване на това разпределение е μ, а дисперсията му е σ^2. *Стандартно* нормално разпределение се нарича разпределението $N(0, 1)$, неговата плътност означаваме с $\phi(x)$.

Нормалното разпределение има голямо значение в теория на вероятностите и математическата статистика, което се дължи на твърдението, известно като Централна Гранична Теорема. То гласи, че разпределението на сума от голям брой независими, еднакво разпределени случайни величини клони към нормално разпределение.

Нека устремим към безкрайност броят на опитите в схемата на Бернули. Да означим

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

и поискаме това число да остане "почти постоянно" при $n \to \infty$. Смисълът му е ясен - това е центрираната и нормирана стойност на сл.в. брой на успехи. Ясно е, че тогава (при фиксирана вероятност за успех p) също и $k \to \infty$.

Теорема 7.1

$$\sqrt{npq}b(n,k,p) \longrightarrow \phi(x).$$

Доказателство: За простота ще изпускаме индекса n от означенията в доказателството. Да логаритмуваме биномната вероятност от лявата страна

$$\ln b(n,k_n,p) = \ln n! - \ln k_n! - \ln(n-k_n)! + k_n \ln p + (n-k)\ln q. \tag{7.2}$$

Ще използуваме представянето на Стирлинг на $\ln n!$:

$$\ln n! = n\ln n + \frac{1}{2}\ln(2\pi n) - n + \alpha(n), \tag{7.3}$$

където $\alpha(n) = O(1/n)$. Да означим $m_n = n - k_n$. Тъй като $m_n, k_n \to \infty$, то от (7.2) и (7.3) следва

$$\begin{aligned} &\ln b(n,k,p) - \frac{1}{2}\ln 2\pi \\ &= n\ln n - k\ln k - m\ln m + k\ln p + m\ln q + -\frac{1}{2}\ln\frac{km}{n} + \beta_n = \\ &= -(np+\sigma x)\ln(1+\frac{xq}{\sigma}) - (nq-\sigma x)\ln(1-\frac{xp}{\sigma}) - \frac{1}{2}\ln\frac{km}{n} + +\beta_n, \end{aligned} \tag{7.4}$$

където $\beta_n = \alpha(n) - \alpha(k_n) - \alpha(m_n) = O(1/n)$. $k = np + \sigma x$ и $m = n - k = nq - \sigma x$. Тук ще се отклоним малко да разгледаме дробта:

$$\frac{k(1-k)}{n\sigma^2} = \frac{(np+\sigma x)(nq-\sigma x)}{n\sigma^2} = (1 + \frac{(q-p)x}{\sigma} - \frac{pqx^2}{\sigma^2}).$$

Така лесно ще можем да получим израз за третия логаритъм в (7.4) чрез σ.

$$\frac{1}{2}\ln\frac{mk}{n} = \frac{1}{2}\ln\sigma + O(\frac{1}{\sigma}).$$

За първите два логаритъма ще използуваме разложението $\ln(1+x) = x - x^2/2 + O(x^3)$. Заместваме, съкращаваме и получаваме окончателно:

$$\begin{aligned} \ln\sigma + \ln b(n,k_n,p) &= \frac{1}{2}\ln 2\pi + \\ -(np+\sigma x)(\frac{xq}{\sigma} - \frac{x^2q^2}{2\sigma^2})) - (nq-\sigma x)(-\frac{xp}{\sigma} - \frac{x^2p^2}{2\sigma^2})) + \gamma_n \\ &= \frac{1}{2}\ln 2\pi - \frac{x^2}{2} + \gamma_n. \end{aligned}$$

където $\gamma_n = O(\sigma^{-1}) = O(n^{-1/2})$. ■

Тази теорема ни дава възможност да пресмятаме лесно конкретни биномни вероятности. При големи стойности на n това са твърде малки числа. За да можем да пресмятаме суми от Биномни вероятности си служим със следната интегрална теорема на Муавър - Лаплас.

Теорема 7.2

$$\mathbf{P}(\frac{k-np}{\sqrt{npq}} < x) = \sum_{k=0}^{[np+x\sqrt{npq}]} b(n,k,p) \longrightarrow \int_{-\infty}^{x} \phi(y)dy.$$

Доказателство: Виж [Янев, Димитров (1990)]. Тя лесно следва и от централната гранична теорема за еднакво разпределени събираеми. ■

7.2 Статистически приложения

В този параграф за първи път ще се запознаем с един статистически извод — твърдение за неизвестното разпределение на стойностите в изучавано множество от обекти въз основа на ограничената информация, получена от една случайна крайна извадка.

Нека направим някои упростяващи предположения:

- А. Множеството от изучавани обекти (генералната съвкупност) е много голямо —"почти безкрайно".
- Б. Разполагаме с механизъм позволяващ всеки обект от генералната съвкупност да бъде избран с еднакъв шанс. Можем да прилагаме този механизъм многократно и неговите качества от това няма да се изменят.

За съжаление тези предположения едва ли са достатъчни за една пълна формализация. За това ще упростим още нещата и ще предположим допълнително, че наблюденията могат да се разглеждат като набор от независими сл.в. с еднаква неизвестна плътност.

7.2.1 Доверителен интервал за медиана

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.). Така наредени те се наричат *вариационен ред*.

Теорема 7.3 *За всяко* $i < n/2$

$$\mathbf{P}(\xi_{(i)} \leq \mu \leq \xi_{(n-i+1)}) = 1 - 2(\frac{1}{2})^n \sum_{k=0}^{i-1} \binom{n}{k} \tag{7.5}$$

Доказателство: Имаме равенствата:

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

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 (7.5) стойностите на наблюденията, ние получаваме *доверителен интервал* за неизвестната медиана. Вероятността в дясно се нарича *ниво на доверие* и би трябвало да се избира достатъчно висока за да имат хората някакво доверие в нашите твърдения. Като едно разумно ниво в статистиката е прието нивото на доверие 0.95. При големи стойности на n е затруднително пресмятането на суми от биномни коефициенти. Тогава се използува интегралната теорема на Моавър - Лаплас (теорема 7.2). 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. От друга страна, извадъчната медиана (средното на 50 и 51 наблюдения) е една оценка на теоретичната медиана на разпределението.

7.3 Доверителен интервал за вероятност

Нека си поставим за цел по n наблюдавани стойности да кажем нещо за неизвестната вероятност p на поява на даден признак. Нека с k означим резултата от нашите наблюдения – броят на поява на признака. Ще използуваме интегралната теорема на Муавър – Лаплас (теорема 7.2).

$$\int_{-x}^{x} \phi(y)dy = \mathbf{P}(-x < \frac{k - np}{\sqrt{npq}} < x) = \mathbf{P}(|\frac{k}{n} - p| < x\sqrt{pq/n}).$$

Можем да подберем числото $x = 1.96$, тогава вероятността е 0.95. Тъй като $max(pq) = 1/4$, получаваме:

$$0.95 \leq \mathbf{P}(|\frac{k}{n} - p| < \frac{1.96}{\sqrt{4n}}).$$

Тема 8

Непараметрични методи

Тук ще разгледаме някои съвсем прости непараметрични методи, които не се нуждаят от особено силни предположения и, съответно, не притежават други добри качества освен простотата си.

8.1 Тест на знаците

Тук ще проверим хипотезата H_0, че медианата μ на разпределението е равна на 0 срещу различни контра хипотези. Статистиката Z_n, която се предлага, е броят на наблюденията с положителен знак. Ясно е, че при изпълнена нулева хипотеза нейното разпределение е биномно с вероятност $p = 0.5$. Да разгледаме проверката на тази хипотеза при различните алтернативи.

8.1.1 Насочени алтернативи

Нека алтернативата $H_1 : \mu > 0$. Тогава броят на наблюденията с положителен знак би следвало да нарастне. Следователно, критичната област ще бъде локализирана в дясната част на биномното разпределение:

$$W = \{Z_n : Z_n > i\}, \quad \mathbf{P}(W) = \sum_{k=i}^{n} b(n, k, 0.5).$$

При големи стойности на n се използува интегралната теорема на Моавър - Лаплас. Това ни дава лесна възможност да намерим необходимото i.

8.1.2 Доверителен интервал за медиана

Когато алтернативата е $H_1 : \mu \neq 0$ проверката на H_0 е еквивалентна на проверката, че нулата влиза в доверителния интервал за неизвестната медиана (7.5).

8.2 Независими извадки

При отсъствие на нормалност не можем да използуваме тестовете на Стюдент и Фишер и прибягваме към по-слабите рангови тестове.

8.2.1 Тест на Ман-Уитни или Уилкоксън

Нека са дадени две независими извадки от различни съвкупности $x_1, x_2, ..., x_{n_x}$ и $y_1, y_2, ..., y_{n_y}$ 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

$$U_x = \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \delta_{ij}, \tag{8.1}$$

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

$$U_x + U_y = n_x n_y$$

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

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

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

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

8.3 Тестове за сдвоени наблюдения

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

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

8.3.1 Тест на Стюдент за сдвоени наблюдения

Проверяваме хипотезата, че двете съвкупности са еднакви срещу алтернативата, че едната съвкупност има разпределение с по-малки стойности. Ако имаме увереността, че двете съвкупности са с нормално разпределение можем да приложим теста на Стюдент в следната модификация.

Разглеждаме разликите между съответните наблюдения: $z_i = x_i - y_i$. Ако имаме основание да предположим нормалност, прилагаме обикновения тест за тези разлики.

8.3.2 Тест на знаците

Когато не сме уверени в нормалността на разпределенията, прилагаме теста на знаците. Проверяваме хипотезата, че разликите са с нулева медиана $H_0 : \mu_z = 0$ срещу една или друга алтернатива. Например, че едната медиана е по-голяма от другата $H_1 : \mu_x < \mu_y$.

Тема 9

Трансформация на сл.в.

В тази лекция ще определим многомерни функция на разпределение и плътности. Ще изведем формулата за пресмятане на плътност при трансформация на сл.в. - аналог на смяна на променливите под знака на интеграла.

9.1 Многомерни функции на разпределение

Многомерната функция на разпределение на сл.в. $\vec{\xi} \in R^n$. се определя просто:

$$F(\vec{x}) = F(x_1, x_2, \ldots, x_n) = \mathbf{P}(\bigcap_{i=1}^{n} \{\xi_i < x_i\}) \tag{9.1}$$

Тя притежава следните очевидни свойства:

1. $F(-\infty, x_2, \ldots, x_n) = 0$;
2. нормираност - $F(\infty, \infty, \ldots, \infty) = 1$;
3. монотонност - ако $x_1' < x_1''$, то $F(x_1', x_2, \ldots, x_n) \leq F(x_1'', x_2, \ldots, x_n)$;
4. ако $\xi_1 \perp \xi_2 \perp \ldots \perp \xi_n$, то $F_{\vec{\xi}}(\vec{x}) = \prod_{i=1}^{n} F_{\xi}(x_i)$;
5. маргиналното разпределение на сл.в. ξ_1 се възстановява лесно:

$$\mathbf{P}(\xi_1 < x) = F_{\xi_1}(x) = F(x, \infty, \ldots, \infty).$$

Многомерната плътност на разпределение на сл.в. $\vec{x}$ (когато съществува) се определя просто:

$$f(\vec{x}) = \frac{\partial^n F(x_1, x_2, \ldots, x_n)}{\partial x_1 \partial x_2 \ldots \partial x_n} \tag{9.2}$$

При това функцията на рзпределение се възстановява от плътността:

$$F(\vec{x}) = \int_{\vec{y}<\vec{x}} f(y)dy = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_n} f(y_1, y_2, \ldots, y_n)dy_1dy_2\ldots dy_n.$$

Плътността притежава следните очевидни свойства:

1. $f(\vec{x}) \geq 0$;
2. $\int_{R^n} f(y)dy = 1$;
3. ако сл.в. ξ_i са независими $f_{\vec{\xi}}(\vec{x}) = \prod_{i=1}^{n} f_\xi(x_i)$.
4. маргиналната плътност на сл.в. ξ_1 се възстановява лесно от многомерната плътност:
$$f_{\xi_1}(x) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(x, y_2, \ldots, y_n)dy_2, dy_3\ldots dy_n.$$

9.2 Смяна на променливите

Теорема 9.1 *Нека $U : A \longrightarrow B$, където $A, B \in R^n$ са отворени множества, U взаимнооднозначно съответствие и $V = U^{-1}$. Нека функцията $V(x)$ притежава непрекъснати производни на B. Нека ξ е сл.в. със стойности в A и тя има плътност $f_\xi(x)$. Тогава сл.в. $\eta = U(\xi)$ има плътност $f_\eta(y)$, която се задава по формулата:*

$$f_\eta(x) = |J(V)(x)|f_\xi(V(x)), \tag{9.3}$$

където с $J(V)$ е якобианът на трансформацията V, т.е. детерминантата на матрицата:

$$\left\| \begin{array}{cccc} \frac{\partial V_1}{\partial x_1} & \frac{\partial V_1}{\partial x_2} & \cdots & \frac{\partial V_1}{\partial x_n} \\ \frac{\partial V_2}{\partial x_1} & \frac{\partial V_2}{\partial x_2} & \cdots & \frac{\partial V_2}{\partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial V_n}{\partial x_1} & \frac{\partial V_n}{\partial x_2} & \cdots & \frac{\partial V_n}{\partial x_n} \end{array} \right\|.$$

Тази теорема няма да доказваме — тя е следствие от стандартните теореми на анализа за смяна на променливите.

Пример 9.1 *Многомерно нормално разпределение*

Плътността на стандартното нормално разпределение $N(0,I)$ в R^n има вида:

$$\phi(\vec{x}) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}} e^{-x_i^2/2} = \frac{1}{(2\pi)^{n/2}} e^{-||\vec{x}||^2/2},$$

където $\vec{x} \in R^n$.

Нека сл.в. $\xi \in N(0,I)$. Ще разгледаме линейната трансформация $\eta = A\xi + b$. Тук A е неизродена $n \times n$ матрица, а $b \in R^n$. Тогава плътността на η ще се изчисли по формулата (9.3).

$$f_\eta(x) = |J(V)| f_\xi(V(x)) = \frac{|A^{-1}|}{(2\pi)^{n/2}} e^{-\frac{1}{2}(x-b)'(AA')^{-1}(x-b)}.$$

Като означим матрицата $C = AA'$, получаваме стандартния вид на многомерното нормално разпределение $N(b,C)$ с параметри $\mathbf{E}\,\eta = b$ и $cov(\eta) = C$:

$$\phi(x,b,C) = \frac{1}{|C|^{1/2}(2\pi)^{n/2}} e^{-\frac{1}{2}(x-b)'C^{-1}(x-b)}. \tag{9.4}$$

Да проверим тези равенства за параметрите:

$$\mathbf{E}\,\eta = A\mathbf{E}\,\xi + b = b, \quad cov(\eta) = \mathbf{E}\,(\eta - b)(\eta - b)' = A(\mathbf{E}\,\xi\xi')A' = AA'.$$

9.3 Конволюция на плътности

Ще приложим формулата 9.3 към следната задача:

Теорема 9.2 *Нека са дадени две независими сл.в. ξ и η с положителни плътности на разпределение. Тогава са изпълнени следните формули:*

$$f_{\xi+\eta}(x) = \int_{-\infty}^{\infty} f_\xi(x-y) f_\eta(y) dy \tag{9.5}$$

$$Ако \quad \xi, \eta > 0, mo\ f_{\xi*\eta}(x) = \int_{-\infty}^{\infty} \frac{1}{y} f_\xi(\frac{x}{y}) f_\eta(y) dy \tag{9.6}$$

$$Ако \quad \xi, \eta > 0, mo\ f_{\xi/\eta}(x) = \int_{-\infty}^{\infty} y f_\xi(x*y) f_\eta(y) dy \tag{9.7}$$

Доказателство: Да докажем формула (9.5). Разглеждаме двумерната сл.в. $\{\xi, \eta\}$. Тя има плътност $f(x,y) = f_\xi(x) f_\eta(y)$ защото двете сл.в. са независими. Нека разгледаме сега трансформациите:

$$U = \begin{cases} u = & x+y, \\ v = & y \end{cases} \qquad \text{и} \qquad V = U^{-1} = \begin{cases} x = & u-v, \\ y = & v \end{cases}$$

и приложим формула (9.3). Тъй като якобианът на V е равен на 1, получаваме за двумерната плътност на $U(\{\xi,\eta\})$ формулата:

$$f(u,v) = f_\xi(u-v)f_\eta(v).$$

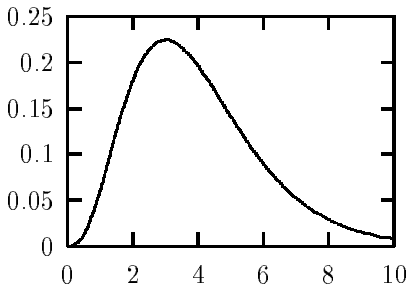
За да получим плътността на първата сл.в. $\xi+\eta$, трябва да интегрираме по втората променлива y. Формули (9.6) и (9.7) се доказват аналогично. ■

9.4 Гама и Бета разпределения

Тук ще се запознаем накратко с две много популярни семейства разпределения.

Определение 9.1 *Наричаме Гама-разпределение , (a,λ) разпределение с плътност:*

$$f(x) = \frac{\lambda^a}{,(a)} x^{a-1} e^{-\lambda x}, \quad x > 0. \tag{9.8}$$



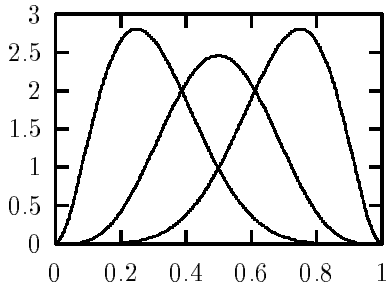
Фигура 9.1: , $(4,1)$

Това семейство е популярно в статистиката, защото е тясно свързано с нормалното. При стойности на a кратни на $1/2$ е известно като Хи-квадрат разпределение и описва разпределението на сума от квадрати на центрирани независими еднакво нормално разпределени сл.в. Параметърът a, който определя формата му, има смисъла на степени на свобода - колкото по-голям е, толкова по-неопределени са стойностите на сл.в. Гама-разпределението има винаги положителна асиметрия, но тя клони към нула при нарастване на a. Вторият параметър λ е мащабен – той не оказва влияние на ексцеса и асиметрията. При $a \to \infty$ центрираното и нормирано Гама-разпределение клони към нормалното.

Определение 9.2 *Наричаме Бета-разпределение разпределение с плътност:*

$$f(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1. \tag{9.9}$$

Тук с $B(a,b)$ сме означили бета-функцията.



Фигура 9.2: B(3,7), B(5,5), B(7,3)

На фиг.9.2 са показани три различни плътности от семейството на бета разпределенията. Вижда се, че те могат да имат различна по знак асиметрия. С нарастването на параметрите a и b, разпределението се изражда (дисперсията му клони към 0). Ако скоростта на нарастване е еднаква и то е правилно нормирано, бета разпределението също клони към нормалното.

Ще приложим формулата (9.3) за да опишем връзката между Гама и Бета рапределенията.

Теорема 9.3 *Нека* $\xi \in \,,(a,\lambda)$ *и* $\eta \in \,,(b,\lambda)$ *са независими Гама - разпределени сл.в. Тогава*

1. *сл.в.* $\zeta = \xi + \eta \in \,,(a+b,\lambda)$;
2. *сл.в.* $\theta = \frac{\xi}{\xi+\eta} \in \mathbf{B}(a,b)$;
3. *сл.в.* $\theta \perp \zeta$.

Доказателство: Разпределението на двумерната сл.в. $\{\xi,\eta\}$ е

$$f(x,y) = \frac{\lambda^a}{\,,(a)} x^{a-1} e^{-\lambda x} \frac{\lambda^b}{\,,(b)} y^{b-1} e^{-\lambda y}.$$

Да разгледаме сега трансформациите:

$$U = \begin{cases} u = & x+y, \\ v = & \frac{x}{x+y} \end{cases} \qquad \text{и} \qquad V = U^{-1} = \begin{cases} x = & uv, \\ y = & u*(1-v) \end{cases}$$

и приложим формула (9.3). Тъй като якобианът на V е равен на u, получаваме за двумерната плътност на $\{\zeta,\theta\}$ формулата:

$$f(u,v) = \left(\frac{\lambda^{a+b}}{\,,(a+b)} u^{a+b-1} e^{-\lambda u}\right)\left(\frac{\,,(a+b)}{\,,(a)\,,(b)} v^{a-1}(1-v)^{b-1}\right),$$

откъдето следват всички твърдения на теоремата. ■

Тема 10

Правдоподобие

10.1 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 примат две "стойности" — истина и неистина, задачата всъщност е да намерим вероятността едно заключение да бъде верно.

10.1.1 Лема на Нейман–Пирсън

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

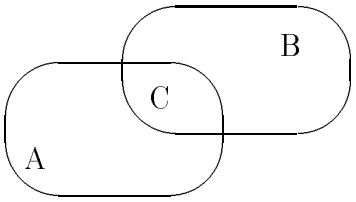
Лема 10.1 *(Нейман–Пирсън) Нека са дадени две плътности $f_0(x), f_1(x)$. Тогава решението на разпределителната задача:*

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

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

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

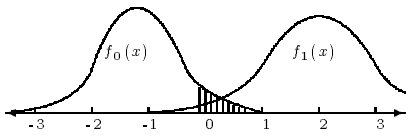
$$\begin{array}{lll} \int_W f_1(x)dx - \int_{W'} f_1(x)dx & = \int_A f_1(x)dx - \int_C f_1(x)dx & \geq \\ \int_A cf_0(x)dx - \int_C cf_0(x)dx & = c(\int_W f_0(x)dx - \int_{W'} f_0(x)dx) & = 0. \end{array}$$



Фигура 10.1: Лема на Нейман-Пирсън

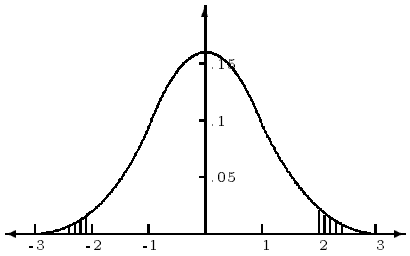
Тук сме означили $A = W \setminus W'$,$B = W' \setminus W$,$C = WW'$ или $W = A + C$,$W' = B + C$, както това е показано на фигурата.■

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



Фигура 10.2: Едностранен критерий

Числото α наричаме *грешка от първи род*, а числото $1 - \alpha$ - *ниво на доверие*. На фиг. 10.2 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 наляво.



Фигура 10.3: 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".

Възможна е и обратната грешка β– *грешка от втори род* – да приемем хипотезата, когато тя не е вярна. Естествено е нашето желание да търсим критичната си област така, че запазвайки α да минимизираме β. Лемата на Нейман - Пирсън ни дава средство лесно да строим *оптимални* критични

области. Тя може да се използува и за произволни функции от наблюденията. Числото $1 - \beta$ се нарича *мощност* на критерия (критичната област) и е различно за всяка конкретна алтернатива.

10.2 Хипотези за м.о. и дисперсия

Пример 10.1 *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 с ниво на доверие $1-\alpha$.*

Решение. Статистиката $\bar{x}$ има за плътности и при двете хипотези нормална плътност с еднаква дисперсия 1, но различни средни стойности. От лемата 10.1 следва, че оптималната критична област има вида:

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

Определяме константата от уравнението $1 - \alpha = \Phi(c\sqrt{n})$. Решенията на уравнението $\alpha = \Phi(z_\alpha)$ се наричат квантили на нормалното разпределение и вземат от таблица. Така определяме $c = z_{1-\alpha}/\sqrt{n}$ и оптималната ни критична област за критично ниво α става:

$$W = \{\bar{x} > \frac{1}{\sqrt{n}} z_{1-\alpha}\} \tag{10.1}$$

Ако разгледаме обратната (лява) алтернатива $H_1 : \xi \in N(-1,1)$, ще получим критична област отляво:

$$W = \{\bar{x} < \frac{1}{\sqrt{n}} z_{\alpha}\} \tag{10.2}$$

■

Нормалното разпределение е симетрично, така че $z_\alpha = -z_{1-\alpha}$. Затова в таблиците са дадени само квантилите за стойности на нивата на доверие $1 - \alpha > 0.5$. ■

Пример 10.2 *Нека $H_0 : \xi \in N(0, \sigma^2)$, а $H_1 : \xi \in N(1, \sigma^2)$, където $\sigma > 0$ е известно число. Нека сме направили n наблюдения. Намерете оптималния критерий.*

Решение.

$$W = \{\bar{x} > \frac{1}{\sqrt{n}} \sigma z_{1-\alpha}\} \tag{10.3}$$

■

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 10.1 се вижда, че за всички "десни" алтернативи (с м.о. по - високо от 0) решението ще бъде същото.

Определение 10.1 *Казваме, че критерият е равномерно най - мощен за дадено множество алтернативи, ако той е оптимален за всяка алтернатива поотделно.*

Така на фигура 10.2 е показан критерий, които е равномерно най - мощен за всички "десни" алтернативи.

Пример 10.3 *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 съшествува равномерно най - мощен критерий за това множество алтернативи.*

10.2.1 Разпределения, свързани с нормалното

Нека сл.в $\xi, \xi_1, \xi_2, \ldots, \xi_n \in N(0,1)$ са независими.

Определение 10.2 *Случайната величина* $\eta = \sum_{i=1}^{n} \xi_i^2$ *има разпределение* $\chi^2(n)$ *с* n *степени на свобода.*

Определение 10.3 *Частното*

$$t = \frac{\xi}{\sqrt{\eta/n}},$$

където ξ *и* η *са две независими сл.в.* $\xi \in N(0,1)$ *и* $\eta \in \chi^2(n)$ *с* n *степени на свобода има разпределение на Стюдент* $T(n)$ *с* n *степени на свобода .*

Определение 10.4 *Частното*

$$f = \frac{\xi/m}{\eta/n},$$

където ξ *и* η *са независими сл.в.* $\xi \in \chi^2(m)$ *и* $\eta \in \chi^2(n)$ *с* m *и* n *степени на свобода има разпределение на Фишер* $F(m,n)$ *с* m, n *степени на свобода.*

10.2.2 Критерий на Фишер

Нека проверим хипотезата за равенство на дисперсии на две различни генерални съвупности (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:

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

Всяка от тези статистики има хи-квадрат разпределение χ^2, умножено със съответната σ^2. Ако за нулева изберем хипотезата: $H_0 : \sigma(x) = \sigma(y)$, то 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, само критичните области ще са от различни страни на разпределението;
- За алтернативата $H_1 : \sigma(x) \neq \sigma(y)$ не съществува равномерно най - мощен критерий.

10.2.3 Критерий на Стюдент

Определението 10.3 дава възможност след като табулираме квантилите на разпределението на Стюдент за различен брой степени на свобода и нива на доверие, да създадем следните критерии. Да разгледаме статистиката:

$$t = \frac{\sqrt{n}\bar{x}}{\sqrt{S^2(x)/(n-1)}}$$

Нека сега предположим, че наблюденията са от $N(\mu, \sigma^2)$. Ако за нулева изберем хипотезата: $H_0 : \mu = 0$, то съгласно това определение статистиката t ще разпределение на Стюдент независимо от σ. Това ни дава възможност да проверяваме нулевата хипотеза срещу различни алтернативи:

- За алтернативите $H_1 : \mu < 0$ или $H_1 : \mu > 0$ критерият ще бъде равномерно най - мощен, само критичните области ще са от различни страни на разпределението:

$$W_{\mu>0} = \{t > t_{1-\alpha}\} = \{\bar{x} > \frac{1}{\sqrt{n}} s_n t_{1-\alpha}\},$$
$$W_{\mu<0} = \{t < t_{\alpha}\} = \{\bar{x} < -\frac{1}{\sqrt{n}} s_n t_{1-\alpha}\},$$
$$s_n^2 = \frac{1}{n-1} S^2(x)$$

  Тук с t_α сме означили квантила на разпределението на Стюдент със съответния брой степени на свобода. То е симетрично, както и нормалното, така че $t_\alpha = -t_{1-\alpha}$. При голям брой на наблюденията (ст.св.) $n > 120$ разпределението на Стюдент клони към стандартно нормално, така че тези критични области съвпадат с определените с формули (10.1) и (10.2) при $\sigma = 1$, защото $s_n \to \sigma$ п.с.

- За алтернативата $H_1 : \mu \neq 0$ не съществува равномерно най - мощен критерий, но можем да използуваме доверителните интервали, както ще видим в следващата лекция.

Тема 11

Оценяване на параметри

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*.

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

11.1 Определения

Определение 11.1 *Наричаме функция на правдоподобие $f(x, \theta)$ плътността на наблюдаваната сл.в. ξ, когато тя зависи от неизвестен параметър.*

Определение 11.2 *Казваме, че статистиката $\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$ е оценка на параметъра θ, ако $\hat{\theta}$ не зависи от стойността на параметъра.*

Определение 11.3 *Казваме, че оценката $\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$ на параметъра θ е неизместена, ако $\mathbf{E}\,\hat{\theta} = \theta$.*

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

Определение 11.4 *Казваме, че оценката $\hat{\theta}$ на параметъра θ е ефективна, ако е с минимална дисперсия сред всички неизместени оценки на този параметър.*

Определение 11.5 *Казваме, че редицата от статистики $\hat{\theta}_n$ е състоятелна оценка на параметъра θ, ако $\hat{\theta}_n \xrightarrow{p} \theta$ при увеличаване на броя n на наблюденията.*

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

11.2 Доверителни области и интервали

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

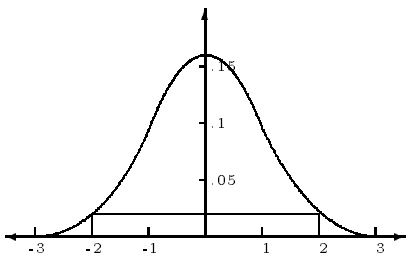
В такава постановка задачата много прилича на лемата на Нейман - Пирсън. Първоначално ще построим доверителна област за наблюдението, така че тя да има минимален обем. В последствие (при подходящи условия) тя ще се превърне в доверителна област за параметъра.

Лема 11.1 *Нека е дадена семейството плътности $f(x,\theta)$. Тогава решението на разпределителната задача:*

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

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

Доказателство: Абсолютно същото като на оригиналната лема.■



Фигура 11.1: Доверителен 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 доверителния интервал така, че плътността да бъде равна в двата края. Обикновено това е достатъчно за проверка на оптималността (минималната дължина) на така построения доверителен интервал.

Пример 11.1 *Нека $\xi \in N(\theta, 1)$. Нека сме направили n наблюдения. Намерете оптимална доверителна област за θ.*

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

$$|\bar{x} - \theta| < z/\sqrt{(n)}.$$

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 11.1 и е с минимална дължина. ■

Пример 11.2 *Нека $\xi \in N(0, \theta)$. Нека сме направили n наблюдения. Намерете оптимална доверителна област за θ.*

Решение Статистиката $S^2 = \sum_{i=1}^{n} x_i^2$ има разпределение $\theta\chi_n^2$ (определение 10.2). От лема 11.1 следва, че доверителна област с минимална дължина за θ се дава от неравенствата:

$$q_l \leq \frac{S^2}{\theta} \leq q_u.$$

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

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

$$f(q_l) = f(q_u),$$

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

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.

Пример 11.3 *Нека $\xi \in N(\theta, \sigma^2)$. Нека сме направили n наблюдения. Намерете оптимална доверителна област за θ. Тук втория параметър σ се смята за неизвестен.*

Решение Статистиката $t = \sqrt{n}(\bar{x} - \theta)/s$ има T-разпределение с $(n-1)$ степени на свобода и това разпределение не зависи от σ. Тук сме означили

$$s^2(x) = \frac{1}{n-1} \sum (x_i - \bar{x})^2.$$

От лема 11.1 следва, че оптималната доверителна област за θ е

$$|\bar{x} - \theta| < t_{1-\alpha/2} s/\sqrt{(n-1)}.$$

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

11.3 Н.О.М.Д.

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

Теорема 11.1 *(Рао - Блекуел) Неизместената оценка с минимална дисперсия (н.о.м.д.) е единствена (п.с.).*

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

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

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

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

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

11.4 Неравенство на Рао - Крамер

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

Теорема 11.2 *(Рао - Крамер) Ако θ е едномерен параметър и правдоподобието $f(x,\theta)$ удовлетворява:*

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

то е валидно следното неравенство:

$$\mathbf{D}\,(\hat{\theta}) \geq \frac{1}{\mathbf{E}\,((\frac{\partial \log f(x,\theta)}{\partial \theta})^2)}. \tag{11.1}$$

При това, равенство се достига само ако

$$\frac{\partial \log f(x,\theta)}{\partial \theta} = k(\theta)(\hat{\theta} - \theta). \tag{11.2}$$

За доказателството виж например [Янев, Димитров (1990)].

Горните две теореми дават лесно средство за проверка на ефективността на оценките - достатъчно е да се достигне равенство в неравенството на Рао - Крамер, т.е. да бъде изпълнено равенството (11.2).

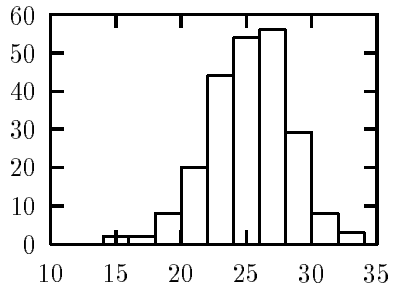
Тема 12

χ^2-критерий

Този знаменит критерий всъщност е съвокупност от статистически процедури основани на свойствата на χ^2 разпределението. В тази лекция ще се запознаем с двата най-популярни в практиката критерии - за съгласуваност на разпределения и за анализ на (двумерни) честотни таблици.

12.1 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.



Фигура 12.1:

Нека ни е зададена някаква теоретична плътност $p(x)$ и имаме за задача да проверим съгласуваността й с извадката. Нека разполагаме с n наблюдения. Разделяме множеството от стойности на сл.в. (или носителя на плътността) на k интервал 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 интервли да се обединяват за да може то да се удовлетвори. Съществува и вариант, когато интервалите се избират рвновероятни, т.е. $p_i = p_j$.

Нека означим с n_i броя на наблюденията попаднали в H_i. Пресмятаме статистиката

$$h = \sum_{i=1}^{k} \frac{(np_i - n_i)^2}{np_i}. \tag{12.1}$$

Теорема 12.1 *(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 (12.1) представлява квадрата на центрирана асимптотично нормална сл.в. Действително, $np_i = \mathbf{E}\, n_i$. За съжаление, тези величини са зависими – $\sum n_i = n$. Оказва се, че условното разпределение на сл. гаусов вектор $\xi \in N(0, I)$ в R^n при условие $(\xi, 1) = 0$ е същото като асимптотичното съвместно разпределение на сл.в. n_i, съответно центрирани и нормирани. ■

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

12.2 Независимост на частотни таблици

В много случаи данните се сумират в така наречените частотни таблици. Нека за всеки обект (наблюдение) си отбелязваме проявата на някакви признаци. Ако признаците са два можем да сумираме наблюденията си в таблица.

Ще илюстрираме концепцията със следния пример.

Пример 12.1 *(The Case of Luddersby Hall) 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 $\{$ял, неял$\}\times\{$болен, здрав$\}$. За болни ще смятаме студентите, които повръщат и тези които имат бактерии. Така получаваме следните три подтаблички:

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

Сега нека за всяка от табличките си формулираме задачата в термини на проверка на статистическата хипотеза:

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

срещу алтернативата:

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

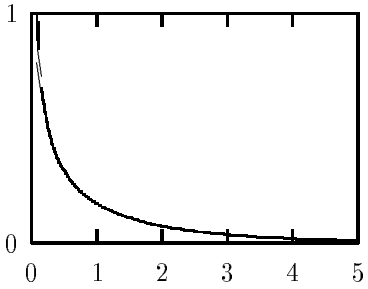
Прието е проверката на такава хипотеза да се прави по следния начин. Нека означим с $p_{i,j}$ вероятността случайно избран студент да попадне в клетката (i, j) на таблицата. Но тогава, ако признаците бяха независими, би трябвало тази вероятност да е произведение от двете маргинални вероятности: $p'_i = (\sum_k p_{i,k})$ и $p''_j = (\sum_k p_{k,j})$.

Така получаваме две извадъчни разпределения, за съгласуваността на които можем да използуваме χ^2-критерий. Нека първият признак има k, а вторият – m категории. Да пресметнем статистиката:

$$h = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(nn'_i n''_j - n_{i,j})^2}{nn'_i n''_j}. \tag{12.2}$$

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

Ако се върнем към нашия пример, 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 фигурата, а 0.95 квантила му е равен 3.84. Така, когато статистиката h, пресметната по формула (12.2), надхвърли критичната стойност 3.84 би трябвало да отхвърлим нулевата хипотеза.



Фигура 12.2: χ^2_1

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 13

Регресионен анализ

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

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

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

13.1 Линейни модели с гаусова грешка

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

$$y \quad = \quad z + \epsilon. \tag{13.1}$$

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

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

Теорема 13.1 *За модела (13.1) са изпълнени свойствата:*

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

$$\hat{y} = \mathop{\mathbf{arg\,min}}_{z \in \hat{Y}} ||z - y||^2;$$

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

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

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

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

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

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

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

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

Теорема 13.2 *Нека се налага да проверим хипотезата*

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

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

$$f_{m-k,n-m} = \frac{||y_1 - y_0||^2/(m-k)}{||y - y_1||^2/(n-m)} > F_{1-\alpha}, \tag{13.2}$$

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

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

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

Проверката на неравенството $\lambda(y) > c$ е еквивалентна на критичната област определена от неравенството (13.2). Твърдението за разпределението е пряко следствие от теоремата на Кокрън или от Питагоровата теорема:

$$||y - y_0||^2 = ||y - y_1||^2 + ||y_1 - y_0||^2. \tag{13.3}$$

■

Когато към модела (13.1) добавяме предположения за параметризация на Z, получаваме различните форми на, т.н. в литературата, общ линеен модел с гаусова грешка. Някои от тях ще разгледаме сега.

13.2 Нормална линейна регресия

Нека изследваният модел е от вида

$$y = Xa + e, \tag{13.4}$$

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

$$\hat{a} = (X'X)^{-1}X'y \tag{13.5}$$

$$cov(\hat{a}) = \sigma^2(X'X)^{-1} \tag{13.6}$$

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 13.1 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 (13.5). Това решение съществува и е единствено поради пълния ранг на X.

Като заместим y в (13.5) получаваме

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

което влече неизместеността на $\hat{a}$. От същото представяне следва и представянето на $cov(\hat{a})$ в (13.6). ■

От теорема 13.1 веднага получаваме, че неизместена оценка на σ^2 ще получим по формулата:

$$\hat{\sigma^2} = \frac{1}{n-k}||y - X\hat{a}||^2. \tag{13.7}$$

Тази оценка, обаче, не е максимално правдоподобна.

Тема 14

Проверки на хипотези в регресията

В тази лекция ще експлоатираме безпощадно теорема 13.2 и ще конструираме множество популярни хипотези в линейната регресия. В някои частни случаи конструираните доверителни области (поради естествените "широки" алтернативни хипотези) ще станат и доверителни интервали за неизвестните параметри.

В тази секция ще разгледаме модела в уравнение (13.4) и процедурата продиктувана от теорема 13.3. Това ще ни помогне при анализа на други аналогични модели. Ще разгледаме накратко хипотезите в реда, в които те се използуват на практика.

14.1 Коефициент на детерминация

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

Нека разгледаме регресионен модел със свободен член:

$$y = Xa + b\vec{1} + e, \tag{14.1}$$

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 $H_0 : a = 0$. Естествената контра хипотеза е $H_1 : a \neq 0$. Следователно, Z_0 има размерност $k = 1$, а Z_1 е с размерност $m = dim(a) + 1$. От теорема 13.2 получаваме, че критичната област за проверка на хипотезата $H_0 : z \in Z_0$ срещу хипотезата $H_1 : z \in Z_1 \backslash Z_0$ се определя от неравенството:

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

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

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

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

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

Частното

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

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

Проверката за равенство на нула на R^2 е първото действие, което изследователят трябва да предприеме, когато започва анализа на някой модел. Наистина трудно можем да се надяваме, че моделът е добър, ако няма значима връзка между предикторите и отклика.

14.2 Проверка за равенство на нула на някой от коефициентите

Нека е вярна хипотезата $H_0 : a_1 = 0$. Естествената контра хипотеза е $H_1 : a_1 \neq 0$. Следователно, Z_0 има размерност $k = dim(a) - 1$, а Z_1 - размерност $m = dim(a)$. От теорема 13.2 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

$$t_i = \frac{\hat{a}_i}{\hat{\sigma}^2((X'X)^{-1}_{ii})^{1/2}} \tag{14.2}$$

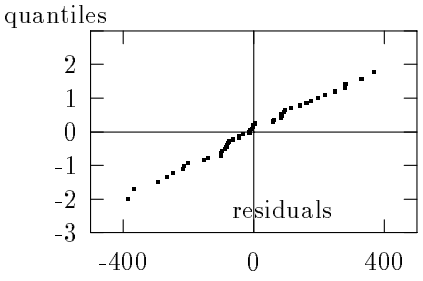
имат разпределение на Стюдент с $n - m$ степени на свобода при изпълнена хипотеза $H_0 : a_i = 0$. Естествено, със същото разпределение се пресмятат и доверителните интервали около оценките за неизвестните параметри (при изпълнена H_1). Това следва от неизместеността им и от това, че оценките на параметрите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

14.3 Анализ на остатъците

В тази група влизат различни, главно визуални средства за проверка на адекватността на модела. Те са основани на различни графики - хистограма, scatter plot, нормална хартия.

- нормална хартия;

Проверява се предположението за нормалност на остатъците. За това може да се използува обикновена хистограма, но може и т.н. нормална хартия. Съответните квантили на гаусовото разпределение. се рисуват по отношение на вариационният им ред. Силни или закономерни отклонения от правата линия са свидетелство за нарушение на това предположение.

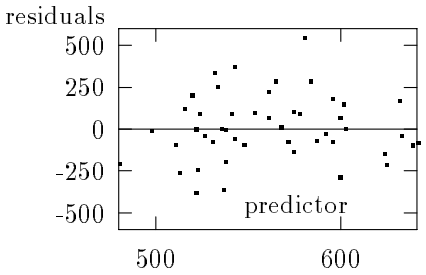


Фигура 14.1: Нормална хартия

- хетеро-скедастичност, нелинейност;

Тук въз основа на нарисуваните остатъци по отношение на някой предиктор се проверява предположението за постоянство на дисперсията – тя не трябва да зависи от стойността на предиктора.

Разположение на остатъците в определена нелинейност говори за необходимост от преразглеждане на модела и включването в него на някоя подходяща функция на този предиктор.



Фигура 14.2: Остатъци/предиктор

- стандартизирани остатъци;

  Това са нормирани с помощта на стандартното си отклонение остатъци (виж формула (14.4)). С тяхна помощ се анализират конкретни наблюдения, при които предположението на модела се нарушава. Наблюденията с голям стандартизиран остатък са подозрителни. Това може да се дължи на груби грешки при измерване на отклика или на други причини. За тези наблюдения моделът ни може да не е верен.

- jack-knife остатъци;

  Това е остатък на i-тото наблюдение, получен от модел, оценен без помощта на това наблюдение. Така намираме наблюдения, които оказват прекомерно голямо влияние на модела – тяхното отстраняване тотално го променя и следователно техният jack-knife остатък ще бъде голям. Това може да се дължи на груби грешки при измерване на някои от предикторите. За тези наблюдения моделът ни може да не е верен.

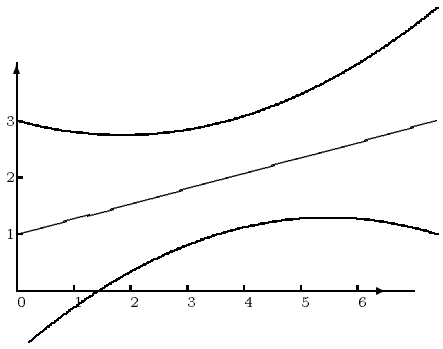
14.4 Доверителен интервал за прогноза

За произволни стойности x на предикторите от областта, за която е верен модела (14.1), случайната величина $\hat{y} = x'\hat{a} + \hat{b}$ е неизместена оценка за $E(y|x)$ и

$$D(\hat{y}|x) = \sigma^2(\frac{1}{n} + (x - \bar{X})'(\tilde{X}'\tilde{X})^{-1}(x - \bar{X})). \tag{14.3}$$

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

$$\sigma_y^2(x) = \sigma^2(1 + \frac{1}{n} + (x - \bar{X})'(\tilde{X}'\tilde{X})^{-1}(x - \bar{X})). \tag{14.4}$$



Фигура 14.3: Проста линейна регресия

Проверете уравнения (14.3) и (14.4).

На фигурата е нарисувана апроксимиращата права при простия линеен модел $y = ax + b + \epsilon$. С двете параболи са отбелязани доверителните граници за наблюдаваната стойност съгласно формула (14.4). С аналогична форма, но значително по - тесен е коридорът за модела – формула (14.3). Така се вижда колко опасни (и понякога безсмислени) могат да бъдат прогнози за далечното бъдеще, основани на тенденция, наблюдавана в краен интервал от време.

14.5 Проверка на адекватността на модела

Проверката за адекватност на модела в регресионния анализ е възможна само в два случая: ако е известна σ^2 или ако разполагаме с независима от SSR и от параметрите на модела нейна оценка.

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

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

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

Най-популярния начин за получаване на независима оценка за σ^2 е да се провеждат повторни наблюдения при фиксирани стойности на предикторите. При такива наблюдения сумата SSR също се разлага на две независими събираеми, от които се конструира

статистика, която има разпределение на Фишер, в случай че моделът е адекватен. Обикновено тази задача се решава със средствата на еднофакторния дисперсионен анализ. Отделните експериментални точки x се разглеждат като нива на фактор (групираща променлива). За всяко x имаме по n_x наблюдения $y_i(x)$. Имаме равенството:

$$SSR = \sum_x (y_i(x) - \bar{y}(x))^2 + \sum_x n_x(\bar{y}(x) - \hat{y}(x))^2 = SSI + SSM. \tag{14.5}$$

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

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

Опишете подпространствата Z_0 и Z_1 в този случай и изведете уравнение (14.5). Постройте критичната област.

Както видяхме тази процедура не винаги е възможна. Когато, обаче това е възможно, то е необходима да се проведе тази проверка. Понякога се налага да се разделят данните и моделирането да се проведе поотделно за различните извадки. Това дава възможност със стандартните методи за независими извадки да се провери съгласуваността на моделите.

Тема 15

Апроксимация на плътности

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

15.1 Криви на Пирсън

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

$$\frac{dp(x)}{dx} = \frac{x-a}{b_0 + b_1 x + b_2 x^2} p(x) \tag{15.1}$$

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

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

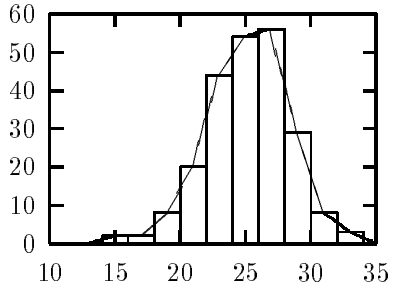
Коефициентите в уравнението (15.1) се определят еднозначно от първите 4 момента на разпределението. Това дава възможност, замествайки теоретичните с извадъчните моменти и решавайки уравнението, да получим смислена оценка на плътността, тъй като те - м.о.,дисперсията, асиметрията и ексцеса - доста прилично описват формата на разпределението.

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

Подробно описание на типовете криви на Пирсън и методите за оценка на параметрите им може да се намери у [Поллард (1967)], [Митропольский, (1963)]

15.2 Изглаждане на хистограми

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



Фигура 15.1: Съдържания на апатит

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

15.3 Ядра на Розенблат - Парзен

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

$$\hat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K(\frac{x_i - x}{h_n}), \tag{15.2}$$

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

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

$$f(x) = f(x_0) + \sum_{i=1}^{k} \frac{f^i(x_0)}{i!}(x - x_0)^i + o(|x - x_0|^k) \tag{15.3}$$

Разбира се, то има смисъл, ако съществуват производните на неизвестната плътност f до ред k в точката x_0. Като поставим $x - x_0 = yh_n$ и използуваме (15.3), получаваме, че изместването B_n на оценката е

$$B_n = \mathbf{E}\,\hat{f}_n(x_0) - f(x_0) = \int K(y)(f(x_0 + y * h_n) - f(x_0))dy =$$
$$f'(x_0)h_n \int yK(y)dy + f''(x_0)\frac{h_n}{2} \int y^2 K(y)dy + \cdots = O(h_n^2)$$

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

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

Така като използуваме равенството

$$\mathbf{E}\,(\hat{f}_n(x_0) - f(x_0))^2 = D_n + B_n^2 = O(\frac{1}{nh_n}) + O(h_n^4), \tag{15.4}$$

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

Приложение А

Таблици

$df \mid p$	.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 А.1: Хи-квадрат разпределение (квантили)

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

Таблица А.2: Нормално разпределение

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

Таблица А.3: Разпределение на Стюдент (квантили)

Библиография

[Янев, Димитров (1990)] Б.Димитров, Н.Янев, *Теория на вероятностите и математическа статистика*, С.1990.

[Чобанов (1992)] Г. Чобанов, *Теория на вероятностите за физици,* С.,1992

[Уилкс (1967)] Уилкс С., *Математическая статистика*, М., Наука, 1967.

[Поллард (1967)] Дж. Поллард, *Справочник по вычислительным методам статистики*, М., Финансы и статистика, 1982.

[Вындев, Матеев (1988)] Вындев Д., Матеев П., *Статистика с Правец*, Наука и изкуство, С., 1988.

[Афифи (1982)] А. Афифи, С. Айзен, *Статистический анализ. Подход с использованием ЭВМ.*, Мир, М., 1982.

[Митропольский, (1963)] Митропольский Г., *Техника статистических вычислений,* М.,Наука,1963.

[Кокс, Снелл (1984)] Кокс Д., Снелл Э., *Прикладная статистика. Принципы и примеры,* М., Мир, 1984, стр. 183 - 185.

[Дрейпер, Смит (1973)] Дрейпер Н., Смит Г., *Прикладной регресионный анализ*, Статистика, М., 1973.

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

[Себер (1976)] Себер Дж., *Линейный регресионный анализ*, Мир, М., 1976.

[Идье (1976)] Идье,В., Драйард, Д., Джеймс, Ф., Садуле Б., *Статистические методы в экспериментальной физике*, М.,Атомиздат, 1976.