Откриване на знания в данните
или
Полезните статистически методи: теория, софтуер, приложения

Веска Нончева

Април 2010

Съдържание

Глава 1

Вместо увод

Всеки от нас се е сблъсквал с несигурността. Малцина са щастливците, които могат винаги да взимат решения при достатъчна информация.

Докато един инженер може да измери колко е широк един конструктивен елемент с рулетка (да проведе "експеримент"), то един специалист по маркетинг няма такъв уред, с който да "измери" желанието на хората да потребяват конкретния продукт, а в някои области дори това понятие "експеримент" е силно ограничено като обхват - един икономист не може да повтори предната ситуация на пазара за да изследва влиянието на взетото от него решение как да се развива фирмата.

В такива ситуации на помощ идва науката "статистика" и методите й за извличане на информация и знания от данни.

Разбира се, така посочените примери са една много малка част от ползите, които човек може да извлече чрез прилагането на статистиката. Практически всяка област на съвременната наука и технология критично зависи от използването на един или друг вид вероятностен метод за оценка на резултатите от експериментите, за прогнозиране на бъдещи резултати и управление на наличните ресурси в съответствие с тези прогнози.

1.1 За книгата

Тази книга е предназначена за настоящи и бъдещи бизнес лидери и за всички, които имат кураж да си задават трудни въпроси и готовност да търсят техните отговори.

С четенето на тази книга читателят ще види силата на статистиката, ще усвои основните понятия и полезни статистически методи, ще разбере за съществуването на мощен и достъпен статистически софтуер и ще започне да го използва.

Читателят ще се научи сам да решава някои често срещани в неговата практика задачи.

Тази книга е един опит да се излезе от рамките на стандартното, да се мисли различно и работи ефективно.

1.2 За автора

Авторът има опит в прилагането на стохастични методи за откриване на измами, натрупан по време на работата му като национален експерт в Изследователски център на Европейската комисия, за повишаване на ефективността на селскостопански ферми, натрупан по време на работата по проект с Правителството на Азорите, за намиране на синоними в огромни масиви от текстове, натрупан по време на специализация в Нов Лисабонски университет, за прогнозиране на тежки метали в организма на рибите и при изследване на ДНК редици. Той чете лекции по анализ на данни пред различни аудитории.

Настоящата книга в голяма степен съдържа материали представени на лекциите, семинарните и лабораторните занятия по вероятности, статистика и статистически софтуер със студентите от бакалавърските и магистърските програми във Факултета по математика и информатика на Пловдивския университет. В нея са включени и материали от различни курсове, които авторът е чел през последните години пред аудитории от други специалности в университети на други европейски страни.

1.3 Благодарности

Изказвам искрена благодарност на рецензента ... за направените уместни забележки и ценни препоръки, както и на моите студенти за помощта, която много спомогна за подобряване качеството на книгата. ...

Глава 2

Защо си заслужава да се учим да анализираме данни

Средата, в която живеят компаниите днес, се характеризира с повишена турбулентност и хаос. Едни цени се качват драматично, други падат със същата скорост в рамките на кратки времеви цикли. Това налага да се реагира за кратко време, време, по-малко от времето за което компаниите са реагирали до сега. Тази реакция изисква познаването и използването на различни математически модели както в периода на рецесия, така и в периода на подем.

Новата реалност изисква от бизнес лидерите възприемане на нови стратегически модели на поведение с цел минимизиране на рисковете и увеличаване на възможностите пред компаниите им. Съвременните бизнес ръководители в периодите на променливост се нуждаят от обмислени действия, насочени към откриване и коригиране на слабостите и неефективностите, както и от знания за нови поведенчески модели, основаващи се на анализ на фирмени данни.

Как ще разберете какво Вашите клиенти ценят най-много? Попитайте ги, анализирайте техните отговори като използвате научнообосновани методи и вземете правилното решение.

За да насърчим компаниите в съвременния бизнес свят да мислят позитивно и да търсят новите възможности, които турбулентността може да създаде, ще обединим идеите за анализ на фирмени данни със съвременен софтуер и съвети за прагматични действия, които съвременните бизнес лидери могат да предприемат. Предлагаме

следните три конкретни стъпки:

1. Направете стратегическото планиране по-динамично, вместено в по-кратки времеви цикли.

2. Поощрявайте вътрешно фирменото взимане на решения, подпомогнато от съвременните методи за анализ на данни, за да намирате по-добрите решения. За тази цел дайте на вашите експерти нови знания и съвременен софтуер.

3. Повишете професионалните и личните умения за извличане на знания от дани на целия управленчески екип, за да се подобри качеството на решенията.

Непосредствената задача сега на бизнес лидерите и изпълнителните им екипи е да установят нови поведенчески модели и стратегии, подходящи за новите условия.

Увеличаващата се турбулентност вече е факт от реалността! В нормални обстоятелства проявите на слаба ефективност са търпими, но във времена на икономическа турбулентност неефективните елементи увеличават уязвимостта на компанията. Преди да пристъпят към крути мерки - съкращения на служители, преструктуриране или закриване на отдели, ръководителите трябва да открият неефективните звена.

Най-ефективният начин да решим тази задача е да възприемем прагматичния подход – подходът на смислов анализ на данни от всички ключови бизнес действия с цел правене на прогнози, на чиято основа бизнес лидерите да понижат рисковете от неприятни изненади по време на криза.

Глава 3

Как да четем тази книга

Тази книга, подобно на всеки друг учебник, може да бъде четена по различни начини в зависимост от интересите и профила на читателя.

Незапознатите с теорията на вероятностите е важно да прегледат глава 4. В тях са изложени основните елементи на теорията и практиката. Но дори запознатите могат да научат нещо ново, тъй като навсякъде сме се стремили да даваме илюстрации на смисъла на дефинициите.

В глава 5 са представени основните понятия на статистиката и началният метод за обработка на данните във всяка задачата - графичният.

За по-нататък, ако сте решили да се учите от тази книга, настойчиво ви съветваме да я четете на удобно място, близко до компютър, на който са инсталирани средата *R* с пакетите *rggobi*, *cluster* и *MASS*, и да пробвате дадените съвети. Инсталирането на софтуера е показано в приложение **??**.

Оттук нататък реда на четене е въпрос на личен избор. В текста сме се стремили да оставяме достатъчно препратки към теорията, така че когато разчитаме на материал от предни глави читателят ще може да се ориентира къде да го търси. Разбира се, подредбата на главите е съобразена с това, така че читателят, които реши да чете книгата като обикновен учебник ще може да го направи без проблеми.

В статистическата литература се срещат много начини за изра-

зяване на един и същи термин. Надяваме се читателя да намери за полезни азбучния указател в края на книгата, както и краткия речник на термините на английски език с техния превод.

3.1 Означения, използвани в книгата

В книгата има много примери за приложението на статистиката в различни области от практиката. Те са отделени визуално от текста по този начин:

Пример 1 (Реален пример)**.** *и неговото описание.*

Решението на поставената задача се намира вътре в самия текст.

Веднага след дефинирането на задачата се представя теорията, необходима за нейното решаване. Обикновенно това е един статистически метод.

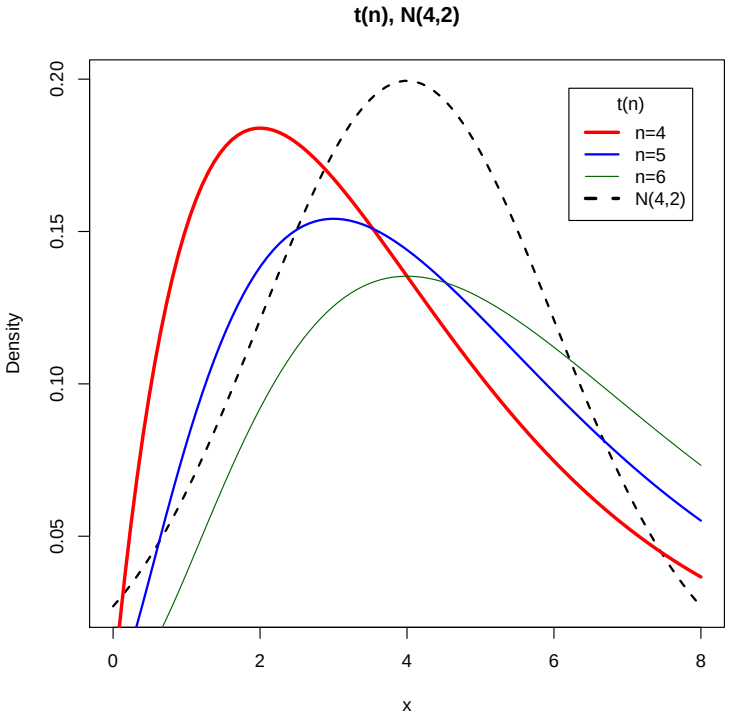
Някои методи решават често срещани в практиката задачи и имат свои имена. Те са цитирани с името си. Други методи се използват рядко и нямат име. Те са цитирани описателно. Например методът за проверка на хипотези за средното на популация, която има нормално вероятностно разпределение с неизвестна дисперсия, е известен с името *критерий на Стюдент*. А методът, който решава същата задача при известна дисперсия няма свое име и затова в книгата се цитира описателно.

Тази книга може да бъде полезно учебно пособие за всеки, който иска да се научи да анализира данни. За тази цел са дадени както теоретичните основи на методите, така и командите на R, реализиращи разгледаните статистически модели, така и кратки скриптове, реализиращи показаните статистически процедури и модели. Кодът на R ще бъде представян в такива полета, заедно с текстовия изход от програмата:

```
x <- seq(0, 8, length=100)
hx <- dnorm(x, mean = 4, sd = 2)
degf <- c(4, 5, 6)
colors <- c("red", "blue", "darkgreen", "black")
labels <- c("n=4", "n=5", "n=6", "N(4,2)")
lwdd <- c(3,2,1)
plot(x, hx, type="l", lty=2, lwd=2, xlab="x",
  ylab="Density", main="t(n), N(4,2)")
for (i in 1:3){
  lines(x, dchisq(x,degf[i]), lwd=lwdd[i], col=colors[i])
```

```
}
legend("topright", inset=.05, title="t(n)",
labels, lwd=lwdd, lty=c(1, 1, 1, 2), col=colors)

dev.copy2pdf(file="F:\\chi-distributions-1.pdf")
```



Фигура 3.1: Сравняване на положително асиметрични плътности с нормалната

Когато резултата е отпечатана на екрана графика, тя ще се вижда като тази на фигура 3.1.

Математическите символи, използвани в книгата: (??? DA GI SLOJA ???)

Глава 4

Ако не знаем точно, знаем с вероятност

Има ли случайности около нас? Кое е случайност и кое не е?

Да погледнем хазартната индустрия. Компаниите влагат милиони долари, за да изграждат хотели, които привличат клиенти към техните казина. Пример за това е Лас Вегас. Казината там правят постоянна печалба ден след ден и година след година, използвайки събития, които имат напълно случаен изход. Хазартните корпорации са като всички други корпорации. Как да си обясним факта, че те оправдават влагането на огромни по размер суми в хотели и казина, чиято основна функция е да създават доход от събитие, което има напълно случаен изход?

Класическата икономическа теория казва, че няма начин дори професионалният спекулант в борсовата търговия предварително да знае какво ще бъде поведението на другите участници и как това поведение ще повлияе върху сделката му. Следователно изходът от сделката му е неопределен, случаен. [1] В същото време управляващите на инвестиционните фондове могат да намалят риска от загубата чрез подходящо моделиране на пазара с инструментите на теорията на вероятностите.

И далеч не на последно място, при всяко изследване (на пазара, в лаборатория, маркетингово и т.н.) има известна степен на не-

[1] Повечето спекуланти (погрешно) вярват, че движението на пазарът може да се предвижда. Това не е неоходимо за целите на управлението на риска.

сигурност, породена както от обективни причини (ограничения на експерименталните условия), така и от присъщата на реалния свят неопределеност в един или друг смисъл. Понякога е непрактично тези неопределености да се игнорират, и естествен апарат за тяхното моделиране се явява теорията, изложена по-долу.

4.1 Случайни събития

В тази глава ще се научим да пресмятаме шанса за сбъдване на събития.

Първите опити да се построи математически модел са свързани с понятието равен "шанс". Предполага се, че даден опит има краен брой равновероятни изходи. При провеждане на опита се сбъдва някой от тях, при това всеки може да се случи с еднакъв "шанс". Най-простите примери за такава концепция са свързани с хазартните игри, където се хвърлят зарове или се използва добре разбъркано тесте карти.

Дефинираме понятието класическа вероятност p по следния начин:

$$p = \frac{m}{n}$$

където n е брой на всички възможни изходи, m е брой на благоприятните изходи. Тази дефиниция е известна като *класическа дефиниция за вероятност*.

Теорията на вероятностите става строга математическа теория едва след въвеждането на аксиоматика през 1939 г. от А. Н. Колмогоров. Ще започнем с представянето на Теория на вероятностите като въведем събития и действия с тях, определим какво разбираме под вероятностно пространство и дадем примери.

Елементарно събитие w е първично понятие – подобно на точка в геометрията. В примера с тестето карти това е всяко събитие от вида *изтеглили сме дама купа*.

Събитие наричаме множество от елементарни събития A. Например събитие може да бъде *изтеглили сме дама*, като неговите елементи са събитията *изтеглили сме дама купа*, *изтеглили сме дама каро*, *изтеглили сме дама пика*, *изтеглили сме дама спатия*.

Казваме, че събитието A е настъпило, ако е настъпило някое от

елементарните събития $\omega \in A$.

Множеството от всички елементарни събития логично наричаме *достоверно събитие* и означаваме с Ω. Празното множество бележим с ϕ и наричаме *невъзможно събитие*.

Тъй като всички събития са подмножества на Ω, то с тях могат да се извършват обичайните действия в теория на множествата: допълнение, обединение, сечение. Резултата от операцията е отново множество от елементарни събития (и следователно е събитие!), което има връзка с началните събития:

- Допълнението $\Omega \setminus A$ на множеството A в Ω означаваме с $\overline{A}$ и наричаме *допълнително събитие* (или *отрицание*) на събитието A.
- Сечението на множествата A и B означаваме с $A \cap B$ (а понякога за краткост с AB) и казваме, че събитията A и B са се сбъднали съвместно.
- Обединението на множествата A и B означаваме с $A \cup B$ и казваме, че се е сбъднало поне едно от събитията A или B. За краткост това се произнася "сбъднало се е A или B". За несъвместими събития вместо $A \cup B$ обикновено използваме знака за събиране и пишем $A + B$.

Когато A и B са свързани с релацията $A \subset B$ 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 законите на Де Морган.

Събитията A и B наричаме *несъвместими*, ако е изпълнено $AB = \emptyset$.

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

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

1. $\Omega \in \mathcal{A}$;
2. ако $A \in \mathbf{A}$, то $\overline{A} \in \mathbf{A}$;

3. ако $A \in \mathbf{A}$, $B \in \mathbf{A}$, то $A \cup B \in \mathbf{A}$.

Буловата алгебра не е длъжна да бъде затворена относно операции с безкраен брой множества. Булова алгебра $\mathbf{A}$, която е затворена относно изброимите операции обединение и сечение, се нарича булова σ - алгебра. Следователно, ако $A_k \in \mathbf{A}$ (k=1,2,...), то $\cup A_k \in \mathbf{A}$ и $\cap A_k \in \mathbf{A}$.

Определение 1. *Елементите на буловата σ - алгебра $\boldsymbol{A}$ се наричат случайни събития.*

Определение 2. *Реалната функция $\mathbb{P}$, определена върху елементите на буловата σ- алгебра $\boldsymbol{A}$, се нарича вероятност, ако удовлетворява условията:*

1. *неотрицателност: $\mathbb{P}(A) \geq 0$, за всяко $A \in \boldsymbol{A}$;*
2. *нормираност: $\mathbb{P}(\Omega) = 1$;*
3. *адитивност: Ако $A_i \cap A_j = \phi$ когато $i \neq j$, то $\mathbb{P}(A_1 + A_2 + ...) = \mathbb{P}(A_1) + P(A_2) + ...$*

Определение 3. *Тройката $(\Omega, \boldsymbol{A}, \mathbb{P})$ наричаме вероятностно пространство.*

От аксиомите лесно следват следните свойства на случайните събития.

Нека A и B са произволни случайни събития. Тогава вероятността за настъпване на поне едно от тях се пресмята по следната формула: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$, известна като *формула за събиране на вероятности.*

Определение 4. *Нека B е случайно събитие с положителна вероятност. За всяко случайно събитие A числото $\mathbb{P}(A|B) = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)}$ се нарича условна вероятност на събитието A при условие събитието B.*

Теорема 1. *Нека B е случайно събитие с положителна вероятност. Тогава функцията $\mathbb{P}(.|B)$ е вероятност. т.е. $\mathbb{P}(A|B)$ е вероятност за всяко случайно събитие A.*

Следователно, ако фиксираме условието (т.е. събитието B), условната вероятност притежава всичките свойства на безусловната. Събитието B (и всички съдържащи го събития) притежава условна вероятност 1. Несъвместимите с B събития стават "невъзможни". Така върху същата σ- алгебра е породена нова вероятност отразяваща факта за настъпването на събитието B. Нека я означим с $\mathbb{P}_B$. Тогава всички твърдения за безусловната вероятност $\mathbb{P}$ са в сила и за условната вероятност $\mathbb{P}_B$.

Нека $A_1, A_2, ..., A_n$ са събития. Вероятността за съвместното им настъпване се пресмята по следната формула

$$\mathbb{P}(A_1A_2...A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_1A_2)...\mathbb{P}(A_n|A_1A_2...A_{n-1})$$

известна като *формула за умножение на вероятности.*

Независимостта е най-фундаменталното понятие на теорията на вероятностите. Независимостта в действителност лежи в основата ѝ. Независимостта, като строго понятие от математиката, се оказва неимоверно близка до нормалните, езикови и човешки представи за същото, когато едно събитие оказва (или не оказва) някакво влияние върху възможността друго събитие да настъпи. Но независимостта има и недостатъци - изискванията са толкова строги, че стават на практика непроверяеми. Когато казваме, че две събития са независими, ние влагаме много повече вяра, отколкото бихме могли да проверим с формални средства.

Регистрацията на настъпването на дадено случайно събитие променя състоянието на вероятностното пространство. Вече е невъзможно настъпването на елементарни събития извън това събитие, т.е. на елементарни събития невлачещи това събитие. Тази ситуация е отразена в изменението на вероятността на другите събития. Условната им вероятност невинаги е равна на безусловната.

Пример 2 (Вредители по земеделските култури)**.** *От началото на текущия месец е регистриран летеж на основния неприятел по черешата - черешовата муха. Масив от череши се третира срещу този икономически важен вредител с два препарата. Известно е, че действието на всеки от препаратите не зависи от това, дали е използван другия. Фирмите производители твърдят, че вероятността първият препарат да е ефективен е 80%, а вероятността*

вторият да е ефективен е 70%. Да се намери вероятността нито един от препаратите да не се окаже ефективен.

В някои редки случаи, обаче настъпването на някои събития не оказва влияние върху шансовете на други събития да настъпят. 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 5. *Казваме, че събитието A и събитието B, за което $\mathbb{P}(B) > 0$, са независими събития, ако е изпълнено равенството $\mathbb{P}(A|B) = \mathbb{P}(A)$. Ще бележим независимите събития по следния начин: $A \perp B$.*

Теорема 2. *Нека $\mathbb{P}(A) > 0$ или $\mathbb{P}(B) > 0$. Необходимото и достатъчно условие събитията A и B да бъдат независими е $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$.*

От тази теорема веднага следва, че условната вероятност на всяко от двете събития при условие, че е настъпило другото, е равна на безусловната му вероятност. С други думи, вероятността да настъпи събитието A не зависи от това, дали е настъпило или не, събитието B.

Нека A=\{Първият препарат не е ефективен\}, B=\{Вторият препарат не е ефективен\}. Тогава $\mathbb{P}(A) = 1-0.8 = 0.2$ и $\mathbb{P}(B) = 1-0.7 = 0.3$. Търсената вероятност е $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B) = 0.06$.

Определение 6. *Казваме, че събитията $A_k, k = 1, 2, ..., n$ 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 силно опростяване на формулата за умножение на вероятности, когато събитията са независими в съвкупност: $\mathbb{P}(A_1A_2...A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2)...\mathbb{P}(A_n)$

Нещо повече, тя е изпълнена и за произволно избрани k събития от тези n събития ($k < n$).

Пример 3 (Коя е печалившата стратегия?)**.** *Вие сте в телевизионно предаване, в което Ви се дава възможност да изберете една от три врати. Зад една от вратите стои кола, а зад останалите две стоят кози. Вие избирате врата. Да я означим с номер* 1*. Тогава водещият, който знае къде е колата, ви отваря една от останалите две врати зад която стои коза. Да означим отворената врата с номер* 3*. Тогава водещият Ви предлага да изберете врата номер* 2*. Но Вие вече сте избрали врата номер* 1*. Заслужава ли си да смените своя първоначален избор на врата номер* 1 *с врата номер* 2*?*

Определение 7. *Казваме, че събитията* $(H_1, H_2, ..., H_n)$ *образуват пълна група от събития в* Ω*, когато събитията са несъвместими и изчерпват достоверното събитие (т.е.* $H_1 + H_2 + ... + H_n = \Omega$*).*

Прието е събитията от пълната група да се наричат *хипотези.*

Нека е зададена пълната група събития $(H_1, H_2, ..., H_n)$. Тогава за всяко случайно събитие A е изпълнена следната *формула за пълната вероятност* $\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(H_i)\mathbb{P}(A|H_i)$.

Следната знаменита *формула на Бейс* намира широко приложение в статистиката, информатиката и в много други области:

$$\mathbb{P}(H_i|A) = \frac{\mathbb{P}(H_i)\mathbb{P}(A|H_i)}{\sum_{j=1}^{n} \mathbb{P}(H_j)\mathbb{P}(A|H_j)}$$

където събитията $(H_1, H_2, ..., H_n)$ образуват пълна група от събития. Тя изчислява обновените (апостериорните) вероятности за настъпване на хипотезата H_i, при условие, че вече е настъпило събитието A.

Да се върнем към телевизионната игра. Нека въведем събитията $H_i = \{$ колата е зад врата $i\}$, $i = 1, 2, 3$. В началото на играта $\mathbb{P}(H_i) = 1/3$, $i = 1, 2, 3$. Събитията H_i, $i = 1, 2, 3$, образуват пълна група от събития. Нека означим $A_i = \{$водещият отваря врата $i\}$, $i = 1, 2, 3$.

Нашата цел е да пресметнем условните вероятности $\mathbb{P}(H_1|A_3)$ и $\mathbb{P}(H_2|A_3)$.

Нека вероятността водещият да отвори врата 3, когато колата е зад врата 1, е $\mathbb{P}(A_3|H_1) = a$, $0 < a < 1$. Тогава вероятността водещият да отвори врата 2, когато колата е зад врата 1, е $\mathbb{P}(A_2|H_1) = 1 - a$,

защото вероятността водещият да отвори врата 1, когато колата е зад врата 1, е $\mathbb{P}(A_1|H_1) = 0$.

Ако водещият избира по случаен начин една от вратите зад която стои коза, когато колата е зад врата 1, то $a = 1/2$.

По условие вероятността $\mathbb{P}(A_3|H_3) = 0$. Вероятността $\mathbb{P}(A_3|H_2) = 1$.

Вероятността водещият да отвори врата 3 от формулата за пълната вероятност е

$$\mathbb{P}(A_3) = \mathbb{P}(H_1)\mathbb{P}(A_3|H_1) + \mathbb{P}(H_2)\mathbb{P}(A_3|H_2) + \mathbb{P}(H_3)\mathbb{P}(A_3|H_3)$$
$$= \frac{1}{3}a + \frac{1}{3}1 + \frac{1}{3}0 = \frac{a+1}{3}$$

От формулата на Бейс получаваме търсените вероятности

$$\mathbb{P}(H_1|A_3) = \frac{\mathbb{P}(H_1)\mathbb{P}(A_3|H_1)}{\mathbb{P}(A_3)} = \frac{a}{a+1}, \mathbb{P}(H_2|A_3) = \frac{\mathbb{P}(H_2)\mathbb{P}(A_3|H_2)}{\mathbb{P}(A_3)} = \frac{1}{a+1}$$

От предположението, че водещият избира по случаен начин една от вратите 2 или 3 зад която стои коза, когато колата е зад врата 1, получаваме $a = 1/2$. Тогава вероятността колата да е зад първата врата, при условие, че водещият е отворил третата врата е $1/3$. Вероятността колата да е зад втората врата, при условие, че водещият е отворил третата врата е $2/3$.

Следователно смяна на вратата е по-добрата стратегия. Тъй като водещият винаги отваря врата, зад която стои коза, играчът трябва да се поучи от новата информация, която в случая е, че зад третата врата няма кола.

Пример 4 (Битка със спам)**.** *Напоследък в Интернет пространството се наблюдава натрапване на нежелана информация на потребителите. Спам (spam) е нежелано съобщение, изпратено по електронните комуникации. Най-известната форма на спам е съобщение с рекламно съдържание, изпратено по електронната поща.*

За съжаление засега няма система за борба със спам, която да дава 100% *гаранция. Още по-неприятното е, че ако направим една система за защита твърде "подозрителна", има голям шанс някое важно писмо да отиде при нежеланите.*

Съществуват анти-спам филтри за електронна поща, основаващи се на формулата на Бейс. Тези програми изчисляват вероятността дадено електронно съобщение, което съдържа определени думи, да е спам по следния начин:

$$\mathbb{P}(spam|words) = \frac{\mathbb{P}(spam)\mathbb{P}(words|spam)}{\mathbb{P}(words)}$$

където $\mathbb{P}(spam|words)$ е вероятността дадено съобщение да е спам, при положение че съдържа определени думи или изрази, $\mathbb{P}(words|spam)$ е вероятността тези думи или изрази да се съдържат в едно спам-съобщение, $\mathbb{P}(spam)$ е вероятността едно съобщение да е спам, а $\mathbb{P}(words)$ е вероятността тези думи да бъдат намерени в произволно електронно съобщение, изчислена по формулата за пълната вероятност.

Способностите на филтъра зависят най-вече от капацитета на сървъра, на който е инсталиран. Идеята е предложена за пръв път от английския програмист и предприемач, изобретател на езика за програмиране Lisp, Пол Греъм. Освен това Пол Греъм е и писател с въображение.

4.2 Случайни величини

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

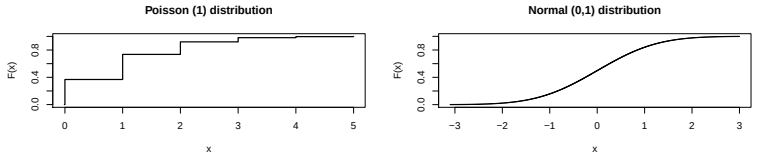
Определение 8. *Ще казваме, че функцията $\xi(.)$, определена в Ω със стойности върху реалната права R^1, е случайна величина, ако $\{\omega : \xi(\omega) < x\}$ е случайно събитие, където $x \in \mathbb{R}^1$ и $\omega \in \Omega$.*

В сила е следната теорема:

Теорема 3. *Линейна комбинация от случайни величини, произведение и измерима функция от случайни величини е случайна величина.*

Казваме, че случайните величини ξ и η са независими, ако за всеки две числа x и y е в сила $\mathbb{P}(\xi < x, \eta < y) = \mathbb{P}(\xi < x)\mathbb{P}(\eta < y)$.

Определение 9. *Функцията $F(x) = \mathbb{P}\{\omega : \xi(\omega) < x\}$ ще наричаме функция на разпределение на случайната величина ξ.*



Фигура 4.1: Графики на функции на разпределение, съответно на дискретна и на непрекъсната случайна величина

Може да се докаже, че функцията $F(x)$ е монотонно ненамаляваща и непрекъсната отляво. Освен това $F(-\infty) = 0$, $F(\infty) = 1$.

В термини на разпределението си случайните величини се класифицират като дискретни и непрекъснати.

Определение 10. *Случайната величина, която приема краен брой или изброимо много стойности $x_1, x_2, x_3, ...$ с вероятности съответно $p_1, p_2, p_3, ...$ се нарича дискретна.*

Естествено $\sum_{i=1}^{\infty} p_i = 1$ и $p_i \geq 0$. Тогава функцията на разпределение има само скокове в точките x_i, навсякъде другаде е константа. В точката x_i скокът й е равен точно на числото p_i.

Определение 11. *Ако съществува функция $f(x) \geq 0$ такава, че за всяко x да е изпълнено $\int_{-\infty}^{\infty} f(x)dx = 1$, то случайната величина наричаме непрекъсната, а функцията $f(x)$ наричаме плътност на случайна величина .*

Тогава, разбира се, $F(x)$ е непрекъсната, т.е. няма никакви скокове, и плътността е производна на функцията на разпределение, т.е. $f(x) = F'(x)$ за почти всяко x.

Графиките на две функции на разпределение, съответно на дискретна и на непрекъсната случайна величина, са дадени на фигура 4.1.

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

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

Вероятностното разпределение представя стойностите, които една случайна величина може да приеме и вероятностите, с които тя приема тези стойности. Казваме, че познаваме една случайна величина, ако знаем нейното вероятностно разпределение.

Ще разгледаме някои числови характеристики на случайните величини, които се пресмятат само по разпределението.

Определение 12. *Математическото очакване (средно) на непрекъснатата случайна величина ξ се пресмята като интеграла $\mathbb{E}\xi = \int xf(x)dx$, а на дискретна случайна величина като сумата $\mathbb{E}\xi = \sum_i x_i p_i$, когато това е възможно.*

Нека a и b са константи, а ξ и η са случайни величини. Тогава лесно се доказва следната теорема:

Теорема 4. *Математическото очакване притежава следните свойства:*

- *линейност:* $\mathbb{E}(a\xi + b\eta) = a\mathbb{E}\xi + b\mathbb{E}\eta$
- *мултипликативност: Ако* $\xi \perp \eta$, *то* $\mathbb{E}\xi\eta = \mathbb{E}\xi\mathbb{E}\eta$.

Определение 13. *Моменти от ред k на случайна величина наричаме следните числови величини (когато съществуват): обикновен момент - $\mathbb{E}\xi^k$, абсолютен - $\mathbb{E}|\xi|^k$, централен - $\mathbb{E}(\xi - \mathbb{E}\xi)^k$.*

Определение 14. *Медианата μ се определя като решение на уравнението $F(\mu) = \frac{1}{2}$.*

Тя описва положението на средата на разпределението върху числовата ос. Когато решението не е единствено, се взема средата на интервала от решения.

Определение 15. *Квантил q_α (квантил с ниво α) на дадено вероятностно разпределение F се определя като решение на уравнението $F(q_\alpha) = \alpha$.*

В статистиката е прието квантилите на вероятности кратни на $\frac{1}{4}$ да се наричат квартили, тези на $\frac{1}{10}$ - децили, а на $\frac{1}{100}$ - процентили. Така $\mu = q_{\frac{1}{2}}$ е втори квартил, пети децил, петдесети процентил.

Определение 16. *Мода се определя като най-вероятното число за дискретни случайни величини, а за непрекъснати - като координатата на максимума на плътността.*

Естествено, разпределенията могат и да не притежават единствена мода. За едномодалните симетрични разпределения, очевидно трите характеристики: мода, медиана и математическо очакване, съвпадат. ???

Определение 17. *Дисперсия на случайна величина ξ се определя като числото $\mathbb{D}\xi = \mathbb{E}(\xi - \mathbb{E}\xi)^2$ (когато съществува).*

Дисперсията е най-важната характеристика за разсейване на стойностите на случайната величина. Дисперсията може да се окаже и безкрайна. Когато дисперсията е безкрайна, за "определяне" на мащаба се използува т.нар. интерквартилен размах.

Определение 18. *Интерквартилен размах r наричаме разликата между третия и първия квартили: $r = q_{\frac{3}{4}} - q_{\frac{1}{4}}$.*

Фактически вместо дисперсията, както в числовите, така и в аналитичните сметки, се използува стандартно отклонение. Това е $\sigma(\xi) = \sqrt{D\xi}$. Тази характеристика се мери в същите физически единици като случайната величина и може да бъде съответно интерпретирана.

Когато искаме да се отървем от размерността, например за да сравним разпределенията на две различни случайни величини, прилагаме т.н. центриране и нормиране. Вместо случайната величина ξ разглеждаме центрираната и нормирана величина $\widetilde{\xi} = \frac{\xi - \mathbb{E}\xi}{\sigma(\xi)}$.

Следващите две характеристики на разпределенията не зависят от мерните единици, с които са отчитани съответните случайни величини, 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 19. *Асиметрия на случайната величина ξ ще наричаме числото (когато съществува):*

$$As(\xi) = \frac{\mathbb{E}(\xi - \mathbb{E}\xi)^3}{\sigma^3(\xi)} = \mathbb{E}\left\{\widetilde{\xi}\right\}^3$$

На фигура 3.1 на страница 16 е дадено сравнение на положително асиметрични плътности с плътността на нормалното разпределение, 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 разпределения с положителна асиметрия те се нареждат в посочения ред, а за тези с отрицателна асиметрия - в обратния. Това правило, разбира се, е вярно само за унимодални разпределения с проста аналитична форма на плътността.

4.3 Някои важни вероятностни разпределения

Нормално разпределение

Нормалното разпределение е било изучено още през 17 век, когато се е наблюдавало, че грешките от измерване имат симетрично и камбановидно разпределение. Неговата математическа формула е получена за първи път през 1733г. от Моавър като граница на биномното разпределение. То е известно още като гаусово, защото Гаус първи го е публикувал през 1809г.

Определение 20. *Казваме, че случайната величина ξ с непрекъснато вероятностно разпределение е нормално разпределена и означаваме $\xi \in N(\mu, \sigma^2)$, ако нейната плътност има вида:*

$$f(x, \mu, \sigma) = \frac{1}{(2\pi)^{1/2}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Математическото очакване на това разпределение е μ, а дисперсията му е σ^2. Стандартно нормално разпределение се нарича разпределението $N(0, 1)$, а 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 и еднакво разпределени случайни величини клони към нормално разпределение.

В сила е следната теорема, която ни позволява от гаусово разпределение с произволни параметри да получим стандартното нормално разпределение:

Теорема 5. *Ако* $\xi \in N(\mu, \sigma^2)$, *то* $\frac{\xi-\mu}{\sigma} \in N(0,1)$.

χ^2 разпределение

Определение 21. *Нека* $\xi_1, \xi_2, ..., \xi_n$ *са независими еднакво стандартно нормално разпределени случайни величини. Случайната величина* $\sum_{i=1}^{n} \xi_i^2$ *има* $\chi^2(n)$ *разпределение с параметър* n, *наречен степен на свобода.*

Средната стойност (математическото очакване) на $\chi^2(n)$ е n, а дисперсията е равна на $2n$.

Разпределение на Фишер

Определение 22. *Разпределение на Фишер (F разпределение) с* r *и* s *степени на свобода има частното* $\frac{\frac{\chi_r^2}{r}}{\frac{\chi_s^2}{s}}$, *където* χ_r^2 *и* χ_s^2 *са независими случайни величини с* χ^2 *разпределение с* r *и* s *степени на свобода, съответно.*

Разпределение на Стюдент

Определение 23. *Разпределение на Стюдент* $T(n)$ *с* n *степени на свобода има частното* $\frac{\xi}{\sqrt{\frac{\chi_n^2}{n}}}$, *където* ξ *и* χ_n^2 *са независими случайни величини,които имат съответно стандартно нормално разпределение и* χ^2 *разпределение с* n *степени на свобода.*

Връзката между T и F разпределенията може да се представи по следния начин: $T(n)^2 = F(1,n)$, т.е. квадратът на случайна величина, която е $T(n)$ разпределена е $F(1,n)$ разпределена случайна величина.

Неравенството за смесените моменти Неравенството за смесените моменти на две случайни величини има вида:

$$\mathbb{E}\xi\eta \leq \left(\mathbb{E}\xi^2\mathbb{E}\eta^2\right)^{\frac{1}{2}}$$

Коефициент на корелация

Определение 24. *Коефициент на корелация на случайните величини ξ и η с крайни втори моменти наричаме числото*

$$R(\xi,\eta)=\frac{\mathbb{E}(\xi-\mathbb{E}\xi)(\eta-\mathbb{E}\eta)}{\sigma(\xi)\sigma(\eta)}=\mathbb{E}\widetilde{\xi}\widetilde{\eta}$$

От това определение е ясно, че при независими случайни величини коефициентът на корелация е нула. Когато вместо ξ и η в неравенството за смесените моменти поставим центрираните и нормирани случайни величини $\widetilde{\xi}$ и $\widetilde{\eta}$, получаваме $|R(\xi,\eta)|\leq 1$.

В сила е и следната теорема:

Теорема 6. *Ако коефициентът на корелация $|R(\xi,\eta)|=1$, то между случайните величини ξ и η съществува линейна връзка, т.е.*

$$\eta=a\xi+b$$

където a и b са константи.

Коефициентът на корелация може да се разглежда като мярка за зависимост между случайните величини, което често се прави на практика.

4.4 Многомерни случайни величини

До сега разглеждахме едномерни случайни величини. Сега ще разгледаме многомерни случайни величини (т.е. случайни вектори). Ще определим понятията многомерна функция на разпределение и многомерна плътност. Ще разгледаме и условни разпределения.

Нека $\xi=(\xi_1,\xi_2,...,\xi_n)$ е случаен вектор.

Определение 25. *Многомерната функция на разпределение на случайната величина $\xi\in\mathbb{R}^n$ се определя така:*

$$F(x_1,x_2,...,x_n)=P\left\{\bigcap_{i=1}^{n}(\xi_i<x_i)\right\}$$

Многомерната функция на разпределение притежава следните очевидни свойства:

1. $F(-\infty, x_2, ..., x_n) = 0$;
2. нормираност: $F(\infty, \infty, ..., \infty) = 1$;
3. монотонност: ако $x_1' < x_1''$, то $F(x_1', x_2, ..., x_n) \leq F(x_1'', x_2, ..., x_n)$;
4. Ако $\xi_1 \perp \xi_2 \perp ... \perp \xi_n$, то $F(x) = \prod_{i=1}^{n} F(x_i)$.
5. Маргиналното разпределение на едномерната случайна величина ξ_1 се възстановява лесно от многомерната функция на разпределение по следния начин $P(\xi_1 < x) = F(x, \infty, ..., \infty)$.

Казваме, че съществува плътност на разпределение $f(x_1, x_2, ..., x_n)$ на случайната величина $\xi = (\xi_1, \xi_2, ...\xi_n)$, ако:

$$F\left(x_1, x_2, ..., x_n\right) = \int_{-\infty}^{x_1} ... \int_{-\infty}^{x_n} f(y_1, y_2, ..., y_n) dy_1 dy_2 ... dy_n$$

Плътността се възстановява от функцията на разпределение:

$$f(x_1, x_2, ..., x_n) = \frac{\partial^n F\left(x_1, x_2, ..., x_n\right)}{\partial x_1 \partial x_2 ... \partial x_n}$$

Плътността притежава следните свойства:

1. неотрицателност: $f(x_1, x_2, ..., x_n) \geq 0$;
2. нормираност: $\int_{R^n} f(x_1, x_2, ..., x_n) dx_1 dx_2 ... dx_n = 1$;
3. Ако случайните величини ξ_i са независими, то

$$f(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f(x_i)$$

4. Маргиналната плътност на случайната величина ξ_1 се възстановява лесно от многомерната плътност:

$$f_{\xi_1}(x) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f(x, y_2, ..., y_n) dy_2 ... dy_n$$

Условни разпределения

Да разгледаме първо две целочислени случайни величини ξ и η. Тяхното съвместно разпределение се задава с таблицата:

$$p_{ij} = \mathbb{P}\left(\xi = i, \eta = j\right)$$

Условно разпределение на случайната величина ξ при условие η ще наричаме разпределението: $\mathbb{P}(\xi = i|\eta = j) = \frac{p_{ij}}{\mathbb{P}(\eta=j)}$, $\mathbb{P}(\eta = j) = \sum_k p_{kj}$.

Условните разпределения могат да се определят само за "действителните" стойности на случайната величина η, т.е. тези с ненулева вероятност.

Нека сега разгледаме две непрекъснати случайни величини със съвместна плътност $f(x, y) > 0$. Тук също се оказва възможно определянето на условни разпределения във формата на плътности.

Многомерно нормално разпределение

Определение 26. *Плътността на стандартното нормално разпределение в* $\mathbb{R}^n$ *има вида:*

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

където $x \in \mathbb{R}^n$.

От определението се вижда, че тази плътност зависи само от нормата на вектора x и следователно е инвариантна относно всякакви ортогонални трансформации защото те запазват нормата. Също така е ясно, че тя може да се представи като произведение на n едномерни стандартни нормални плътности. Всички маргинални разпределения (или проекции в произволна размерност) са нормални. Произволна линейна функция от (зависими или независисми) нормални случайни величини е нормална случайна величина. Вярно е и че условните разпределения (при линейни ограничения от типа на равенството) са гаусови.

4.5 Редици от случайни величини

Разгледахме случайна величина и случаен вектор, който по същество е крайна редица от случайни величини. Сега ще разгледаме безкрайни редици от случайни величини. Ще дефинираме различни видове сходимост. Ще се интересуваме кога, при какви условия, случайностите изчезват.

Определение 27. *Казваме, че редицата от случайни величини $\xi_1, \xi_2, ...$ клони по разпределение към случайната величина ξ, ако редицата от функциите на разпределение F_n клони към F във всяка точка на непрекъснатост на F. Означаваме $\xi_n \xrightarrow{d} \xi$.*

Определение 28. *Казваме, че редицата от случайни величини $\xi_1, \xi_2, ...$ клони по вероятност към случайната величина ξ и означаваме $\xi_n \xrightarrow{p} \xi$, ако $\forall \epsilon > 0$ е в сила $P(|\xi_n - \xi| > \epsilon) \to 0$.*

Определение 29. *Казваме, че редицата от случайни величини $\xi_1, \xi_2, ...$ е сходяща към случайната величина ξ почти сигурно (или с вероятност 1), ако $P(|\xi_n - \xi| \to 0) = 1$. Означаваме $\xi_n \xrightarrow{p=1} \xi$.*

Теорема 7. *Сходимост почти сигурно влече сходимост по вероятност. Сходимост по вероятност влече сходимост по разпределение.*

Закони за големите числа

В своя труд "Изкуство на предположенията"("Ars conjectandi"), публикуван посмъртно през 1713г, Яков Бернули е показал, че при неограничено нарастване на броя на опитите, които се провеждат по схемата на Бернули, относителната честота на настъпването на събитието A клони към $\mathbb{P}(A)$. Бернули доказва това твърдение за сходимост по вероятност. В кореспонденция с Лайбниц Яков Бернули пише: "Законът на големите числа е правило, което дори и най-големият глупак разбира от само себе си, по някакъв природен инстинкт, без предварителни обяснения". По-късно Борел доказва, че това твърдение е вярно и за сходимост с вероятност единица. Фактът, че относителната честота клони към вероятността, прави примамлива идеята вероятността да се дефинира като граница на относителните честоти. Основното предимство на тази дефиниция е, че тя е близка до експеримента.

Ще дадем обобщение на резултатите, получени от Бернули и Борел.

Нека е дадена редица от случайни величини $\xi_1, \xi_2, ...$ Нека с $S_n = \sum_{i=1}^{n} \xi_i$ означим редицата от парциалните суми. Предмет на законите за големите числа са следните задачи:

Слаб закон за големите числа. Интересуваме се при какви условия редицата $\frac{1}{n}(S_n - \mathbb{E}S_n)$ клони по вероятност към 0. Забележете, че разглеждаме слаба сходимост – тази по вероятност.

Определение 30. *Когато горната сходимост е изпълнена за дадена редица, казваме, че за тази редица е в сила слаб закон за големите числа.*

В сила е следната теорема:

Теорема 8. *(Теорема на Марков) Нека $\xi_1, \xi_2, \ldots$ е редица от случайни величини. Ако $\frac{1}{n^2}\mathbb{D}\sum_{i=1}^{n}\xi_i \to 0$ когато $n \to \infty$, то за тази редица е в сила Слаб закон за големите числа.*

Следствие:

Теорема 9. *Ако $\xi_1, \xi_2, \ldots$ е редица от еднакво разпределени случайни величини, то $\frac{1}{n}\sum_{i=1}^{n}\xi_i \xrightarrow{p} a$, където $a = \mathbb{E}\xi_1$.*

Последните две твърдения позволяват следната интерпретация: Да предположим, че с даден уред многократно измерваме някаква физична величина, като всяко от измерванията извършваме независимо от останалите. Нека измерваната величина е величината a. Измерванията можем да разглеждаме като независими случайни величини $\xi_1, \xi_2, \ldots$ с едно и също вероятностно разпределение, понеже уредът е един и същ. Приемаме, че $\mathbb{E}\xi_1 = a$ и $\mathbb{D}\xi_1 = \sigma^2$. При посочените условия вероятността случайната величина $\frac{1}{n}\sum_{i=1}^{n}\xi_i$ да се окаже на произволно разстояние от измерваната величина a за голям брой опити е близка до единица. В този смисъл $\frac{1}{n}\sum_{i=1}^{n}\xi_i$ е оценка на неизвестния параметър a. Поради този факт теоремата (по-точно следствието) стои в основата на теорията за оценяване на параметри, представена в ???. Но тези условия могат да бъдат отслабени, като се предположи, че в резултат на износване точността на уреда от измерване към измерване може да започне да намалява, което означава, че дисперсията ще расте. Това обаче не бива да става прекалено бързо. Дисперсията $\mathbb{D}\xi_i$ може да расте, но все пак ще искаме $\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{D}\xi_i \to 0$ когато $n \to \infty$. Когато случайните величини са произволни, този факт е строго изразен в теоремата на Марков.

Теорема 10. *(Теорема на Хинчин) Ако $\xi_1, \xi_2, \ldots$ е редица от независими и еднакво разпределени случайни величини с крайно математическо очакване, то за тази редица е в сила слабия закон за големите числа.*

Теоремата на Хинчин показва, че когато случайните величини имат едно и също вероятностно разпределение, изискването за съществуване на дисперсия се оказва излишно.

Усилен закон за големите числа. Разглеждаме сходимост почти сигурно. Интересуваме се при какви условия редицата $\frac{1}{n}(S_n - \mathbb{E}S_n)$ е сходяща почти сигурно към 0.

Определение 31. *Когато горната сходимост е изпълнена за дадена редица от случайни величини, казваме, че за тази редица е в сила Усилен закон за големите числа.*

В сила е следната теорема:

Теорема 11. *(Теорема на Колмогоров)Разглеждаме редица от независими и еднакво разпределени случайни величини $\xi_1, \xi_2, ...$ Тогава за тази редица е в сила Усилен закон за големите числа тогава и само тогава, когато е ограничен първият абсолютен момент, т.е. $\mathbb{E}|\xi_i| < \infty$.*

Смисълът на закона за големите числа е, че средното аритметично на n случайни величини с нарастване на броя им клони към величина, която не е случайна. Случайностите, при определени условия, взаимно се погасяват, компенсират се. Ценността на законите за големите числа е ясното и точно формулиране на условията, при които случайностите се компенсират.

Централна гранична теорема. Интересуваме се от въпроса: *При какви условия, центрирана и нормирана, тази редица клони по разпределение към гаусова случайна величина?*

Определение 32. *Когато тази сходимост е изпълнена за дадена редица от случайни величини, казваме, че за тази редица е в сила Централната гранична теорема.*

Разглеждаме редица от независими и еднакво разпределени случайни величини $\xi_1, \xi_2, ...$ с крайна дисперсия σ^2 и математическо очакване μ. Числовите характеристики за парциалната сума са $\mathbb{E}S_n = n\mu$ и $\mathbb{D}S_n = n\sigma^2$. Получаваме твърдение, което намира огромно приложение в статистиката.

Теорема 12. *Ако $\xi_1, \xi_2, ...$ е редица от независими и еднакво разпределени случайни величини с крайна дисперсия, то за тази редица е в сила Централната гранична теорема.*

Тълкуването на тази теорема е следното: За фиксирано цяло положително число n случайната величина $\frac{S_n - n\mu}{\sigma\sqrt{n}}$ има приблизително

стандартно нормално разпределение. В приложенията можем да използваме приближената стандартна нормална плътност, вместо точната.

Глава 5

Статистиката - наука за откриване на знания в данните

5.1 Основни понятия

Стохастичното моделиране е мощно средство за решаване на задачи от реалния свят. Швейцарският математик от холандски произход Яков Бернули (1654 – 1705) употребява гръцката дума стохастика като синоним на ars conjectardi – изкуството да отгатваме, да предвиждаме. Той казва, че човек знае нещо, ако му е известно каква е вероятността то вече да е станало или да се случи след време. В съвременния език стохастичен се възприема като синоним на случаен, а стохастика като събирателно за научните дисциплини, изучаващи модели със стохастични елементи. В англоезичната литература със същото значение се използва терминът статистика като събирателно за вероятностните методи и анализ на данни.

И така, данните (наблюденията) носят информация. В тази глава ще видим как от данните с помощта на средствата, предоставени ни от описателната статистика, може да се получи нова информация за изследваното явление. От данните с помощта на средствата, предоставени ни от математическата статистика, могат да се получат и нови знания за явлението, което тези данни представят.

В тази глава от книгата ще разгледаме основните термини и ме-

тоди на статистиката и ще се опитаме да покажем мястото на математиката в нея.

Съвременното разбиране на понятието "статистика" е твърде широко. То включва в себе си както методи на просто преброяване и визуализиране на информацията, така и методи за взимане на решения, основани на строги математически разсъждения. Освен това в разговорната реч с думата "статистика" често означаваме и събраната информация - например статистика за футбола. Основно понятие в науката статистика е понятието популация (генерална съвкупност).

Определение 33. *Популация наричаме множеството от обектите на изследването.*

За едно изследване на Националния статистически институт това могат да бъдат: всички държавни учреждения в България, домакинствата в планинските райони, семействата без деца, всички жители на страната и т.н. За комисията, изследваща качеството на обучението в университета, това са студентите от университета.

Изчерпателни данни наричаме данни, които напълно представят дадено явление. Такива са например данните получени при едно преброяване на населението на страната от Националния статистически институт. Изчерпателни данни можем да получим когато изучаваме малобройна популация – един клас от ученици, една специалност в университета. За съжаление такива данни често не са достъпни, или пък струват прекалено скъпо. Когато не е възможно изследване на всички единици от популацията или изчерпателните данни за интересуващото ни явление не са достъпни, популацията става абстрактно множество от обекти, представляващо цел на нашето изследване. За да постигнем тази цел, да изследваме популацията, работим с извадки.

Извадката е част от популацията и търсените характеристики на популацията се оценяват по данните от извадката.

Следователно основна цел на статистиката е по даден непълен обем данни да направи някакво правдоподобно заключение за популацията като цяло. Този набор от обекти, които всъщност се измерват, се нарича извадка.

Определение 34. *Подмножеството от обекти на популацията, достъпно за наблюдение (за измерване), наричаме извадка.*

Получаването на достоверни изводи за генералната съвкупност по информация от извадката е възможно само ако извадката се състои от типичните данни на популацията, т.е. ако тя съдържа приблизително всички особености на популацията и възпроизвежда нейната структура. Това свойство на извадката се нарича представителност (репрезентативност).

Получаването на представителна извадка може да бъде направено по различни начини. Най-често това се осигурява чрез случаен избор на нейните елементи, и така построената извадка се нарича случайна извадка. При нея се предполага, че шансът на всеки обект от генералната съвкупност да попадне в извадката е равен. Това означава, че всички обекти са равноправни и изборът е напълно случаен. Казваме, че всеки обект от популацията с една и съща вероятност може да попадне в извадката.

Нарушаването на принципа за случаен избор може да доведе до сериозни грешки. Знаменито по своя неуспех е проучването, проведено от американското списание "Литературен преглед", относно изхода на предстоящите президентски избори през 1936 г. Кандидатите били Рузвелт и Ландън. Като избрала по случаен начин 4 милиона адреси от телефонните книги, редакцията на списанието изпратила писма с допитване за президент. Когато получените писма с отговори били обработени, списанието обявило, че за президент ще бъде избран Ландън. Както вече знаем, резултатът от изборите бил противоположен на тази прогноза. Какви са допуснатите грешки? Основната грешка е, че като се използват телефонните указатели не може да се състави репрезентативна извадка от населението на страната поради факта, че абонатите са главно глави на семейства. Освен че през 1936 година това са били главно мъже, по това време само заможните семейства са имали телефон. Това естествено е довело до това, че в извадката, която списанието проучла, са били включени само относително богати хора. Респективно отговор на питането са изпратили не всички анкетирани, а предимно тези, които са привикнали да отговарят на писма, т.е. представителите на деловия свят, за които е било известно, че подкрепят Ландън. Да отбележим, че в социологическите изследвания често се наблюдава изместване на извадката, което се дължи на голям брой откази на отговор. В следствие на тези грешки огромна част от населението не била изследвана и това довело до грешния резултат.

За разлика от злополучното изследване, по същото време социолозите Галъп и Роупър правилно предсказали победата на Рузвелт, като при това използвали само 4 хиляди анкети!

Причина за техния успех е правилното съставяне на извадката. Те отчели факта, че обществото се разделя на социални групи (наречени слоеве или страти), които вътрешно са сравнително еднородни по отношението си към кандидатите за президенти, но между отделните слоеве може да има големи разлики. Затова и случайната извадка от всеки отделен слой е могла да бъде малобройна. Но броят на избраните от всяка страта елементи трябва да се отнася към броя на всички елементи в извадката така, както се отнася обемът на стратата към обема на генералната съвкупност.

Днес използването на такава извадка, наречена стратифицирана извадка, е общоприет подход. Когато извадката не представя цялата генерална съвкупност, а само някакъв неин слой, говорим за изместване на извадката и в този случай имаме изместена извадка, която не е представителна. Изместването на извадката е един от основните източници на грешки при използване на статистическите методи. Вземането на по-голяма извадка в този случай не подобрява качеството на извадката.

Планиране на експеримента В научните изследвания експериментът е основно средство за изучаване на техническите обекти и природните явления. Дълго време обаче организирането на самия експеримент т. е. кога, къде и как да се провеждат наблюденията, се е извършвало интуитивно според наличния опит на експериментатора. В началото на нашия век Р. Фишер достига до извода, че методите за обработка на експерименталните данни в много случаи могат да се окажат практически безполезни. Ако експериментът е поставен и проведен лошо, това не може да се поправи дори и с най-съвършените методи за обработка на опитните данни.

В резултат на разработените от Фишер методи за провеждане, организиране и анализ на експеримента възниква планирането на експеримента като ново научно направление. Основен предмет на изследване в теорията на планиране на експеримента е активният експеримент. Активен експеримент е този, при който на входа на изследвания многофакторен обект въздействат само управляеми фактори. Как да намалим максимално разходите за, и без това, скъпия и продължителен опит? На това ни учи планирането на експеримента,

което е част от науката статистика.

Числови и нечислови данни Информацията, която представляват данните, обикновено се различава по това как се записва. Понякога това са числа: размери, тегла, бройки и т.н. Друг път това са нечислови характеристики като цвят, форма, вид химическо вещество, вид тор 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.

Основни скали на измерване Данните могат да бъдат класифицирани в една от следните четири скали:

- номинална. При тази скала може само да се установи различие между обектите. Примери: наблюденията над променливата пол са или мъжки, или женски, и могат да се кодират с 0 и 1 (или M и F); видът на хранителните стоки е тестени, млечни и месни и може да се кодира с 1, 2 и 3 (или a, b, c).

- ординална (наредена). Обектите могат само да се наредят. Пример: наблюденията над променливата оценка от изпит са слаб, среден, добър, много добър и отличен, които се кодират съответно с числата 2, 3, 4, 5 и 6. Фактът, че числото 6 е два пъти по-голямо от числото 3 не означава, че отличният успех е два пъти по-висок от средния.

  Да вмъкнем тук следната забележка: Ординалната и номиналната скала предоставят по-ограничени средства за анализ на данните. Тук е мястото да споменеем и дихотомните (двоичните, алтернативните) данни. Дихотомните данни могат да бъдат или наредени или ненаредени, но винаги приемат точно две стойности, които могат да бъдат кодирани с 0 и 1, или като неуспех и успех.

- относителна. Възможно е да се направят изводи относно абсолютните и относителните различия. Пример: Ако 5 лв. и 10 лв. са цените на две стоки, то можем да твърдим, че цената на втората стока е с 5 лева по-висока и два пъти по-голяма от цената на първата. Относителната скала на измерване има фиксирано начало (“0”).

- интервална. Възможно е да се правят изводи относно относителните различия. Ако температурата на две различни места е 5^oC и 10^oC , това означава, че на второто място е с 5^o по-топло, но не означава, че е 2 пъти по-топло. Тук началото е избрано произволно.

Относителната и интервалната скала се наричат силни скали, защото обикновено носят повече информация. Те позволяват използването на по-силни статистически методи и следователно получаването на по-силни резултати.

Категорни данни Категорийни променливи са тези, които разбиват популацията на краен брой подпопулации (категории). Например променливата пол разбива популацията на две подпопулации. Информацията, която ни носят данните в този случай, е броя на обектите в едната подпопулация (от мъжки пол) и броя на обектите в другата (от женски пол). Когато разглеждаме нечислов признак на един случайно избран обект от генералната съвкупност, то той съгласно предположенията ни за равнопоставеност на обектите в извадката би трябвало да попадне в дадена категория с вероятност равна на пропорцията на обектите в тази категория от генералната съвкупност.

Описателна статистика Описателната статистика ни дава средства за визуализиране и обобщаване на данните, за получаване на първоначална представа за изследваното явление.

Математическа статистика Основна цел на математическата статистика е изграждането на полезни математически модели с помощта на теорията на вероятностите.

Освен това статистиката дава средства за проверката на моделите върху реални данни, както и за интерпретация на резултатите от моделите.

В математическата статистика винаги се разглежда следният вероятностен модел:

Данните се състоят от няколко наблюдения. За всяко наблюдение от извадката се предполага математически модел, който е случайна величина с нейното вероятностно разпределение (виж параграф ???). Ако наблюденията са еднотипни и независими едно от друго, то за модел се използват независими и еднакво разпределени случайни ве-

личини. Когато независимостта е съмнителна, трябва да се изследва съвместното вероятностно разпределение на тези случайни величини (виж параграф ???). Въз основа на така направените предположения се изследват вероятностните свойства на различни удобни за анализ и взимане на решения зависимости в наблюденията. Тези зависимости представляват "полезни" за нас функции от наблюденията, например техните моменти - математическото очакване, дисперсията и т.н. Анализирайки ги ние получаваме нова информация. В науката Статистика е прието функции от наблюденията се наричат статистики. Когато в тези функции се поставят стойностите на реалните наблюдения, се получават конкретни числови стойности, въз основа на които се правят заключения. Наричаме ги статистически изводи, и можем да оценим колко вероятни са те при положение че модела ни е верен.

Основните въпроси тук са:

- Какви са параметрите на модела?
- Доколко моделът е правдоподобен (адекватен)?
- Как да използваме получения модел за да прогнозираме?

Когато моделът се окаже неадекватен, се строи друг. И това е стандартния подход в науката при прилагане на подхода, известен като Математическо моделиране. Основния момент е как може да се използва полученият модел за прогнозиране на бъдещи събития.

5.2 Откриване на информация в данните

В ерата на компютърните технологии, ние живеем в море от данни. Естественият проблем, който възниква, е да извлечем информация за интересуващото ни явление от масивите с думи и числа, съхранени в компютъра или достъпни в Интернет.

Визуализацията на данните е основна стъпка в процеса на откриване на нова информация от данните. Тези картини (графики) ни дават възможност да получим информация за явлението което изследваме. Такива картини са се съставяли на ръка от над 200 години.

Компютрите ни дават уникалната възможност да представим информацията в графики. Освен способността да визуализират огром-

ни по обем данни, те ни дават възможност и да раздвижим и "съживим" тези графики. С компютърна визуализация можем да видим данните от различни гледни точки, под различни ъгли, което ни помага да сравним различните изгледи, да поставим въпроси и бързо да получим отговори, да открием образци (шаблони) и интересни особености. Това е изключително полезно при проучване на данните и при строенето на модела на изследваното явление.

Както вече споменахме, множеството от данни не е просто набор от числа. То има своя вътрешна структура и често целта на анализа на тези данни е да извлечем знания за тази структура.

Често, но далеч не винаги, данните са числови. Обобщените числови характеристики на данните, като средно, дисперсия, и др., могат да бъдат много полезни и се използват в съвременните подходи за моделиране, но някои особености в данните могат да се забележат само от подходящото графично представяне (визуализиране) на данните.

Както вече подчертахме, построяването на математически модел, който добре описва данните, е мощен подход в разбирането какво точно ни казват данните и прави възможно прогнозирането на наблюдаваното явление. Но ако не направим подходящи предположения, няма да можем да построим адекватен математически модел.

Ако предположенията ни са верни, колкото те са по-силни, толкова по-надеждни са методите, които можем да прилагаме, и съответно по-точни са резултатите, които ще получим. Но ако предположенията са грешни, тогава и моделът ще бъде неадекватен.

Една част от графичните методи са развити с цел да се проверят доколкото е възможно предположенията на които градим модела (например независимост на наблюденията). Пред всеки статистик, който изследва явлението стои дилемата в каква степен да позволи на данните да влияят на избора на модел, и в каква степен ще разчита на известни вече факти за физическата същност на явлението.

Друга част от графичните методи са развити с цел да се провери доколко вече построеният модел е адекватен (правилен).

Следователно процесът на построяване на модел, описващ данните, започва и завършва с графичен анализ на данни. Докато в началната стъпка на моделирането графично се представят оригиналните данни, в крайната стъпка графично се визуализират разликите на

оригиналните данни и прогнозираните стойности.

Пример 5 (Бакшишът в заведението)**.** *Бакшиш в заведение, в което се сервират храна и напитки, е възнаграждение за обслужващия сервитьор, което е допълнително „дадено на ръка“ над цената на консумацията. В различните култури на бакшишите се гледа по различен начин. Някои ги смятат за задължителни, други – за обидни. В Швейцария, Холандия и Австрия размерът на бакшиша е 3-5% от сумата. В Скандинавските страни и Италия бакшишът се включва в сметката и варира около 7-10%. Във Франция бакшишът се включва в сметката, но е прието да се даде и допълнително дребна сума. В Гърция бакшишът представлява задължителни 10% от сметката. Японските сервитьори възприемат даването на бакшиш за оскърбление, тъй като те смятат, че са длъжни да си вършат работата добре по подразбиране. В Китай даването на бакшиш официално е забранено. В България бакшишът се възприема като награда за добро обслужване. У нас никой не би трябвало да се обиди както от това, че си дал бакшиш, така и от това, че не си дал. Заведенията в САЩ имат регламентиран минимален процент възнаграждение за обслужващия персонал. Очаква се клиентът да плати допълнително такса за сервиране, което е регламентирано в меморандума на Националната асоциация на ресторантьорите в САЩ.*

Сервиторите във всеки ресторант (в САЩ ???) се интересуват от следните въпроси:

- *Колко е очакванията бакшиш?*
- *От какво зависи размерът на възнаграждението за обслужващия сервитьор?*
- *Кои клиенти са склонни по-щедро да покажат, че са доволни от обслужването – жените или мъжете, пушачите или непушачите?*

Един сервитьор е записал сумите, които е получавал, и други наблюдения, за които си мисли, че биха му били от полза, в продължение на няколко месеца докато е работил в един ресторант. Каква информация дават тези данни на сервитьора?

Променлива	Смисъл, който променливата носи
obs	Номер на наблюдението
totbill	Сметката в американски долари
tip	Паричното възнаграждение, дадено на сервитьора от посетителите на ресторанта (в американски долари)
sex	Пол на човека, който плаща сметката (0-мъж, 1-жена)
smoker	Пушач (0-Не, 1-Да)
day	Ден от седмицата (3=Четвъртък, 4=Петък, 5=Събота, 6=Неделя)
time	Време (0- през деня, 1- през нощта)
size	Размер на празненството

Таблица 5.1: Таблица на променливите

Графичният анализ на данни е съвкупност от техники за визуализиране на данните, чиято цел е да позволят на данните да разкажат за себе си и за явлението, което сме наблюдавали. Добрите статистици винаги се възползват от тази възможност и строят графики, които са полезни и могат лесно да се интерпретират.

Графичният анализ на данни може да подскаже подходящия метод за формален анализ. Освен това графичният анализ на данни може да отхвърли използвания метод за формален анализ. Нещо повече, графичният анализ на данни може да покаже нови особености в данните, да даде допълнителна информация, да посочи нови идеи или да начертае нов път за провеждане на изследването. Добрата графика, както добрата книга, дава полезна информация и ясни, точни и ефективни идеи.

Хистограмата е основно графично средство за визуализиране на честотното разпределение на едно множество от данни. Лицето на всяко правоъгълниче на хистограмата е правопропорционално на броя на наблюденията, чиито стойности лежат в интервала, където е разположена широчината на правоъгълника. Симетрична хистограма с един "връх" повдига хипотезата за нормално (гаусово) разпределение на данните (т.е. на наблюдаваната случайна величина).

Променливите, които са наблюдавани, са описани в Таблица 5.1.

Ще пристъпим към решаване на проблема стъпка по стъпка.

Стъпка 1. Формализиране на реалния проблем:

Най-важните въпроси, на които търсим отговор, са: *Кои са факторите, които влияят на размера на бакшиша? Има ли връзка между променливата tip и другите променливи?*

Стъпка 2. Данни [1]

Броят на наблюденията е 244. Броят на променливите е 8, от които 5 са категорийни променливи. Очевидно първата променлива (obs) няма да бъде изследвана. Числовите променливи са *totbill* и *tip*.

Стъпка 3. Описателна статистика

Ние ще изследваме зависимости, в които могат да са включени повече от три променливи. В този смисъл данните са многомерни и не могат да се визуализират в реалното тримерно пространство. От най-голям интерес е променливата *tip* и затова ще я изследваме първа. Започваме, начертаваме хистограмата й.

```
> hist (tip, breaks = 12, col="lightblue", border="pink")
```

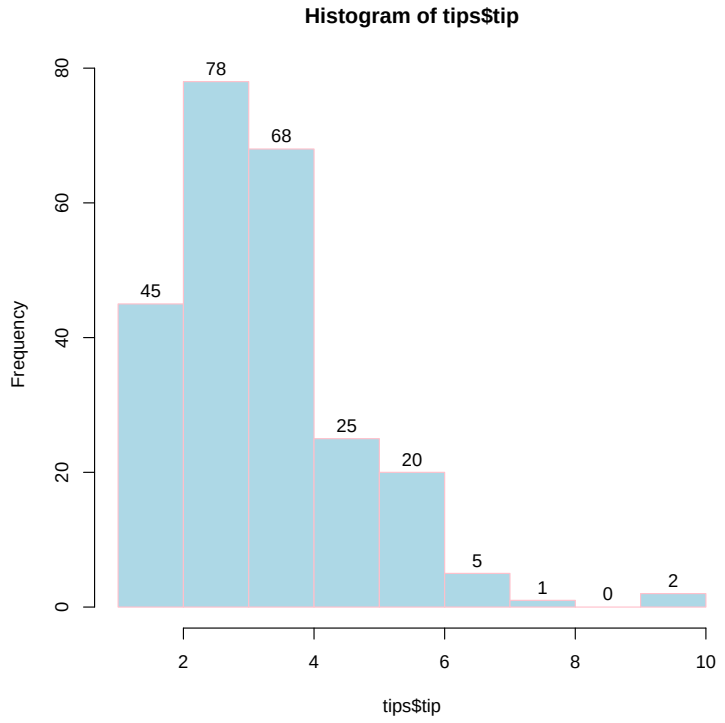
Очевидно разпределението на данните е едномодално (?дефинирано ли е някъде?) (т.е. хистограмата има само един връх). От хистограмата на фигура 5.1 се вижда, че най-често бакшишите са от 2 до 3 долара. Броят на големите бакшиши бързо намалява, което предполага, че този ресторант не е много скъп.

Сега ще начертаем нова хистограма на същите данни, но този път със 101 точки на прекъсване. Така постигаме по-детайлно изображение на случайната величина, като групирането е на по 10 цента.

```
> hist (tip, breaks = 101, col="lightblue", border="pink",
+ labels = FALSE, right = FALSE)
```

Сега разпределението е многомодално с големи върхове при точните суми (2, 3, 4, 5 долара) и по-малки при половинките (1.5, 2.5, 3.5 долара). Това показва, че клиентите обикновено закръглят сумата за бакшиш - по-често до близкото цяло число и по-рядко до

[1] Файлът с данните tips.csv се намира на интернет страницата *http://www.ggobi.org/book/* . Оригиналният източник е Bryant, P. G. and Smith, M. A. (1995), Practical Data Analysis: Case Studies in Business Statistics, Richard D. Irwin Publishing, Homewood, IL.



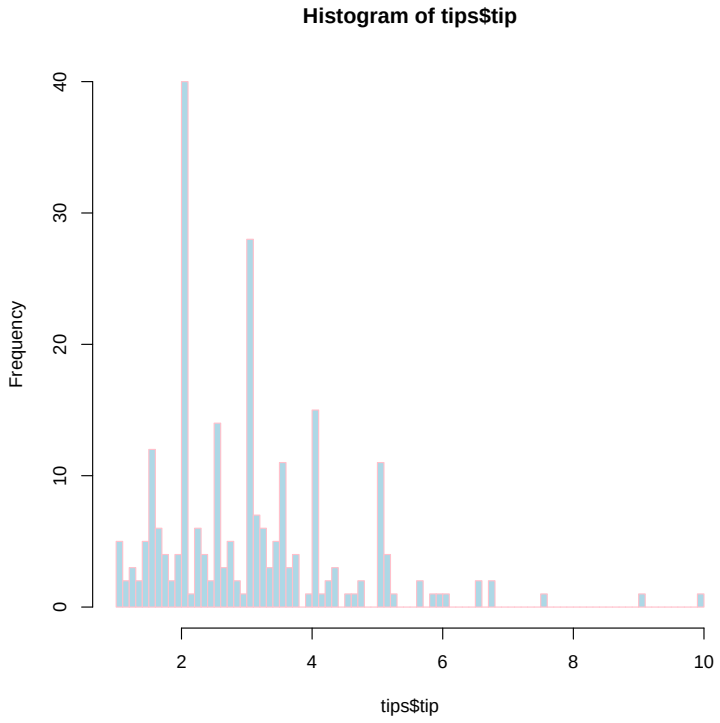
Фигура 5.1: Хистограма с 11 стълба

половин долар.

Като експериментираме с хистограми, начертани с различен брой интервали, забелязваме, че заключенията, направени от хистограмата, силно зависят от броя на интервалите, т.е. от параметър на функцията за начертаване, а не само от данните. Като експериментираме с хистограми, начертани от различни начални точки, забелязваме, че видът на хистограмата зависи и от началната точка.

Причината за това е, че групирането на данните се извършва по различни критерии – например групирайки бакшишите в интервали от 0.00 до 0.10 долара е различно спрямо интервали от вида 0.01 до 0.11 долара, независимо че размера е 10 цента и в двата случая.

До сега не сме дали отговор на основния въпрос, а той е: *Има ли връзка между променливата tip и другите променливи?* Тъй като



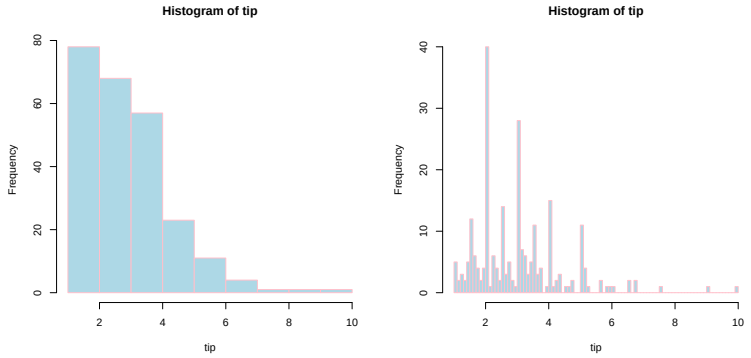
Фигура 5.2: Хистограма със 100 стълба

бакшишът обикновенно се пресмята на базата на сумата от сметката за консумираните ястия, естественно е да предположим връзка между тези двете променливи *tip* и *totbill*, като първо погледнем графиката им.

Графики, изобразяващи едновременно две променливи, се чертаят с командата *plot*().

```
> plot (totbill , tip)
```

С този вид графика получаваме информация за разпръснатостта на данните и за някои основни зависимости между променливите. От получената графика на фигура 5.4 се вижда, че когато сметката расте, нараства и платения бакшиш. (?да оградиш или да покажеш как точно “се вижда”на графиката?)



Фигура 5.3: Хистограми с 11 и със 100 стълба

Основна цел на визуализирането на данни е да се уловят зависимостите между много променливи: две, три, четири и дори повече. Например нека се опитаме да разберем как полът на плащащият и дали е пушач влияят на връзката между размера на сметката и размера на бакшиша, т.е как категорийните променливи *sex* и *smoker* влияят на връзката между променливите *tip* и *totbill*.

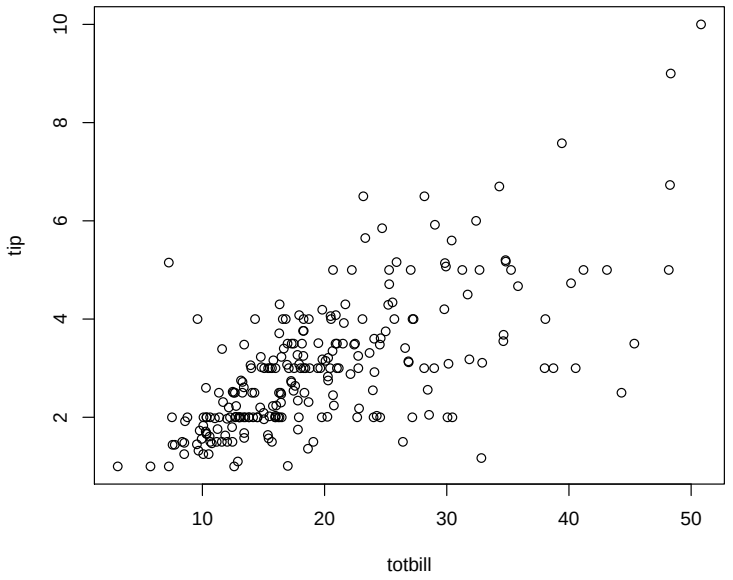
Променливата *sex* е категорийна с две нива (M и F) и е естествено да разделим данните на две групи, в зависимост на пола на този, който плаща сметката. Задаваме си въпроса дали полът има влияние на размера на бакшиша.

Променливата *smoker* също приема две стойности. По този начин разбваме множеството от данни на четири подмножества и чертаем четири графики – по една за всеки от случаите. От тях двете графики на непушачите показват линейна зависимост между изучаваните променливи.

Интересуваме се още и от силата на връзката между променливите *tip* и *totbill* във всяко подмножество. Силата на линейната зависимост между две количествени променливи може да се измери с корелационния коефициент, както споменахме в (?къде? - линейна регресия).

Пресмятаме извадковите корелационни коефициенти за всяко подмножество от данни и ги отпечатваме върху графиката.

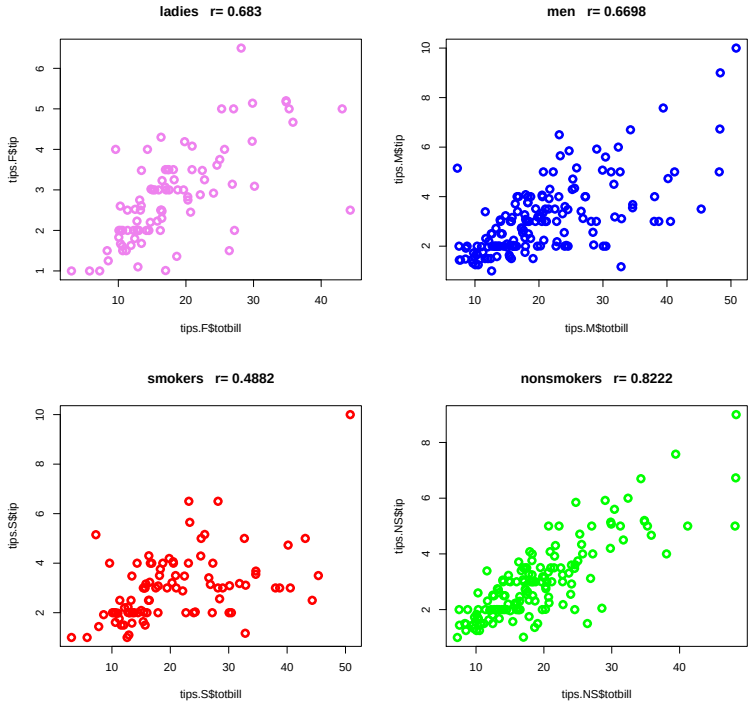
```
> tips.F <- tips [tips$sex=='F',]
```



Фигура 5.4: Данни

```
>  tips.M <- tips [tips$sex=='M',]
>  tips.S <- tips [tips$smoker=='Yes',]
>  tips.NS <- tips [tips$smoker=='No',]
>  op <- par(mfrow=c(2, 2))
>  title <-   paste ( "ladies   r=",
+ signif(cor (tips.F$totbill, tips.F$tip), digits = 4))
>  plot (tips.F$totbill, tips.F$tip, col = "violet",
+ cex = .5, lwd=8, main=title )
>  title <-   paste ( "men   r=",
+ signif(cor (tips.M $totbill, tips.M $tip), digits = 4))
>  plot (tips.M$totbill, tips.M$tip, col = "blue",
+ cex = .5, lwd=8, main=title)
>  title <-   paste ( "smokers   r=",
+ signif(cor (tips.S$totbill, tips.S$tip), digits = 4))
>  plot (tips.S$totbill, tips.S$tip, col = "red",
+ cex = .5, lwd=8, main=title)
>  title <-   paste ( "nonsmokers   r=",
+ signif(cor (tips.NS$totbill, tips.NS$tip), digits = 4))
```

```
> plot(tips.NS$totbill, tips.NS$tip, col = "green",
+ cex = .5, lwd=8, main= title)
> par(op)
```



Фигура 5.5: Данни

От получените графики, показани във фигура 5.5, получаваме следната информация:

За масите с пушачи се наблюдава по-слаба връзка между размера на платения бакшиш и размера на платената сметка. Забелязваме например, че жена непушач, платила сметка в размер 44,3 долара е платила бакшиш 2.5 долара, което е 5% от сметката, и друга, платила сметка 9.6 долара и бакшиш 4 долара (40% от сметката). Също, когато жените непушачи плащат сметката, бакшишът силно зависи от цената на консумираното. Мъжете като цяло са склонни да плащат по-големи бакшиши.

Така показахме как използвайки само елементарни графични сред-

ства, ние освен че получихме информация за една променлива, показахме и зависимости между две, три и четири променливи.

Графика, подходяща за изследване на една променлива е хистограмата. Видяхме, че формата на хистограмата зависи от началната точка и от дължината на интервалите (т.е. дължината на страните на правоъгълниците). Едната хистограма може да визуализира симетрично разпределение, а другата, построена за същите данни – асиметрично, както това се вижда от фигура 5.3.

Този недостатък може да се преодолее с метода, известен като *Усреднено изместената хистограма ASH (average shifted histogram)* . В този метод няколко хистограми са генерирани като се използва една и съща дължина на интервалите (т.е. една и съща широчина на правоъгълниците) и различни начални точки за построяване на хистограмите. Пресметнати са средните честоти и са визуализирани с точки. Този алгоритъм има два параметъра – броя на интервалите (или което е същото, броя на правоъгълниците) който определя широчината на интервалите, и брой на хистограмите, които ще бъдат генерирани. Резултатът от този алгоритъм е една гладка хистограма, визуализирана с точки. (как да вземем графиките от друга система - rggobi???)

Построяването на променящи се във времето графики върху екрана на компютъра е част и от процеса на проучване на данните, и от процеса на построяване на модел на изследваното явление.

Динамични графики са реализирани в пакета *rggobi*. С пакета *ggobi* на *R* могат да се визуализират многомерни данни. Човек може да наблюдава данните в тримерното пространство. За съжаление е трудно дори да си представим визуално четиримерното пространство, да не говорим за пространство с по-голяма размерност. Пакетът rggobi позволява визуализация на многомерни дани (с размерност по-голяма от 3), и визуализирането им в едномерното, двумерното или тримерното пространство, като ги върти и ни дава възможност да ги погледнем от подходящ ъгъл. Други инструменти за визуализиране на данните са представени в книгата Interactive and Dynamic Graphics for Data Analysis: With Examples Using R and GGobi от Dianne Cook и Deborah F. Swayne и на сайта *http://www.ggobi.org/book/* .

С графичните средства за представяне на данните търсим:

- Съмнителни наблюдения. Това са рязко отклоняващите се наблюдения. Те могат да се дължат на грешки при записване на данните. Понякога съмнителните наблюдения показват, че някои от предположенията на модела са нарушени.

- Асиметричност на разпределението. Ако стойностите, близки до минималната стойност, са разположени близко една до друга, а големите стойности са разпръснати една от друга, и всички стойности са положителни, тогава логаритмичната трансформация на данните често помага да получим данни със симетрично разпределение.

- Променливост в разпръскването на данните. Много статистически модели зависят от предположението, че разпръскването на данните в групите е еднакво (т.е. груповите дисперсии са равни). Когато разпръскването нараства когато стойностите на данните нарастват, данните могат да се логаритмуват и често тази трансформация на данните помага да премахнем променливостта в разпръскването на данните.

- Клъстери (групиране на данните). Групите в графиката на разсейване предполагат структура в данните, която вероятно е неочаквана (или очаквана). Когато започнем да строим модел, описващ данните, тази структура трябва да бъде взета под внимание.

- Нелинейност. Не бива да се опитваме да строим линеен модел когато данните показват нелинейност.

Глава 6

Оценки

При моделиранеието на явления с апарата на теорията на вероятностите обикновено имаме някакви теоретични съображения за вида на разпределението на случайната величина, но все пак остават някакви свободни параметри, които трябва да оценим за да получим точното разпределение.

Например в класическите модели за движението на цената на акция се предполага, че дневните изменения на цената представляват случайна величина, която е нормално (гаусово) разпределена, но параметрите на това гаусово разпределение - μ и σ^4 - не могат да бъдат изведени от теоретични съображения. Това очевидно е критично важно за икономиста, който иска да открие подходящата акция за портфйл. Единственото, което може да се направи, е те да бъдат "оценени" по някакъв начин.

6.1 Точкови оценки

Ще представим някои от основните елементи на теорията на точковите оценки.

Ще предполагаме, че моделът на наблюденията са случайните величини $X_1, X_2, \ldots, X_n$ с многомерна плътност на разпределение $f(x, \theta)$ (виж глава **??**), за която знаем всичко, освен един неизвестен (едномерен или многомерен) параметър θ. Пример за това е нормално разпределена случайна величина, за която знаем дисперсията σ^2, но не знаем средното μ.

Точкова оценка на неизвестния параметър се нарича всяка функция от наблюденията, която имаме основание да приемем като стойност на неизвестния параметър. Точковите оценки ще означаваме с $\widehat{\theta} = \widehat{\theta}(x_1, x_2, \ldots, x_n)$.

Определение 35. *(??? някак си да вкарам горните неща вътре ... ???) Казваме, че статистиката $\widehat{\theta}$ е оценка (приближена стойност) на неизвестния параметър θ, ако не зависи от стойността на параметъра θ.*

Но x_i е стойност на наблюдение на случайна величина, която можем да моделираме като такава със същото разпределение като X, но независима от нея (??? тук да сложа референция може би към основния модел? Да се дефинира там X, X_i, x_i... ???). Така че можем да разглеждаме $\widehat{\theta} = \widehat{\theta}(X_1, X_2, \ldots, X_n)$ като функция на случайни величини, следователно самата тя е случайна величина. Затова ще изследваме свойствата и поведението на оценките, използвайки апарата на теорията на вероятностите.

Вече знаем, че в математическата статистика всяка функция от случайни величини се нарича статистика. (??? това нз как да го вържа ???)

Стандартно ще означаваме с $\mathbb{E}\widehat{\theta}$ математическото очакване на случайната величина $\widehat{\theta}$, а с $\mathbb{D}\widehat{\theta}$ нейната дисперсия (виж глава 4).

Ще определим някои полезни свойства на точковите оценки - неизместеност, състоятелност и ефективност.

Нека имаме една оценка $\widehat{\theta}$. Доколко има смисъл да заместим истинската стойност на θ с нея? Едно очевидно изискване е очакваната стойност на $\widehat{\theta}$ да е точно истинската θ. Това формализираме като

Определение 36. *Казваме, че оценката на параметъра $\widehat{\theta}$ е неизместена, ако $\mathbb{E}\widehat{\theta} = \theta$.*

Голяма полза от неизместеността на оценката можем да получим като вземем предвид Централната гранична теорема на стр. **??** - че средното на n на брой измервания на една случайни величини клони към очакването и. Иначе казано, с достатъчно на брой експерименти ще можем да намерим *истинската* стойност на θ с произволно голяма точност!

За илюстрация на неизместените оценки ще използваме конкретен пример.

Пример 6. *(Фабрични дефекти) Разглеждаме популация от дефектни и недефектни изделия, при които вероятността всяко изделие да е дефектно сме означили с p. Случайната величина X определяме като 1, ако изделието е дефектно, и 0, ако не е. Иначе казано, X има бернулиево разпределение с параметър p.*

Трябва да оценим колко е този неизвестен параметър, т.е. каква е вероятността случайно избрано изделие да е дефектно.

За тези цели имаме извадка от n обекта, която сме проверили за дефекти.

Нека вземем за оценка $\widehat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$. Питаме се дали тя е неизместена? Проверяваме го чрез свойствата на математическото очакване:

$$\mathbb{E}\widehat{p} = \mathbb{E}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{\sum_{i=1}^{n} \mathbb{E}X_i}{n} = \frac{np}{n} = p$$

Тоест оценката е неизместена. Забележете, че ние *не знаем* колко точно е вероятността p, но това не пречи да покажем, че тя е равна именно на очакването на $\widehat{p}$!

Оценката $\widehat{p}$ интерпретираме като относителната честота на дефектните изделия в една извадка.

Нека X сега е случайна величина, модел на измервания. Нека $f(x, \mu)$ е плътността на случайната величина X, описваща популацията, а $\mu = \mathbb{E}X$ е неизвестен популационен параметър. Нека $X_1, X_2, \ldots, X_n$ е случайна извадка с обем n.

Ще докажем, че $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ е неизместена оценка на популационния параметър $\mu = \mathbb{E}X$. За тази цел ще пресметнем математическото очакване на $\overline{X}$

$$\mathbb{E}\overline{X} = \frac{1}{n} n \mathbb{E}X = \mathbb{E}X = \mu$$

Следователно $\overline{X}$ е неизместена оценка на популационното средно μ при произволно разпределение на случайната величина, описваща популацията. Оценката $\overline{X}$ наричаме извадково средно.

Изяснихме факта, че математическото очакване на извадковото средно $\overline{X}$ не зависи от броя на наблюденията в извадката и винаги е равно на средното на популацията.

Ще покажем, че

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

не е неизместена оценка на популационната дисперсия $\mathbb{D}X$.

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{2}{n}\overline{X}\sum_{i=1}^{n} X_i + \frac{n}{n}\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$$

$$\mathbb{E}\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}X_i^2 - \mathbb{E}\overline{X}^2 = \mathbb{E}X^2 - \frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^{n} X_i\right)^2 =$$

$$= \mathbb{E}X^2 - \frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^{n} X_i^2 + \sum_{i\neq j} X_i X_j\right) = \mathbb{E}X^2 - \frac{1}{n^2}n\mathbb{E}X^2 - \frac{1}{n^2}\sum_{i\neq j}\mathbb{E}X_i\mathbb{E}X_j =$$

$$= \mathbb{E}X^2 - \frac{1}{n}\mathbb{E}X^2 - \frac{n(n-1)}{n^2}\left(\mathbb{E}X\right)^2 = \frac{n-1}{n}\mathbb{E}X^2 - \frac{n-1}{n}\left(\mathbb{E}X\right)^2 = \frac{n-1}{n}\mathbb{D}X$$

Следователно

$$\mathbb{E}\widehat{\sigma}^2 \neq \mathbb{D}X$$

Нека да означим

$$s^2 = \frac{n}{n-1}\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

Тогава

$$\mathbb{E}s^2 = \frac{1}{n-1}\mathbb{E}\widehat{\sigma}^2 = \frac{n}{n-1}\frac{n-1}{n}\mathbb{D}X = \mathbb{D}X$$

Следователно s^2 е неизместена оценка на $\mathbb{D}X$.

Оценката е случайна величина. Естествено е да искаме стойностите, които тази случайна величина приема, да са близки до нейното средно. Естествено е да търсим оценки с малка дисперсия. От неравенството на Чебишов (виж параграф ???) за неизместената оценка $\widehat{\theta}$ е в сила:

$$\mathbb{P}\left(|\widehat{\theta} - \theta| > \epsilon\right) \leq \frac{1}{\epsilon^2}\mathbb{D}\widehat{\theta}$$

Затова изискването да се търсят неизместени оценки с минимална дисперсия се налага напълно естествено.

Определение 37. *Казваме, че оценката $\widehat{\theta}$ на параметъра θ е ефективна, ако е неизместена и с минимална дисперсия сред всички неизместени оценки на този параметър.*

В сила е следното твърдение:

Теорема 13. *(Рао - Блекуел) Неизместената оценка с минимална дисперсия ако съществува, е единствена.*

Нека имаме безкрайна редица от наблюдения и оценката на неизвестния параметър е построена по първите наблюдения.

Определение 38. *Казваме, че редицата от статистики $\widehat{\theta}_n$ е състоятелна оценка на параметъра θ, ако $\widehat{\theta}_n \xrightarrow{p} \theta$ при увеличаване на броя на наблюденията n.*

Съществува и по-силен вариант - строга състоятелност, където сходимостта е почти сигурно.

Забележете, че $\overline{X} \xrightarrow{p} \mathbb{E}X$ според слабия закон за големите числа. Следователно $\overline{X}$ е състоятелна оценка на популационния параметър $\mu = \mathbb{E}X$. Освен това $\overline{X}$ е силно състоятелна оценка на популационния параметър $\mu = \mathbb{E}X$, което следва от усиления закон за големите числа.

6.2 Един метод за построяване на точкови оценки

Вече използвахме конкретни точкови оценки на някои параметри и показахме техните ценни свойства. Естествен е въпросът как, по какъв начин, точкови оценки могат да бъдат получени.

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

- изключително стройна и завършена теория,
- добри асимптотични свойства на построените оценки.

Нека да предположим, че разпределението на популацията има плътност $f(x,\theta)$, известна с точност до неизвестен едномерен или многомерен параметър $\theta \in \Theta$. Тогава извадката $(X_1, X_2, \ldots, X_n)$ като вектор от независими случайни величини ще има плътност в извадковото пространство R^n от вида

$$L_n(x,\theta) = \prod_{i=1}^{n} f(x_i,\theta)$$

Съвместната плътност на наблюденията, разглеждана като функция на неизвестния параметър θ, наричаме функция на правдоподобие.

Определение 39. *Казваме, че оценката $\widehat{\theta}(x)$ удовлетворява принципа на максимално правдоподобие, ако*

$$L_n(x, \widehat{\theta}(x)) = max_\theta L_n(x,\theta)$$

Максимум на правдоподобието L_n се достига в същата точка и за логаритъма $LL_n(x,\theta) = \log L_n(x,\theta)$. Затова е удобно при намирането му да решаваме "уравненията на правдоподобие"

$$\frac{\partial LL_n(x,\theta)}{\partial\theta} = 0$$

Определение 40. *Казваме, че една оценка е максимално-правдоподобна, ако функцията на правдоподобие е диференцируема и оценката удовлетворява уравненията на правдоподобие.*

Доказва се, че ефективните оценки са максимално правдоподобни, но обратното не винаги е вярно. При достатъчно общи предположения относно функцията на правдоподобие се доказва, че максимално правдоподобните оценки са състоятелни, асимптотично (т.е. при достатъчно голям обем на извадката) ефективни и асимптотично нормални (виж параграф ???).

Ще използваме представения теоретичен метод за построяване на точкови оценки за да намерим оценки на параметрите на нормална популация. Т.е. нека $X \in N(\mu,\sigma^2)$ е случайната величина, която описва популацичта. Нека сме направили n наблюдения. Да намерим максимално-правдоподобните оценки на неизвестните параметри μ , σ^2.

Решение: Натуралният логаритъм от функцията на правдоподобие има вида:

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

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

$$0 = \frac{\partial LL_n(x,\mu,\sigma)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)$$

и

$$0 = \frac{\partial LL_n(x,\mu,\sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2$$

Така получаваме двете оценки $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$ и $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_i})^2$.

Разбира се, редно е да се убедим, че това са истински максимуми на LL_n.

Получените оценки имат следните свойства: Оценката $\overline{x}$ е неизместена и ефективна, но $\widehat{\sigma}^2$ е изместена. И двете оценки обаче са състоятелни.

6.3 Точкови оценки с R

Пример 7 (Застраховка "Гражданска отговорност"). *Цените на онлайн застраховка "Гражданска отговорност" на българския пазар за 2010 г. за 5 местен лек автомобил с обем на двигателя 1000 куб.см., регистриран 2009 година, са дадени в таблица 6.1. Какви средства трябва да планира едно младо семейство за плащане на този данък за личния си автомобил преди да си избере застрахователната компания с която ще работи? Т.е. искаме да оценим очакваната цена на онлайн застраховката "Гражданска отговорност" на българския пазар за 2010 г.*

На R извадковото средно се пресмята с функцията $mean()$.

Застрахователна компания	Цени на застраховката
1	151,45
2	361,14
3	110,94
4	211,2
5	258,9
6	159,54
7	242,97
8	154,9
9	230,5
10	207,1
11	197,06
12	210,9
13	188,61

Таблица 6.1: Цени на онлайн застраховка “Гражданска отговорност” (в лева)

```
> price <- c(151.45, 361.14, 110.94, 211.2, 258.9, 159.54,
      242.97, 154.9, 230.5, 207.1, 197.06, 210.9, 188.61)
> mean(price)
[1] 206.5546
```

Изводът е, че младото семейство трябва да планира 207 лева за застраховката "Гражданска отговорност" през 2010 г.

6.4 Доверителни интервали

Сега целта е да се построи един "случаен интервал", който ще съдържа истинската стойност на неизвестна константа с вероятност близка до 1.

Нека $X_1, X_2, \ldots, X_n$ са n случайни величини. Допускаме, че те имат съвместно дискретно или съвместно непрекъснато вероятностно разпределение с плътност $f(x_1, x_2, \ldots, x_n|\theta_1, \theta_2, \ldots, \theta_k)$, която зависи от k неизвестни параметъра $\theta_1, \theta_2, \ldots, \theta_k$. Целта ни е, като имаме наблюдения, да можем да кажем нещо за стойността на неизвестния параметър θ_1, каквито и да са стойностите на останалите параметри $\theta_2, \ldots, \theta_k$. По-точно, искаме да намерим интервала с минимална дължина, който с голяма вероятност покрива истинската стойност на неизвестния параметър. Т.е. трябва да намерим две функции $\underline{\theta_1}$ и $\overline{\theta_1}$ такива, че въз основа на наблюденията $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ да можем да изчислим стойностите $\underline{\theta_1}(X_1, X_2, \ldots, X_n)$ и $\overline{\theta_1}(X_1, X_2, \ldots, X_n)$ и след това да твърдим, че $\underline{\theta_1}(X_1, X_2, \ldots, X_n) \leq \theta_1 \leq \overline{\theta_1}(X_1, X_2, \ldots, X_n)$.

(??? $<$ или $\leq$)

6.4.1 Доверителен интервал за средна стойност

6.4.2 Задача

Пример 8 (Млади семейства спестяват от облекчението за ипотечни кредити)**.** *Облекчението за млади семейства, което през 2010 година се ползва за първи път, позволява от облагаемия доход да се приспаднат лихвените плащания по кредитите за покупка на*

жилище. Повече от 5 000 млади семейства са се възползвали от възможността да спестят от данъци върху изплащаните от тях ипотечни кредити, съобщи 22 юни 2010г. Националната агенция за приходите (НАП), цитирана от БТА.

Всяко младо семейство се интересува колко лева ще получи обратно когато от облагаемия им доход се приспаднат лихвените плащания по кредита за покупка на жилище. По-точно, в какъв интервал с 99% сигурност ще бъдат върнатите им пари?

За да намерим отговора на този въпрос сме направили случайна извадка от 93 млади семейства и сме записали получените данъчни облекчения на всяко от тях. Получени са следните данни: 360.12, 344.08, 346.17, 386.01, 344.96, 335.14, 353.79, 373.19, 341.51, 326.64, 386.70, 369.73, 376.77, 388.31, 369.74, 339.09, 385.96, 365.36, 383.97, 331.54, 337.65, 363.89, 361.88, 350.28, 353.78, 353.85, 334.88, 357.47, 327.38, 360.30, 368.42, 359.20, 381.56, 360.86, 398.42, 382.68, 362.96, 384.38, 359.51, 344.14, 370.85, 361.76, 355.06, 365.27, 353.44, 369.42, 369.31, 386.67, 349.48, 375.48, 357.93, 351.29, 375.12, 364.87, 373.64, 354.62, 323.37, 381.48, 332.35, 361.42, 365.03, 340.34, 346.55, 398.11, 344.20, 373.73, 336.45, 360.18, 350.29, 330.30, 354.15, 341.18, 337.20, 364.16, 331.32, 364.94, 363.11, 360.81, 342.25, 345.31, 355.11, 369.91, 375.39, 345.91, 369.09, 373.55, 386.44, 366.73, 389.20, 354.23, 363.93, 333.72, 373.01

6.4.3 Модели

В този параграф ще разгледаме три модела.

Разглеждаме нормална (гаусова) популация с неизвестно средно μ и известна дисперсия σ^2. Нека $X_1, X_2, ..., X_n$ е случайна извадка от тази популация. Знаем, че извадковото средно $\overline{X}$ е неизместена, състоятелна и с най-малка дисперсия точкова оценка на неизвестното μ. Ще разгледаме близостта на $\overline{X}$ до μ. За тази цел ще използваме вероятностното разпределение на $\overline{X}$. По-точно, ще използваме следното твърдение:

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \in N(0,1)$$

(??? къде сме го доказали ???)

Искаме, когато α е произволно малко число, например $\alpha = 0.05$, да е изпълнено равенството:

$$\mathbb{P}\left(-z_{1-\frac{\alpha}{2}} < \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

където $z_{1-\frac{\alpha}{2}}$ е $1 - \frac{\alpha}{2}$ квантила на стандартното нормално разпределение.

Чрез еквивалентни преобразувания получаваме:

$$\mathbb{P}\left(\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

където $z_{1-\frac{\alpha}{2}}$ е $1 - \frac{\alpha}{2}$ квантила на нормалното разпределение $N(0, 1)$.

Интервалът, чиито граници са случайни величини, наричаме случаен интервал. Следователно, вероятността случайният интервал

$$\left(\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}; \overline{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

да съдържа неизвестното средно μ е $1 - \alpha$.

Направили сме случайна извадка. Като използваме числовите стойности на наблюденията пресмятаме извадковото средно $\overline{x}$. Като заместим извадковата статистика $\overline{X}$ със съответната й числова стойност $\overline{x}$ получаваме границите на един числов интервал и те са:

$$\left(\overline{x} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}; \overline{x} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

Тъй като, преди да е направена извадката вероятността съответният случаен интервал да покрива μ е $1 - \alpha$, получения числов интервал наричаме $(1 - \alpha).100\%$ доверителен интервал за средното μ на нормалната популация с известна дисперсия σ^2.

Интерпретацията на доверителния интервал е следната:

С $(1-\alpha).100\%$ сигурност може да твърдим, че неизвестната средна μ е по-голяма от $\overline{x} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$ и по-малка от $\overline{x} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$.

По аналогичен начин получаваме, че

$$\left(\overline{x} - t_{1-\frac{\alpha}{2}}(n-1)\frac{\sigma}{\sqrt{n}}; \overline{x} + t_{1-\frac{\alpha}{2}}(n-1)\frac{\sigma}{\sqrt{n}}\right)$$

където $t_{1-\frac{\alpha}{2}}(n-1)$ е $1-\frac{\alpha}{2}$ квантила на $T(n-1)$ разпределението, а s^2 е неизместена и състоятелна оценка на неизвестната дисперсия,

е $(1-\alpha).100\%$ доверителен интервал за средното μ на нормалната популация с неизвестна дисперсия.

Ще построим $(1-\alpha).100\%$ доверителен интервал за средното μ на популация с неизвестно разпределение когато случайната извадка е с голям обем. Знаем, че от Централната гранична теорема следва твърдението

$$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \approx N(0,1)$$

Тогава

$$\left(\overline{x} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}; \overline{x} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

е $(1-\alpha).100\%$ доверителен интервал за средното μ на популация с неизвестно разпределение и с извадка с голям обем.

Като използваме числовите стойности на наблюденията и съответните квантили, изчисляваме границите на $(1-\alpha).100\%$ доверителен интервал за неизвестното популационно средно μ.

6.4.4 Построяване на доверителен интервал за средно с R

```
> relief <- c(360.12, 344.08, 346.17, 386.01, 344.96,
335.14, 353.79, 373.19, 341.51, 326.64, 386.70, 369.73,
376.77, 388.31, 369.74, 339.09, 385.96, 365.36, 383.97,
331.54, 337.65, 363.89, 361.88, 350.28, 353.78, 353.85,
334.88, 357.47, 327.38, 360.30, 368.42, 359.20, 381.56,
360.86, 398.42, 382.68, 362.96, 384.38, 359.51, 344.14,
370.85, 361.76, 355.06, 365.27, 353.44, 369.42, 369.31,
386.67, 349.48, 375.48, 357.93, 351.29, 375.12, 364.87,
373.64, 354.62, 323.37, 381.48, 332.35, 361.42, 365.03,
340.34, 346.55, 398.11, 344.20, 373.73, 336.45, 360.18,
350.29, 330.30, 354.15, 341.18, 337.20, 364.16, 331.32,
364.94, 363.11, 360.81, 342.25, 345.31, 355.11, 369.91,
375.39, 345.91, 369.09, 373.55, 386.44, 366.73, 389.20,
354.23, 363.93, 333.72, 373.01 )

> t.test (relief, conf.level=0.99)
```

```
        One Sample t-test

data:  relief
t = 198.3558, df = 92, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 354.9177 364.4571
sample estimates:
mean of x
 359.6874
```

С 99% сигурност едно младо семейство ще спести данъци поне 354.92 лева и не повече от 364.46 лева с облекчението за млади семейства.

6.4.5 Доверителни интервали за разликата на две средни

Разглеждаме две нормални популации с неизвестни средни, които се описват със случайните величини $X \in N(\mu_1, \sigma_1)$ и $Y \in N(\mu_2, \sigma_2)$. Нека $X_1, X_2, ..., X_n$ и $Y_1, Y_2, ..., Y_m$ са две независими случайни извадки с обем n и m, съответно. Знаем, че

$$\overline{X} \in N\left(\mu_1, \frac{\sigma_1^2}{n}\right), \overline{Y} \in N\left(\mu_2, \frac{\sigma_2^2}{m}\right), \overline{X} - \overline{Y} \in N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$$

Освен това, когато обемите на извадките са достатъчно големи, е вярно и твърдението

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \approx N(0, 1)$$

което е резултат от Централната гранична теорема.

Тогава $(1 - \alpha).100\%$ доверителен интервал за разликата на две средни $\mu_1 - \mu_2$ когато обемите на извадките са достатъчно големи се дава с формулата

$$\left((\overline{x} - \overline{y}) - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}; (\overline{x} - \overline{y}) + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\right)$$

6.4.6 Доверителен интервал за дисперсия

Нека

$$X \in N(\mu, \sigma^2)$$

е случайна величина, която моделира популацията. Нека средното μ е известно и $X_1, X_2, ..., X_n$ е случайна извадка. Ще намерим доверителна област за неизвестната дисперсия σ^2. Знаем, че статистиката

$$\frac{ns_n^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \in \chi^2(n)$$

Искаме

$$\mathbb{P}\left(q_l < \frac{ns_n^2}{\sigma^2} < q_u \right) = 1 - \alpha$$

Квантилите на χ^2 разпределението q_l и q_u се определят от уравнението $F(q_l) + 1 - F(q_u) = \alpha$. Следователно

$$\frac{ns_n^2}{q_u} \leq \sigma^2 \leq \frac{ns_n^2}{q_l}$$

където $s_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$ и $\frac{ns_n^2}{\sigma_2} \in \chi^2(n)$ при известно популационно средно μ.

Следователно, когато популационното средно μ е известно, $(1 - \alpha).100\%$ доверителен интервал за неизвестната дисперсия σ^2 се дава с формулата

$$\left(\frac{ns_n^2}{\chi^2_{1-\frac{\alpha}{2}}(n)} < \sigma^2 < \frac{ns_n^2}{\chi^2_{\frac{\alpha}{2}}(n)} \right)$$

където s_n^2 е неизместена и състоятелна оценка на неизвестната популационна дисперсия σ^2, а $\chi^2_{\frac{\alpha}{2}}$ и $\chi^2_{1-\frac{\alpha}{2}}$ са съответно $\frac{\alpha}{2}$ и $1 - \frac{\alpha}{2}$ квантилите на χ^2 разпределението с n степени на свобода.

В случая, когато популационното средно μ е неизвестно за оценка на неизвестната популационна дисперсия използваме $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2$.

Тогава

$$\frac{(n-1)s^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2 \in \chi^2(n-1)$$

Когато популационното средно е неизвестно, $(1-\alpha).100\%$ доверителен интервал за неизвестната дисперсия се дава с формулата

$$\frac{(n-1)s_n^2}{q_u} \leq \sigma^2 \leq \frac{(n-1)s_n^2}{q_l}$$

На практика така определения интервал с минимална дължина се използва рядко. По-често се приравняват вероятностите на двете опашки $F(q_l) = 1 - F(q_u) = \frac{\alpha}{2}$.

$$\left(\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}; \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)} \right)$$

6.4.7 Доверителен интервал за отношение на две дисперсии

Първи случай. Разглеждаме две нормални популации с неизвестни средни, които се описват със случайните величини $X \in N(\mu_1, \sigma_1)$ и $Y \in N(\mu_2, \sigma_2)$. Нека $X_1, X_2, ..., X_n$ и $Y_1, Y_2, ..., Y_m$ са две независими случайни извадки с обем n и m, съответно.

Известно е, че

$$F = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}} \in F(n-1, m-1)$$

За дадени стойности за m и n и предварително зададена вероятност $1-\alpha$ (например $1-\alpha = 0,95$) от таблиците на F разпределението можем да определим числата $0 < a < b$, така че

$$\mathbb{P}\left(a < \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}} < b \right) = 1 - \alpha$$

$$\mathbb{P}\left(a\frac{s_2^2}{s_1^2} < \frac{\sigma_2^2}{\sigma_1^2} < b\frac{s_2^2}{s_1^2} \right) = 1 - \alpha$$

Интервалът $\left(a\frac{s_2^2}{s_1^2}; b\frac{s_2^2}{s_1^2} \right)$ е случаен. Той съдържа фиксираната, но неизвестна точка $\frac{\sigma_2^2}{\sigma_1^2}$ с вероятност $1 - \alpha$.

Нека $X_1, X_2, ..., X_n$ и $Y_1, Y_2, ..., Y_m$ са съответните наблюдения. Тогава интервалът с известни крайни точки $\left(a\frac{s_2^2}{s_1^2}; b\frac{s_2^2}{s_1^2}\right)$ е $(1-\alpha).100\%$ доверителен интервал за отношението $\frac{\sigma_2^2}{\sigma_1^2}$ на двете неизвестни дисперсии.

Определянето на a и b не е еднозначно. Обикновено числата a и b определяме така, че $\mathbb{P}(F < a) = \mathbb{P}(F > b) = \frac{\alpha}{2}$, където $F \in F(n-1, m-1)$. Следователно a е решението на уравнението $F(a) = \frac{\alpha}{2}$, т.е. a е $\frac{\alpha}{2}$ квантила на F разпределението.

$$F(b) = \mathbb{P}(F < b) = 1 - \frac{\alpha}{2}$$

Следователно b е $1 - \frac{\alpha}{2}$ квантила на F разпределението.

Следователно когато популационните дисперсии не са известни, $(1-\alpha).100\%$ доверителен интервал за отношението $\frac{\sigma_2^2}{\sigma_1^2}$ е

$$\left(F_{\frac{\alpha}{2}}(n-1, m-1)\frac{s_2^2}{s_1^2}; F_{1-\frac{\alpha}{2}}(n-1, m-1)\frac{s_2^2}{s_1^2}\right)$$

Втори случай. Двете нормални популации са с известни дисперсии. Тогава по аналогичен начин се доказва, че $(1-\alpha).100\%$ доверителен интервал за отношението $\frac{\sigma_2^2}{\sigma_1^2}$ е

$$\left(F_{\frac{\alpha}{2}}(n, m)\frac{s_m^2}{s_n^2}; F_{1-\frac{\alpha}{2}}(n, m)\frac{s_m^2}{s_n^2}\right)$$

където $s_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\overline{X} - \mu_1\right)^2$, $s_m^2 = \frac{1}{m}\sum_{i=1}^{m}\left(\overline{Y} - \mu_2\right)^2$.

Глава 7

Методи за вземане на решения

7.1 Статистически изводи

Статистическите изводи са заключения за различни свойства на популацията, направени въз основа на наблюденията и на различни предположения за популацията.

Така, ако предположенията са верни, нашите твърдения стават функции на извадката, т.е. придобиват случаен характер - стават случайни величини. Тъй като твърденията приемат две "стойности" - истина и неистина, задачата всъщност е да намерим вероятността едно заключение да бъде вярно.

Най-популярната и коректна форма за построяване на статистически извод е статистическата хипотеза. Много често имаме основания да предположим, че неизвестното разпределение на популацията притежава плътност $f(x)$. За основен инструмент ни служи знаменитата Лема на Нейман-Пирсън (виж [**?**]).

Проверката на статистически хипотези е един от главните клонове на математическата статистика, който от една страна е тясно свързан с приложението на статистиката, а от друга е доста теоретичен и абстрактен по съдържание.

7.2 Философия

Има много случаи, когато сме изправени пред въпроса "Кое от две възможни състояния на природата е вярното?". Ние не знаем кое е истинското състояние и от нас се иска да го определим. Например:

- Лекар рентгенолог трябва да вземе решение дали пациентът е болен, или симптомите са случайни.
- Експерт по хранително-вкусова промишленост трябва да провери има ли растителни мазнини в произведения кашкавал.
- Специалист по маркетинг иска да определи кой от два вида колбаси се харесва повече на потребителите.
- Учен трябва да вземе решение дали сред потока частици, получени в експеримента на ускорителя LHC в Женева, присъства поне един Higgs бозон (частицата, която дава маса на материята).

Ние непрекъснато сме изправени пред проблема да решим кое от алтернативните възможни състояния в природата е действително вярното.

Ако лекарят, експертът, ученият успеят да достигнат до едно абсолютно сигурно решение със средствата на техните науки, то с това задачата е решена. Може обаче да се случи така, че те да не бъдат в състояние да решат проблема със средствата и методите на своята дисциплина, ??? например когато експерименталната грешка е значителна, или когато самото измерване не може да бъде направено с достатъчна точност. Тогава един статистик сигурно би могъл да помогне.

7.3 Основи на тестването на хипотези

Естествено е да опишем задачата за тестване на хипотези в термините на теория на вероятностите.

Ако имаме вероятностно пространство, и в него хипотезите H_0 и H_1, то формално можем да построим вероятностите $\mathbb{P}(H_i)$, които в

теорията на вероятностите наричахме вероятността да се случи събитието H_i. Това представляват априорните вероятности съответната хипотеза да е вярна.

Но верността ??? на хипотезите влияе на вероятностите на другите събития в света, т.е. ако едно такова събитие е A, то $\mathbb{P}(A) \neq \mathbb{P}(A|H_i)$ в общия случай. Затова и ??? *верността на събитието A влияе на вероятността на H_i.*

Тогава вместо да се задоволяваме с априорните оценки на вероятности на хипотезите $\mathbb{P}(H_i)$, можем да изследваме света около нас дали събитието A се е случило (т.е. да проведем експеримент), и имайки, че събитието A се е случило да използваме за по-добра оценка на вероятностите на H_i *условната вероятност* $\mathbb{P}(H_i|A)$, а ако не се е случило A ще използваме за по-добра оценка на вероятността на H_i *условната вероятност* $\mathbb{P}(H_i|\neg A)$.

Пример 9 (Лекарство)**.** *Да предположим, че имаме лекарство, разработено от компания, и си задаваме въпроса дали то действа реално на пациентите. Дефинираме хипотезите H_0 : лекарството не действа и H_1 : лекарството действа. Въпросът е как можем да преценим действието на лекарството, тъй като ние не можем директно да наблюдаваме действието му върху организма.*

Ако хипотеза H_0 е вярна, то това би имало различно отражение на света около нас спрямо случая в който H_1 е вярна. Затова се опитваме да построим събитие A, което да проверим дали се е случило (като проведем клинично изследване) и въз основа на неговото случване да определим вероятността на H_0.

Пример за такова събитие A е много повече пациенти са оцелели, отколкото се очаква да го направят при недействащо лекарство (т.е. ако е изпълнено H_0). Такова събитие би променило вероятностите в някаква степен, а точно в каква степен ще разгледаме по-долу.

Така че можем да разгледаме задачата за проверка на хипотези по следния начин - да определим A - събитие - такова, че случването на A да доведе до вероятности $\mathbb{P}(H_i|A)$ близки до 0 или 1, и това H_i,

за което вероятността е "близо" до 1 ще заключим че се е сбъднало. Това е така наречения Бейсов подход.

В практиката обаче пресмятането на $\mathbb{P}(H_i|A)$ е свързано с много трудности (и теоретични, и практически). За това се разглежда близката задача - да се определи такова A, за което $\mathbb{P}(A|H_i)$ е "малко", и да се използва следният аргумент, по същество близък до "допускане на противното".

Нека допуснем, че H_0	Нека допуснем, че H_0
От него следва, че A не е вярно, но ние знаем, че A е вярно.	Тогава следва, че вероятността да се случи A е много малка, но въпреки това то се е случило.
Следователно H_0 е невярно.	Тогава можем да отхвърлим H_0.

Фигура 7.1: Сравнение на вероятностния с детерминистичния аргумент

Това е така наречения класически подход, илюстриран на фигура 7.1. ???

За да го приложи статистикът извършва следното:

1. Конструира опит, изходът от който е случаен. Избира събитие A, което може да настъпи по време на опита или да не настъпи. Когато H_0 е вярно, събитието настъпва с голяма вероятност (близка до 1). Когато H_1 е вярно, събитието A с голяма вероятност *не* настъпва (или което е същото, настъпва допълнителното събитие).

2. Действително извършва опита. Ако събитието A настъпи, статистикът решава, че H_1 е вярното състояние. Ако A не настъпи, той решава, че състоянието H_0 е вярно.

Сега ще разгледаме как строим тези хипотези и експериментите, с които ги проверяваме.

7.3.1 Прости алтернативи ???

Предполагаме, че имаме случайна величина X.

Да предположим, че за двата случая, т.е. за двете възможни състояния S_1 и S_2, имаме две вероятностни плътности, които ще означим съответно с $f_0(x)$ и $f_1(x)$, една от които е истинската плътност на случайната величина X.

Хипотезата H_0 ще отъждествяваме с плътността $f_0(x)$, а алтернативата H_1 - с плътността $f_1(x)$. Това ще записваме по следния начин:

$$H_0 : f_0(x)$$

$$H_1 : f_1(x)$$

Има две възможни грешки, които могат да бъдат направени - да отхвърлим основната хипотеза H_0, когато тя е вярна и да приемем основната хипотеза H_0, когато е вярна алтернативата H_1.

Обикновено се случва така, че колкото по-малка се опитваме да направим вероятността за първата грешка, толкова по-голяма ще бъде вероятността за втората грешка и обратното.

Тогава нашата задача се състои в установяването на една максимална вероятност, с която можем да си позволим да направим по-неприятната от тези две грешки. При това състояние се опитваме да минимизираме вероятността за другата грешка.

Допускаме, че по-лошата грешка е да отхвърлим хипотезата H_0, когато тя е вярна. Тази по-лоша грешка се нарича *грешка от I род*. Другата грешка се нарича грешка от II род. Фиксираме една максимална вероятност α, с която си позволяваме да извършим такава грешка. Статистиците обикно работят с $\alpha = 0.05$.

Тази максимална вероятност, с която можем да си позволим да направим по-лошата от двете грешки, обикновено се нарича *ниво на съгласие* (*ниво на значимост*).

Следователно, една статистическа проверка на хипотезата H_0 се дефинира като едно събитие E. Ако събитието E настъпи, тогава ние отхвърляме хипотезата H_0, ако събитието E не настъпи, тогава нямаме основание да отхвърлим хипотезата H_0. Събитието E трябва да е такова, че да е в сила неравенството $\mathbb{P}(E|H_0) \leq \alpha$.

Събитието $E = \{\text{отхвърляне на } H_0\}$ трябва да се дефинира в зависимост от случайна величина X, т.е. определя се подмножество $W \in \mathrm{R}$ и ако $X \in W$, то отхвърляме H_0.

Множеството W се нарича *критична област*. Тогава ще бъде изпълнено $\mathbb{P}(X \in W|H_0) \leq \alpha$.

Обикновено не е много трудно да се намери една критична област. Често обаче съществуват много критични области. Тогава търсим нова област W, за която е изпълнено:

$$\mathbb{P}(X \in W|H_1) \to \max$$

т.е.

$$\mathbb{P}(X \notin W|H_1) \to \min$$

т.е. търсим нова област W, за която $\mathbb{P}(X \in W)$ е най-голяма когато $f_1(x)$ е истинската плътност. Ако такава критична област съществува, тя се нарича оптимална критична област.

Извод: При проверка на хипотезата $H_0 : f_0(x)$ срещу контрахипотезата $H_1 : f_1(x)$ с ниво на съгласие α, критична област е онова подмножество W на $\mathbb{P}^n$, за което

$$\mathbb{P}(X \in W^*|H_1) \geq \mathbb{P}(X \in W|H_1)$$

Тогава W^* наричаме оптимална критична област. Една оптимална критична област може да бъде получена с лемата на Нейман-Пирсън.

Процедурата за проверка на статистически хипотези се състои от следните пет стъпки:

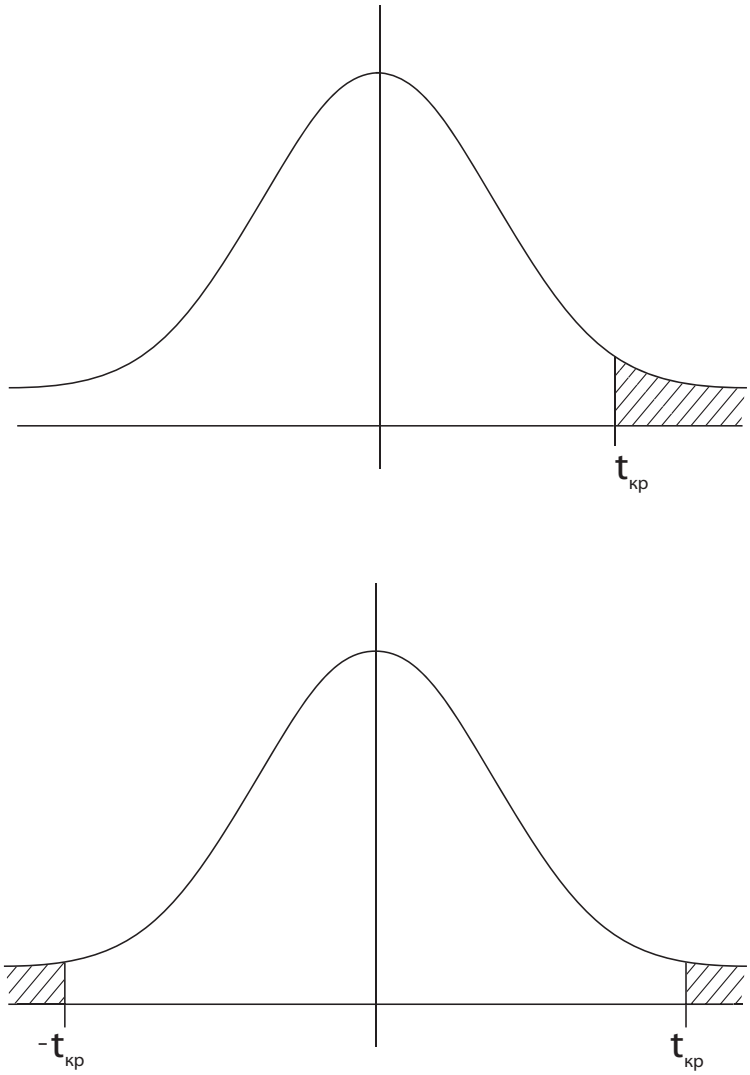
1. Намираме подходяща статистика и нейното вероятностно разпределение. Тази стъпка зависи от вида на основната хипотеза.

2. Построяване на критична област. Тази стъпка зависи от вида на контрахипотезата.

3. Като използваме данните от извадката изчисляваме стойността на статистиката.

4. Статистически извод: Ако получената стойност принадлежи на критичната област, отхвърляме основната хипотеза H_0 в полза на H_1, т.е. приемаме H_1 за вярна. Ако получената стойност не принадлежи на критичната област, нямаме основание да отхвърлим основната хипотеза H_0 и я приемаме за вярна. Това не означава, че основната хипотеза H_0 е вярна, а само че данните от извадката се съгласуват с нея.

5. Извод в термините на приложната област.



Фигура 7.2: Критичната област е обозначена с щрих. Вижда се разликата между едностранна и двустранна критична област при едно и също ниво на доверие

7.4 Проверка на хипотези за средното

Пример 10. *(Домът на щъркела) Щъркеловите гнезда, които птиците вият по електрическите стълбове, понякога са причина за чести аварии по съоръженията на електроразпределителната мрежа. Електроразпределителна компания се стреми да предоставя качествена услуга на своите абонати и в същото време полага грижи за опазване на екологията на района. За да защити птиците, а и да предпази съоръженията от аварии, тя монтира метални платформи за щъркелови гнезда върху електрически стълбове. Известно е, че през дъждовните години, когато реките се разливат и наводняват много райони, птиците отглеждат по 4-5 малки. Когато обаче блатата са сухи, няма риба, жаби и змии за храна на щъркелите и те отглеждат по 1-2 малки.*

За да вземе решение за тежестта, на която тези платформи трябва да издържат, компанията е решила да изследва тегло на щъркелите, които се размножават в България. Белият щъркел, е голяма птица, висока 1*м. с* 2*м. размах на крилата. Но внушителните размери на белия щъркел са измамни за теглото му, тъй като той има кухи кости.*

За да проверят хипотезата, че средното тегло на един щъркел е по-малко от 4 *кг., еколози са измерили теглото на* 89 *щъркела, избрани по случаен начин. Получили са следните данни: 3.58, 3.11, 3.04, 3.30, 4.35, 4.10, 2.77, 3.66, 2.75, 3.29, 3.85, 3.37, 2.87, 3.32, 3.38, 3.20, 3.14, 3.41, 3.35, 3.88, 2.99, 3.05, 2.93, 3.83, 3.53, 3.76, 3.67, 3.71, 3.51, 3.51, 3.18, 4.49, 3.85, 3.29, 2.89, 3.87, 3.41, 2.93, 3.55, 4.05, 4.14, 3.90, 2.56, 3.32, 3.80, 3.16, 2.56, 2.77, 3.65, 4.22, 3.84, 3.54, 3.96, 2.83, 3.52, 3.04, 3.79, 4.51, 3.11, 3.18, 3.53, 2.91, 3.52, 2.89, 3.42, 3.96, 3.44, 3.83, 4.08, 3.42, 4.03, 3.27, 3.33, 3.03, 3.99, 3.15, 2.86, 3.73, 3.48, 3.29, 3.83, 3.78, 3.67, 3.37, 3.94, 3.50, 3.13, 3.63, 3.16.*

Искаме да сме 99% *сигурни в нашия извод.*

Ще формулираме задачата на езика на статистиката. Хипотезата, която искаме да проверим, е:

$$H_0 : \mu = 4$$

т.е. средното тегло на щъркела е 4 кг, срещу алтернативата

$$H_1 : \mu < 4$$

т.е. средното тегло на щъркела е по-малко от 4 кг, с ниво на съгласие $\alpha = 0.01$, т.е. искаме да сме 99% сигурни в нашите изводи.

Ще моделираме ситуацията по следния начин:

Нека $X \in N(\mu, \sigma^2)$, където μ и σ^2 са параметрите на вероятностното разпределение, е случайната величина, която моделира популацията. Съгласно дефинициите в глава **??** ще разглеждаме $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, които са n независими наблюдения над X, т.е $X_1, X_2, \ldots, X_n$ са нашите наблюдавани случайни величини, $x_1, x_2, \ldots, x_n$ са съответно стойностите, които те приемат, а с $\overline{x}$ ще означим извадковото средно:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Задачата е да проверим хипотезите

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

с ниво на съгласие α.

Ще разгледаме два случая в зависимост от това дали знаем колко е дисперсията σ^2 или не.

7.4.1 ... с критерий на Стюдънт

Обикновено на практика не знаем колко е дисперсията на популацията. Затова вместо σ^2 използваме нейната неизместена и състоятелна оценка

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$$

Това равенство можем да представим във вида $\frac{(n-1)s^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{\sigma^2}$. Тогава от теоремата на Кокрън (Кендалл, Стюарт 1966)(Кендалл, М., Стюарт, А. (1966). Теория распределений. Москва: "Наука")

$$\sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{\sigma} \right)^2 \in \chi^2(n-1)$$

защото имаме една линейна зависимост и тя е $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Знаем, че

$$\overline{X} \in N(\mu, \frac{\sigma^2}{n})$$

и от теорема ??? следва, че

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \in N(0,1)$$

Известно е още, че случайните величини $\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ и $\frac{(n-1)s^2}{\sigma^2}$ са независими.

Тогава за частното на тези две случайни величини е в сила твърдението

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{s^2(n-1)}{\sigma^2(n-1)}}} = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \in T(n-1)$$

Тогава при вярна основна хипотеза статистиката

$$t = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}} \in T(n-1)$$

За да проверим хипотезата, съгласно разсъжденията в **??** трябва да използваме двустранна критична област, т.е. да проверим дали наблюдаваната стойност на t-статистиката

$$t_0 = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

е извън интервала, заключен между квантилите на T-разпределението с $n-1$ степени на свобода $\left[t_{(n-1)}(\frac{\alpha}{2}), t_{(n-1)}(1-\frac{\alpha}{2})\right]$. Забележете, че T-разпределението е симетрично спрямо 0 и точката $t_{(n-1)}(\frac{\alpha}{2})$ съвпада с точката $-t_{(n-1)}(1-\frac{\alpha}{2})$.

За дадено t_0 и двустранна критична област, p стойността $p-value$ се дефинира по следния начин:

$$p - value = \mathbb{P}(|t| > |t_0|)$$

където статистиката t има T разпределение с $n-1$ степени на свобода. Малка стойност на $p - value$ съответства на голяма стойност

на $|t_0|$. Този факт се счита за доказателство, че разликата между популационното средно и извадковото средно е значима и води до отхвърляне на основната хипотеза.

Сега да разгледаме задачата за проверка на хипотезата

$$H_0 : \mu = \mu_0$$

срещу алтернативата

$$H_1 : \mu < \mu_0$$

с ниво на съгласие α.

За да проверим хипотезата, съгласно разсъжденията в **??** трябва да използваме лява едностранна критична област, т.е. да проверим дали наблюдаваната стойност на t-статистиката

$$t_0 = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

е в интервала $\left[-\infty, t_{(n-1)}(\alpha)\right]$, който съвпада с интервала $\left[-\infty, -t_{(n-1)}(1-\alpha)\right]$.

За дадено t_0 и лява едностранна критична област p стойността $p - value$ се дефинира като

$$p - value = \mathbb{P}(t < t_0)$$

където статистиката t има T разпределение с $n-1$ степени на свобода. Малка стойност на $p - value$ съответства на голяма стойност на $|t_0|$. Този факт е доказателство, че разликата между популационното средно и извадковото средно е значима и затова отхвърляме основната хипотеза.

Задачата сега е да проверим хипотезата

$$H_0 : \mu = \mu_0$$

срещу контрахипотезата

$$H_1 : \mu > \mu_0$$

с ниво на съгласие α.

За да проверим хипотезата, съгласно разсъжденията в **??** трябва да използваме дясна едностранна критична област, т.е. да проверим дали наблюдаваната стойност на t-статистиката

$$t_0 = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

е в интервала $\left[t_{(n-1)}(1-\alpha), \infty\right]$.

За дадено t_0 и дясна едностранна критична област p стойността $p - value$ се дефинира като

$$p - value = \mathbb{P}(t > t_0)$$

където статистиката t има T разпределение с $n-1$ степени на свобода. Малка стойност на $p - value$ съответства на голяма стойност на t_0. Този факт е доказателство, че разликата между популационното средно и извадковото средно е значима и води до отхвърляне на основната хипотеза.

Следователно и при трите възможни контрахипотези за да направим статистически извод е достатъчно да проверим верността на твърдението $p - value < \alpha$.

Известно е, че телестното тегло зависи от много фактори, всеки от които поотделно има малко влияние върху теглото. Затова телесното тегло се моделира с нормално разпределена случайна величина. Следователно теглото на щъркела ще моделираме с нормално разпределена случайна величина с неизвестни параметри на вероятностното разпределение. За да проверим хипотезата за средното тегло на щъркела, който гнезди в България, ще използваме критерия, чиито теоретични основи току-що представихме, известен като t критерия на Студент.

В системата за статистически изследвания R критерият на Студънт се реализира с функцията t.test().

```
> stork.weight <- c(3.58, 3.11, 3.04, 3.30, 4.35, 4.10,
2.77, 3.66, 2.75, 3.29, 3.85, 3.37, 2.87, 3.32, 3.38, 3.20,
3.14, 3.41, 3.35, 3.88, 2.99, 3.05, 2.93, 3.83, 3.53, 3.76,
3.67, 3.71, 3.51, 3.51, 3.18, 4.49, 3.85, 3.29, 2.89, 3.87,
3.41, 2.93, 3.55, 4.05, 4.14, 3.90, 2.56, 3.32, 3.80, 3.16,
2.56, 2.77, 3.65, 4.22, 3.84, 3.54, 3.96, 2.83, 3.52, 3.04,
3.79, 4.51, 3.11, 3.18, 3.53, 2.91, 3.52, 2.89, 3.42, 3.96,
3.44, 3.83, 4.08, 3.42, 4.03, 3.27, 3.33, 3.03, 3.99, 3.15,
2.86, 3.73, 3.48, 3.29, 3.83, 3.78, 3.67, 3.37, 3.94, 3.50,
3.13, 3.63, 3.16)

> t.test (stork.weight, mu=4, alternative="less")

        One Sample t-test
```

```
data:  stork.weight
t = -11.681, df = 88, p-value < 2.2e-16
alternative hypothesis: true mean is less than 4
95 percent confidence interval:
     -Inf 3.540222
sample estimates:
mean of x
 3.463933
```

От неравнството $p - value < \alpha$ следва, че отхвърляме основната хипотеза в полза на контрахипотезата.

Извод: Теглото на белия щъркел, който гнезди в България, е по-малко от 4 кг.

В този параграф разгледахме един от най-важните критерии в областта на статистическите изводи, известен като критерий на Стюдънт или t – критерий. Той се използва за да се решат голям брой задачи, които се срещат в реалния живот.

Как се решават подобни задачи от практиката когато не може да се предположи, че разпределението на популацията е нормално, ще разберем от параграфите ??? и ???.

7.4.2 ... при известна дисперсия

В някои, макар и редки случаи, дисперсията на популацията е известна. Т.е. знаем, че $\sigma^2 = \sigma_0^2$, където σ_0^2 е известно число.

Ако хипотезата H_0 е вярна, то $X \in N(\mu_0, \sigma_0^2)$. Тогава от допускането, че X_i са независими и еднакво разпределени случайни величини следва, че $\overline{X} \in N(\mu_0, \frac{\sigma_0^2}{n})$. Тогава

$$t = \frac{\overline{X} - \mu_0}{\frac{\sigma_0}{\sqrt{n}}} \in N(0,1)$$

За да проверим хипотезата, съгласно разсъжденията в **??** трябва да използваме двустранна критична област, т.е. да проверим дали наблюдаваната стойност на t-статистиката

$$t_0 = \frac{\overline{x} - \mu_0}{\frac{\sigma_0}{\sqrt{n}}}$$

е извън интервала, заключен между квантилите на нормалното разпределение $\left[\Phi(\frac{\alpha}{2}), \Phi(1-\frac{\alpha}{2})\right]$.

Окончателно, критерият за проверка на хипотези за средното на нормална популация се основава на статистика, която има нормално разпределение, когато популационната дисперсия е известна, и на статистика, която има T разпределение, когато популационната дисперсия не е известна. Във всички случаи ако е изпълнено $p - value < \alpha$, отхвърляме основната хипотеза в полза на алтернативата. Ако това неравенство не е изпълнено нямаме основание да отхвърлим основната хипотеза защото данните се съгласуват с нея.

Отсега нататък когато проверяваме хипотези с помощта на статистически софтуер за да направим статистическия извод ще използваме само стойността на $p - value$.

7.4.3 ... на популация с произволно разпределение

Остана да разгледаме въпроса как можем да проверим хипотеза за средно на популация при произволно разпредление. На помощ ни идва Централната гранична теорема (виж **??**.

Тя гласи, че ако $X_1, X_2, \ldots$ е редица от еднакво разпределени и независими случайни величини с крайна дисперсия σ^2, и дефинираме частичните суми на първите n случайни величини от нея $S_n = \sum_{i=1}^{n} X_i$, то

$$\frac{S_n - \mathbb{E}S_n}{\sqrt{\mathbb{D}S_n}} \to^d X \in N(0,1)$$

Еквивалентното твърдение е:

$$\frac{\sum_{i=1}^{n} X_n - n\mu}{\sqrt{n\sigma^2}} \sim N(0,1)$$

Окончателно, ако H_0 е вярна, т.е. когато случайната величина X, която моделира популацията, има произволно вероятностно разпределение с неизвестно средно μ_0, известна дисперсия σ_0^2 и размера на извадката n е достатъчно голям[1], то статистиката

$$t = \frac{\overline{x} - \mu_0}{\frac{\sigma_0}{\sqrt{n}}}$$

[1] За практически цели е досатъчно $n > 30$

се апроксимира от стандартното нормално разпределение.

Проверката на хипотезата за средното на популацията се основава на твърдението, че при вярна H_0 статистиката има приблизително стандартно нормално разпределение.

За да проверим хипотезата, съгласно разсъжденията в **??** трябва да използваме двустранна критична област, т.е. да проверим дали наблюдаваната стойност на t-статистиката

$$t_0 = \frac{\overline{x} - \mu_0}{\frac{\sigma_0}{\sqrt{n}}}$$

е извън интервала, заключен между квантилите на нормалното разпределение $\left[\Phi(\frac{\alpha}{2}), \Phi(1 - \frac{\alpha}{2})\right]$. Или, което е по-лесно, да проверим дали е вярно твърдението $p - value < \alpha$.

7.4.4 ... с критерий на Уилкоксон

Досега предполагахме, че разпределението на популацията е нормално (гаусово) и проверявахме хипотези за неизвестния параметър μ на това вероятностно разпределение. Данните бяха количествени, измерени в интервалната или относителната скала. Разгледахме и случая, когато популационното разпределение е неизвестно, данните са количествени, но обемът на извадката е достатъчно голям. Тези хипотези и тези методи са известни като *параметрични*. Сега ще предполагаме, че популационното разпределение е неизвестно. Следващите хипотези и методи са известни като *непараметрични*. Сега данните, освен количествени, могат да бъдат и качествени, но измерени в ординалната скала. Тогава медианата дава средната, типичната, очакваната стойност на елементите от популацията.

Пример 11. *(Как да достигнем да нови клиенти в нашия бизнес?)*

Един кинезитерапевт иска да разшири своя бизнес. Той не чете купчината листовки, изпращани му ежедневно по пощата, и не вярва, че това е най-добрия начин за привличане на нови клиенти.

Да напомним, че кинезитерапевтът е здравен специалист, който изследва и оценява функционалното състояние на човека и определя рехабилитационния му потенциал, разработва и провежда ки-

незитерапевтични програми за клинично диагностицирани от лекар заболявания, провежда специализирано лечение, профилактика и рехабилитация.

Нашият човек се запитал "Къде да открия новите си клиенти?" и си спомнил, че повече от пациентите му идвали при него заради болки в гърба. Известно е, че в много от случаите подобни проблеми държат човек буден до късно вечерта. Болката не е чак толкова непоносима, за да се търси неотложна медицинска помощ, но е достатъчно силна за да не може потърпевшият да заспи или да се концентрира върху нещо полезно и затова гледа телевизия да късно.

Кинезитерапевтът решил да изследва до колко часа обикновено са били будни неговите пациенти преди да започнат оздравителните процедури и да пусне поредица от евтини реклами в късните рекламни блокове на местната кабелна телевизия.

Данните от 29 случайно избрани пациенти, попитани в колко часа обикновено заспиват, са:

24, 23, 24, 21, 22, 24, 23, 24, 22, 22, 24, 22, 23, 24, 22, 22, 23, 21, 21, 23, 22, 22, 22, 22, 24, 22, 23, 23, 24

И така, кинезитерапевтът вече е проучил, че рекламите в късните часове са по-евтини и иска да узнае до колко часа неговите потенциални клиенти обикновено са будни.

Когато строим модела на тази задача нямаме основание да направим предположение за нормалност на изследваната популация. Тази задача от практиката се свежда до задачата за проверка на хипотезата за медианата на популацията

$$H_0 : Me = Me_0$$

За да проверим тази хипотеза ще използваме критерия на Уилкоксон, който в R се реализира с функцията $wilcox.test()$.

Алтернативната хипотеза може да бъде една от следните три $H_1 : Me \neq Me_0$, $H_1 : Me < Me_0$ или $H_1 : Me > Me_0$, и в R се определя с параметъра $alternative$ на функцията $wilcox.test()$.

В нашия случай основната хипотеза е $H_1 : Me = 22$, а алтернативата е $H_1 : Me > 22$.

```
> o.clock <- c(24, 23, 24, 21, 22, 24, 23, 24, 22, 22,
24, 22, 23, 24, 22, 22, 23,  21,  21, 23, 22, 22, 22,
22,  24, 22, 23, 23, 24)
> wilcox.test (x=o.clock, mu=22, alternative = "great",
exact = FALSE)

    Wilcoxon signed rank test with continuity correction

data:  o.clock
V = 154.5, p-value = 0.001051
alternative hypothesis: true location is greater than 22
```

Стойността на променливата *p-value* е по-малка от нивото на значимост, което по подразбиране е 0.05. Следователно отхвърляме основната хипотеза в полза на контрахипотезата. Потенциалните клиенти на нашия кинезитерапевт заспиват след 22 часа.

Извод: Потенциалните клиенти на нашия кинезитерапевт обикновено са будни до 22 часа и той спокойно може да рекламира в късните рекламни блокове до 22 часа.

7.5 Проверка на хипотези за отклонението от средното

Нека $X \in N(\mu, \sigma^2)$, където дисперсията σ^2 е неизвестна, е случайната величина, която моделира популацията. Нека $x_1, \ldots, x_n$ са n независими наблюдения над X. Разглеждаме хипотезата

$$H_0 : \sigma^2 = \sigma_0^2$$

срещу една от възможните алтернативи $H_1 : \sigma^2 \neq \sigma_0^2$, $H_1 : \sigma^2 < \sigma_0^2$ или $H_1 : \sigma^2 > \sigma_0^2$.

7.5.1 ... при неизвестно популационно средно

Когато основната хипотеза е вярна, случайната величина $\frac{(n-1)s^2}{\sigma_0^2} \in \chi^2(n-1)$, където $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$. Критерият за проверка на

хипотези за дисперсията на нормална популация с неизвестно средно се основава на статистиката $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$, която при вярна основна хипотеза има χ^2 разпределение с $n-1$ степени на свобода.

7.5.2 ... при известно популационно средно

Когато хипотезата $H_0 : \sigma^2 = \sigma_0^2$ е вярна, случайната величина $\frac{ns_n^2}{\sigma_0^2} \in \chi^2(n)$, където $s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$. Критерият за проверка на хипотези за дисперсията на нормална популация с известно популационното средно μ се основава на статистиката $\chi_n^2 = \frac{ns_n^2}{\sigma_0^2}$, която при вярна основна хипотеза има χ^2 разпределение с n степени на свобода.

Извод: Критериите за проверка на хипотези за дисперсията на нормална популация и при известно и при неизвестно средно се основават на статистики, които имат хи-квадрат разпределение с различни параметри.

7.6 Проверка на хипотези за две популации

7.6.1 Тест на Фишер за равенство на дисперсии

Ще разгледаме следния пример от практиката:

Пример 12. *(Един уникален български продукт) Типичен пример за български принос към здравословно хранене е българското кисело мляко. Технологичните процеси на неговото производство се контролират съгласно система, която осигурява контрол по всички брънки на производствената верига.*

Фирма разработва нова поточна линия, която ще пълни стотици кофички кисело мляко на час.

За да проверят дали производствените стандарти се спазват, специалистите са направили случайна извадка от 30 *кофички и е измерено съдържанието на киселото мляко в тях. Получени са следните данни:*

397, 397, 399, 401, 398, 403, 400, 399, 397, 403, 401, 398, 397, 398, 405, 406, 393, 403, 400, 402, 401, 404, 398, 401, 393, 404, 401, 400, 398, 400

Въпреки, че направените изследвания на специалистите показват, че средното тегло на киселото мляко в кофичката е това, което трябва да бъде по стандарт - 400г., някои от кофичките видимо са по-празни, а други значително по-пълни. Това означава, че дисперсията на случайната величина, моделираща съдържанието на киселото мляко в кофичките, е голяма.

След направени подобрения на поточната линия по случаен начин са избрани 40 кофички и е измерено съдържанието на киселото мляко в тях. Получени са следните данни:

400, 400, 400, 399, 400, 400, 400, 399, 400, 399, 399, 399, 399, 399, 399, 398, 400, 399, 399, 400, 402, 399, 399, 398, 400, 400, 399, 399, 397, 400, 400, 399, 400, 399, 400, 401, 399, 402, 399, 400

Искаме да проверим хипотезата, че дисперсията на случайната величина, моделираща съдържанието на киселото мляко в кофичките вече е по-малка от дисперсията на случайната величина, моделираща съдържанието на киселото мляко в кофичките преди подобренията.

Ще построим статистически модел на задачата.

Нека $X \in N(\mu_x, \sigma_x^2)$ е моделът на съдържанието на киселото мляко в кофичките преди подобренията на поточната линия и $Y \in N(\mu_y, \sigma_y^2)$ е моделът на съдържанието на киселото мляко след подобренията. Двете популационни дисперсии са неизвестни. Освен това X и Y са независими случайни величини.

Искаме да проверим основната хипотеза $H_0 : \sigma_x = \sigma_y$ срещу алтернативната хипотеза $H_1 : \sigma_x > \sigma_y$.

Тези хипотези могат да бъдат записани и в следния еквивалентен вид: $H_0 : \frac{\sigma_x}{\sigma_y} = 1$, $H_1 : \frac{\sigma_x}{\sigma_y} > 1$.

Нека $x_1, \ldots, x_n$ са независими наблюдения над случайната величина X, а $y_1, \ldots, y_m$ са независими наблюдения над случайната величина Y $(n > 1, m > 1)$.

Първо ще разгледаме случая, когато популационните средни μ_x

и μ_y са неизвестни.

Нека $\overline{x}$ и $\overline{y}$ са извадковите средни. Тогава можем да разгледаме неизместените оценки на извадковите дисперсии

$$\widehat{\sigma}_x^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

$$\widehat{\sigma}_y^2 = \frac{\sum_{i=1}^{m}(y_i - \overline{y})^2}{m-1}$$

Естествено е да опитаме да ги приведем до вид, удобен за проверка на хипотезите. За тази цел нормираме с помощта на истинската дисперсия σ_x^2 и σ_y^2, съответно:

$$\frac{(n-1)\widehat{\sigma}_x^2}{\sigma_x^2} = \sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{\sigma_x}\right)^2$$

$$\frac{(m-1)\widehat{\sigma}_y^2}{\sigma_y^2} = \sum_{i=1}^{m}\left(\frac{y_i - \overline{y}}{\sigma_y}\right)^2$$

Известно е, че:

$$\frac{(n-1)\widehat{\sigma}_x^2}{\sigma_x^2} \in \chi^2(n-1)$$

$$\frac{(m-1)\widehat{\sigma}_y^2}{\sigma_y^2} \in \chi^2(m-1)$$

Тогава съгласно **??** частното на двете χ^2 разпределени оценки, делени на степените си на свобода дава F-статистиката

$$F = \frac{\frac{\frac{(n-1)\widehat{\sigma}_x^2}{\sigma_x^2}}{n-1}}{\frac{\frac{(m-1)\widehat{\sigma}_y^2}{\sigma_y^2}}{m-1}} = \frac{\widehat{\sigma}_x^2.\sigma_y^2}{\widehat{\sigma}_y^2.\sigma_x^2} \in F_{n-1,m-1}$$

За да построим критичната област просто трябва да забележим, че при допускане на $H_0 : \sigma_x = \sigma_y$ получаваме, че частното на двете извадкови дисперсии $\frac{\widehat{\sigma}_x^2}{\widehat{\sigma}_y^2} = F \in F_{n-1,m-1}$, докато при алтернативата $H_1 : \sigma_x > \sigma_y$ това частно е по-голям $\frac{\widehat{\sigma}_x^2}{\widehat{\sigma}_y^2} > F \in F_{n-1,m-1}$.

(??? картинка, където да се вижда опашката на F-разпр и критичната област ???)

Това ни позволява да изберем критична област $F > F_{n-1,m-1}(1-\alpha)$ (виж картинка **??**), при която ще можем да отхвърлим H_0.

Този критерий в R се реализира с функцията *var.test*().

По аналогичен начин се показва, че при известни популационни средни μ_x и μ_y и вярна основна хипотеза $H_0 : \sigma_x = \sigma_y$ статистиката

$$F = \frac{s_n^2}{s_m^2} \in F(n, m)$$

където $s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^2$, $s_m^2 = \frac{1}{m}\sum_{i=1}^{m}(y_i - \mu_y)^2$. Статистическата процедура за проверка на хипотези за равенство на дисперсиите на две нормални популации с известни популационни средни се основава на това твърдение.

```
> x <-c(397, 397, 399, 401, 398, 403, 400, 399, 397, 403,
    401, 398, 397, 398, 405, 406, 393, 403, 400, 402, 401,
    404, 398, 401, 393, 404, 401, 400, 398, 400 )
> y <-c(400, 400, 400, 399, 400, 400, 400, 399, 400, 399,
    399, 399, 399, 399, 398, 400, 399, 399, 400, 402,
    399, 399, 398, 400, 400, 399, 399, 397, 400, 400, 399,
    400, 399, 400, 401, 399, 402, 399, 400)
> var.test (x, y, alternative="great")

        F test to compare two variances

data:  x and y
F= 11.261, num df = 29, denom df = 39, p-value = 1.837e-11
alternative hypothesis: true ratio of variances is greater
than 1
95 percent confidence interval:
 6.403221      Inf
sample estimates:
ratio of variances
         11.26095
```

Стойността на променливата *p-value* е по-малка от нивото на значимост, което по подразбиране е 0.05. Следователно отхвърляме основната хипотеза в полза на контрахипотезата.

Извод: Дисперсията на случайната величина, моделираща съдържанието на киселото мляко в кофичките след подобренията е по-

малка от дисперсията на случайната величина, моделираща съдържанието на киселото мляко в кофичките преди подобренията.

7.6.2 Проверка на хипотези за равенство на средните

Когато разпределението на популациите е нормално или обемите на извадките са големи, се използват критерии, базиращи се съответно на разпределението на Стюдент или на нормалното (гаусовото) разпределение. В противен случай се използват непараметрични критерии. (??? тук не сме дали такива?! ???)

Ще разгледаме случая, когато двете популации са нормално разпределени.

Нека $X \in N(\mu_x, \sigma_x)$ и $Y \in N(\mu_y, \sigma_y)$ са две независими случайни величини с неизвестни средни и *известни* дисперсии. Нека $x_1, x_2, \ldots, x_n$ са независими наблюдения над случайна величина X, а $y_1, y_2, \ldots, y_m$ са независими наблюдения над случайна величина Y. Сега нашата цел е да се научим да проверяваме хипотезата $H_0 : \mu_x = \mu_y$. Ако допуснем, че имат равни средни (т.е. хипотезата H_0 е вярна), то статистиката

$$t = \frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \in N(0,1)$$

Когато μ_x расте спрямо μ_y, t-статистиката ще се увеличава, тъй като извадковото средно $\overline{X}$ е неизместена оценка за нея и като такава "се стреми да я следва". Аналогично при обратната ситуация. Това ни дава право да използваме двустранна критична област.

На това твърдение се основава проверката на хипотези за равенство на средните на две нормални популации с известни дисперсии.

Нека сега $X \in N(\mu_x, \sigma_x)$ и $Y \in N(\mu_y, \sigma_y)$ са независими случайни величини с неизвестни средни и *неизвестни*, но равни дисперсии.

Тогава статистиката

$$t = \frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}\left(\frac{1}{n} + \frac{1}{m}\right)}} \in t_{n+m-2}$$

Когато μ_x расте спрямо μ_y, t-статистиката ще се увеличава, тъй като извадковото средно $\overline{X}$ е неизместена оценка за нея и като такава

"се стреми да я следва". Аналогично при обратната ситуация. Това ни дава право да използваме двустранна критична област.

На това твърдение се основава проверката на хипотези за равенство на средните на две нормални популации с неизвестни, но равни дисперсии.

Когато популационните дисперсии са неизвестни можем да използваме функцията $t.test()$.

7.6.3 ... на две популации с произволни вероятностни разпределения

Нека X и Y са две независими случайни величини.

Известно е, че когато обемите на извадките са достатъчно големи, извадковите средни имат приблизително нормално разпределение, т.е. $\overline{X} \sim N\left(\mu_x, \frac{\sigma_x^2}{n}\right)$ и $\overline{Y} \sim N\left(\mu_y, \frac{\sigma_y^2}{m}\right)$. Тогава $\overline{X} - \overline{Y} \sim N\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right)$ и следователно

$$t = \frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \sim N(0,1)$$

При вярна хипотеза $H_0 : \mu_x = \mu_y$ статистиката t има приблизително стандартно нормално разпределение, докато при силно отклонение от това допускане $\overline{X} - \overline{Y}$ ще расте или намалява, понеже е неизместена оценка на очакването на разликата на двете случайни величини. Това ни дава право да използваме двустранен критерий за проверка на хипотезата, с критична област $t \in \left(-\infty, \Phi(\frac{\alpha}{2})\right] \cup \left[\Phi(1 - \frac{\alpha}{2}, \infty\right)$.

Следва да отбележим, че горната статистика само асимптотично клони към стандартно нормалното разпределение. На практика е нужно и достатъчно размерът на извадките да е по-голям от 30.

Когато разпределението на популациите не е нормално и дисперсиите не са известни можем да използваме функцията $wilcox.test()$, която реализира съответния непараметричен критерий.

7.6.4 Проверка на хипотези за равенство на средните на две популации с R

(??? това да се замести с пример, който проверява всички хипотези и се вкаран по-горе в текста! ???)

7.7 Връзка между доверителни интервали и проверка на хипотези

(??? тук е добре да се сложат 1 - картинка с демонстрация на връзката 2 - по-подробно описание 3 - връзка към главата доверителни интервали ???)

7.8 Непараметрични методи ???

7.8.1 Критерий на Уилкоксон при една извадка

7.8.2 Критерий на Ман-Уитни-Уилкоксон при две независими извадки

7.8.3 Критерий на Ансари – Брадли

7.8.4 Критерий на Колмогоров-Смирнов

7.9 Проверка на хипотези за равенство на средните на няколко популации

7.9.1 Задача

Пример 13 (Пример от екологията)**.** *Проучено е съдържанието на тежките метали в органите на сладководните риби, живеещи в силно, антропогенно и химично, замърсените води по поречието на река Арда. Водата, като добър разтворител, е сложен разтвор от различни вещества. Химическият й състав се определя от минерализацията на почвата в местата, където водата извира, тече или*

се събира. Силно е и влиянието на естествените рудни залежи, характерни за този район. Други замърсители са производствените и битови отпадъци.

Задачата е да се отговори на въпроса: Видът риба влияе ли върху съдържанието на тежките метали в органите на сладководните риби? Данните са получени от Велчева, И. Г., 1998. Екологично проучване на съдържанието на кадмий (Cd), олово (Pb) и цинк (Zn) в уклейка (Alburnus alburnus L), шаран (Cyprinus carpio L.) и костур (Perca fluviatilis L.) от язовир "Кърджали" и "Студен кладенец" – автореферат на дисертация за получаване на научна и образователна степен "доктор", Пловдивско университетско издателство, 36 стр. с любезното съдействие на автора.

7.9.2 Модел

Нека имаме b независими нормално разпределени случайни величини $X_1, X_2, ..., X_b$ с неизвестни средни и равни дисперсии, т.е. $X_j \in N(\mu_j, \sigma^2)$, $j = 1, 2, ..., b$. Нека $X_{1j}, X_{2j}, ..., X_{aj}$ е случайна извадка с обем a от j-тото разпределение, т.е. $X_{ij} \in N(\mu_j, \sigma^2), i = 1, 2, ..., a$. Искаме да проверим хипотезата

$H_0 : \mu_1 = \mu_2 = ... = \mu_b = \mu$, където μ не е определено, срещу алтернативата

H_1 : Някои от популационните средни са различни
с ниво на съгласие α.

Ще предположим, че основната хипотеза е вярна и ще намерим максимално правдоподобните оценки на неизвестните параметри. Функцията на правдоподобие има вида:

$$L(x, \mu, \sigma^2) = (\frac{1}{\sqrt{2\pi\sigma^2}})^{ab} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{a}\sum_{j=1}^{b}(x_{ij}-\mu)^2}$$

Логаритмуваме, диференцираме по двата неизвестни параметъра на популационното разпределение и получаваме:

$$\frac{dlnL}{d\mu} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}(x_{ij} - \mu)}{\sigma^2} = 0$$

Следователно

$$\widehat{\mu} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b} x_{ij}}{ab}$$

Знаем, че извадковото средно е

$$\overline{x} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b} x_{ij}}{ab}$$

Така получаваме

$$\widehat{\mu} = \overline{x}$$

От

$$\frac{dlnL}{d\sigma^2} = -\frac{ab}{\sigma^2} + \frac{1}{\sigma^4}\sum_{i=1}^{a}\sum_{j=1}^{b}(x_{ij}-\mu)^2 = 0$$

следва

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}(x_{ij}-\overline{x})^2}{ab}$$

Следователно намерихме максимално правдоподобните оценки за неизвестното популационно средно μ и неизвестната популационна дисперсия σ^2.

Предполагаме, че е вярна контрахипотезата H_1. Ще намерим максимално правдоподобните оценки на неизвестните популационни средни $\mu_j, j = 1, 2, ..., b$. В този случай функцията на правдоподобие има вида:

$$L' = (\frac{1}{2\pi\sigma^2})^{\frac{ab}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{a}\sum_{j=1}^{b}(x_{ij}-\mu_j)^2}$$

Логаритмуваме, диференцираме по неизвестния параметър μ_j и получаваме

$$\frac{dlnL'}{d\mu_j} = \frac{\sum_{i=1}^{a}(x_{ij}-\mu_j)}{\sigma^2} = 0$$

Получаваме

$$\mu_j = \frac{\sum_{i=1}^{a} x_{ij}}{a}$$

Следователно

$$\widehat{\mu_j} = \frac{\sum_{i=1}^{a} x_{ij}}{a} = \overline{x}_{.j}$$

е максимално правдоподобната оценка на средното μ_j на j-тото разпределение.

Сега ще диференцираме по втория неизвестен параметър σ^2.

$$\frac{dlnL'}{d\sigma^2} = -\frac{ab}{\sigma^2} + \frac{1}{\sigma^4}\sum_{i=1}^{a}\sum_{j=1}^{b}(x_{ij}-\mu_j)^2 = 0$$

Получаваме

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}(x_{ij}-\overline{x}_{.j})^2}{ab}$$

Получихме втора максимално правдоподобна оценка на неизвестната дисперсия σ^2.

Въвеждаме следните обозначения:

$Q = \sum_{i=1}^{a}\sum_{j=1}^{b}(x_{ij}-\overline{x})^2$, $Q_1 = \sum_{i=1}^{a}\sum_{j=1}^{b}(x_{ij}-\overline{x}_{.j})^2$ и $Q_2 = a\sum_{j=1}^{b}(\overline{x}_{.j}-\overline{x})^2$.

За тези статистики са в сила твърденията: $Q_1 \perp Q_2$, $\frac{Q_1}{\sigma^2} \in \chi^2(b(a-1))$, $\frac{Q_2}{\sigma^2} \in \chi^2(b-1)$. Тогава

$$F = \frac{\frac{Q_2}{\sigma^2(b-1)}}{\frac{Q_1}{\sigma^2(b(a-1))}} = \frac{\frac{Q_2}{b-1}}{\frac{Q_1}{b(a-1)}} \in F_{(b-1,b(a-1))}$$

Следователно критерият за проверка на сложната хипотеза H_0 : $\mu_1 = \mu_2 = ... = \mu_b = \mu$, когато μ не е определено, срещу всички възможни алтернативи, може да се основава на една F статистика.

От глава ?? раздел ?? знаем, че популационната дисперсия σ^2 може да бъде оценена с една от извадковите дисперсии $s_j^2 = \frac{1}{a-1}\sum_{i=1}^{a}(x_{ij}-\overline{x}_{.j})^2$.

Популационната дисперсия σ^2 може да бъде оценена и със средното на извадковите дисперсии ???

$s_W^2 = \sum_{j=1}^{b}\frac{s_j^2}{b} = \frac{1}{b(a-1)}\sum_{j=1}^{b}\sum_{i=1}^{a}(x_{ij}-\overline{x}_{.j})^2 = \frac{1}{b(a-1)}\sum_{j=1}^{b}\sum_{i=1}^{a}(x_{ij}-\overline{x}_{.j})^2 = \frac{Q_1}{b(a-1)}$.

Забележете, че $Es_W^2 = \sum_{j=1}^{b}\frac{Es_j^2}{b} = \frac{b\sigma^2}{b} = \sigma^2$, защото извадковата дисперсия s_j^2 е неизместена оценка на σ^2. Знаем, че s_j^2 е и състоятелна оценка на σ^2. Следователно s_W^2 е една неизместена оценка на популационната дисперсия σ^2. Оценката s_W^2 се нарича *вътрешногрупова извадкова дисперсия.*

Статистиката

$$s_B^2 = \frac{a}{b-1}\sum_{j=1}^{b}(\overline{x}_{.j} - \overline{x})^2 = \frac{Q_2}{b-1}$$

също оценява неизвестната популационна дисперсия σ^2 ??? и се нарича *междугрупова извадкова дисперсия.*

Ако основната хипотеза H_0 е вярна, вътрешногруповата извадкова дисперсия s_W^2 и междугруповата извадкова дисперсия s_B^2 ще имат близки стойности, защото и двете оценяват неизвестната популационна дисперсия σ^2. Тяхното отношение ще бъде близко до 1. Това е причината нашият критерий да се основава на отношението на междугруповата извадкова дисперсия и вътрешногруповата извадкова дисперсия.

Ако основната хипотеза H_0 не е вярна, междугруповата извадкова дисперсия s_B^2 се очаква да бъде по-голяма от вътрешногруповата извадкова дисперсия s_W^2. Следователно ще отхвърлим основната хипотеза H_0 с ниво на значимост α, ако отношението на дисперсиите надвишава критичната стойност $F_{(b-1,n-b)}(1-\alpha)$ на F разпределението, където b е броя на групите, а $n = ab$ е обема на извадката.

Забележете, че разгледахме случая когато взимаме случайни извадки с равен обем a от всяко нормално разпределение. Извадките могат да имат различни обеми. В този случай моделът се строи по аналогичен начин.

7.9.3 Дисперсионен анализ

Дисперсионният анализ е част от статистиката, изучаваща влиянието на една или няколко групиращи променливи върху една количествена променлива. В дисперсионния анализ е възприето тази зависима количествена променлива да се нарича отклик, групиращата променлива да се нарича фактор, а стойностите ѝ - нива на фактора. Отклоненията на средните стойности на групата от общата средна $\alpha_j = \mu_j - \mu$ се наричат ефекти. Дисперсионният анализ разглежда следния модел:

$$y_{ij} = \mu_j + \epsilon_{ij}, i = 1, 2, ..., a, j = 1, 2, ..., b.$$

където μ_j са средните на групите, а ϵ_{ij} са случайни грешки, т.е. $\epsilon_{ij} \in N(0, \sigma^2)$. Следователно b независими случайни величини, които

имат нормални разпределения с неизвестни средни $\mu_1, \mu_2, ..., \mu_b$ и неизвестна, но еднаква дисперсия σ^2, моделират групите. Този модел е еквивалентен на модела $y_{ij} = \mu + alpha_j + \epsilon_{ij}$, където μ е генералното средно, е α_j е j-тия ефект, а ϵ_{ij} са случайните грешки. Нашата цел е да сравним наблюдаваните извадкови средни. Ако те са близки, тогава разликите могат да бъдат приписани на случайни фактори. Ако наблюдаваните извадкови средни са значимо различни, то има причина да вярваме, че влиянията на различните групи са значими, т.е. ефектите са значими. Проблемът може да бъде формулиран като проверка на хипотези по следния начин: Да се провери хипотезата:

$H_0 : \mu_1 = \mu_2 = ... = \mu_b = \mu$, когато μ не е определено, срещу алтернативата

H_1 : Някои от популационните средни са различни
Следователно, основната задача, която се решава с помощта на дисперсионния анализ, може да се формулира най-просто така: Да се провери хипотезата дали средните стойности на отклика в няколко различни групи от наблюдения съвпадат. Тази задача разгледахме в предишния параграф. Ако тази хипотеза се отхвърли, необходимо е да се оценят различните средни стойности за всяка група.

Прието е резултатите от дисперсионния анализ да се представят в така наречените таблици на дисперсионния анализ *ANOVA (ANalysis Of VAriance)*. В таблицата на еднофакторния дисперсионен анализ първата колонка има име `Df` и съдържа степените на свобода $b - 1$ и $b(a - 1)$, съответно на статистиките Q_2 и Q_1. В колонката с име `Sum Sq` са дадени стойностите на Q_2 и Q_1. Колонката `Mean Sq` показва стойността на междугруповата дисперсия s_B^2 и стойността на вътрешногруповата дисперсия s_W^2. Колонката `F value` показва стойността на F статистиката. Числото `Pr(>F)` е вероятността да отхвърлим основната хипотеза H_0, когато тя е верна. Малките стойности на вероятността `Pr(>F)` показват, че има разлика между груповите средни. Знаем, че нивото на значимост α ни дава максималната вероятност с която може да си позволим да направим грешката да отхвърлим основната хипотеза H_0 когато тя е верна. Затова сравняваме `Pr(>F)` с α. Ако е изпълнено `Pr(>F) <` α, то разликата между груповите средни е значима и затова отхвърляме основната хипотеза H_0 в полза на алтернативата H_1. Ако това не е вярно, нямаме основание да отхвърлим основната хипотеза и я приемаме за вярна.

Една алтернативна параметрична процедура на anova(), която не изисква равенство на дисперсиите на групите е oneway.test().

Ако основното предположение за нормалност на разпределението на отклика е нарушено, то вместо метода на дисперсионния анализ се прилага неговият непараметричен аналог, известен като *Kruskal-Walis тест*. В *R* този тест се реализира с функцията *kruskal.test*(). Предположението за равенство на дисперсиите може да се провери с *критерия на Бартлет*, за който вече знаем, че се реализира с функцията *bartlett.test*().

Проверката на хипотезата, че разпределението на една променлива има една и съща дисперсия във всички групи, може да се извърши с критерия Бартлет. Този критерий не е робастен при отклонения от предположението за нормалност на популационното разпределение, т.е. при отклонения от предположението за нормалност на популационното разпределение критерият не дава добри резултати.

Винаги предполагаме, че групите са независими помежду си.

7.9.4 Дисперсионен анализ с R

Функцията *anova*() реализира дисперсионен анализ в *R*. Прочитаме данните и отпечатваме само 2 реда от таблицата с данните.

```
> z <- read.table ("F:\\Heavy_Metals.txt", header=TRUE,
                   dec=".")
> head (z, n = 2L)
     species Kidney_Pb Liver_Pb Gills_Pb Skin_Pb Bones_Pb
1 A.alburnus    13.250    7.489    1.889   1.761    1.775
2 A.alburnus    23.252   10.447    2.090   2.000    1.906
  Muscle_Pb Kidney_Cd Liver_Cd Gills_Cd Skin_Cd Bones_Cd
1     1.647    9.245    2.105    1.302   1.204    0.241
2     1.701   10.201    2.541    1.368   1.299    0.579
  Muscle_Cd Kidney_Zn Liver_Zn Gills_Zn Skin_Zn Bones_Zn
1     0.298   128.409   83.070    57.46   42.56   54.709
2     0.718   682.586  558.551    66.12   46.02   79.470
  Muscle_Zn
1     7.897
2    23.233
```

Променливата *species* дава информация за биологичния вид риба. В нашия анализ тази променлива ще играе ролята на фактор с

три нива (*A.alburnus*, *C.carpio* и *P.fluviatilis*).

```
> species <- as.factor (z[,1])
```

Променливата *heavy.metal* е матрица, която съдържа информация за съдържанието на трите метала в органите на рибите. Отпечатваме само първите 2 реда от тази матрица.

```
> heavy.metal <- as.matrix (z [,2:19])
> head (heavy.metal, n = 2L)
     Kidney_Pb Liver_Pb Gills_Pb Skin_Pb Bones_Pb Muscle_Pb
[1,]    13.250    7.489    1.889   1.761    1.775     1.647
[2,]    23.252   10.447    2.090   2.000    1.906     1.701
     Kidney_Cd Liver_Cd Gills_Cd Skin_Cd Bones_Cd Muscle_Cd
[1,]     9.245    2.105    1.302   1.204    0.241     0.298
[2,]    10.201    2.541    1.368   1.299    0.579     0.718
     Kidney_Zn Liver_Zn Gills_Zn Skin_Zn Bones_Zn Muscle_Zn
[1,]   128.409   83.070    57.46   42.56   54.709     7.897
[2,]   682.586  558.551    66.12   46.02   79.470    23.233
```

В променливата *Pb.in.Muscle* съхраняваме съдържанието на оловото в мускулите на рибите. Съдържанието на оловото в мускулите зависи от много фактори, всеки от които има малко влияние. Затова предположението за нормалност на популацията е коректно.

```
> Pb.in.Muscle <-heavy.metal[,6]
```

Ще проверим хипотезата, че средното съдържание на олово в мускулите на плантоноядните риби, растителноядните риби и рибите хищници е едно и също, при ниво на значимост $\alpha = 0.05$.

```
> anova (lm (Pb.in.Muscle ~ species))
Analysis of Variance Table

Response: Pb.in.Muscle
          Df  Sum Sq Mean Sq F value  Pr(>F)
species    2  1.9515  0.9758  3.4713 0.03660 *
Residuals 69 19.3953  0.2811
---
Signif. codes:0'***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

В представената таблица на дисперсионния анализ колонката `Df` съдържа степените на свобода $b-1 = 2$ и $b(a-1) = 3(24-1) = 69$, съответно на статистиките Q_2 и Q_1. В колонката с име `Sum Sq` са дадени

стойностите $Q_2 = 1.9515$ и $Q_1 = 19.3953$. Колонката `Mean Sq` показва стойността на междугруповата дисперсия $s_B^2 = 0.9758$ и стойността на вътрешногруповата дисперсия $s_W^2 = 0.2811$. Стойността на F статистиката е $F = 3.4713$. Вероятността `Pr(>F)=0.03660` сравняваме с нивото на значимост $\alpha = 0.05$. Изпълнено е `Pr(>F)<`α и затова отхвърляме основната хипотеза H_0 в полза на алтернативната H_1. Разликата между груповите средни е значима.

Резултатът от приложения дисперсионен анализ е, че отхвърляме хипотезата за равенство на средното съдържание на оловото в мускулите на различните видове риби.

Извод: Факторът *вид риба* влияе върху съдържанието на олово в мускулите на рибите.

В традиционния еднофакторен дисперсионен анализ е в сила предположението за равенство на дисперсиите във всички групи. Ще проверим условието за хомогенност на дисперсията при ниво на значимост $\alpha = 0.05$:

```
> bartlett.test (Pb.in.Muscle ~ species)

        Bartlett test of homogeneity of variances

data: Pb.in.Muscle by species
Bartlett's K-squared = 5.7465, df = 2, p-value = 0.05652
```

Следователно нямаме основание да отхвърлим хипотезата за равенство на груповите дисперсии. Изводът от дисперсионния анализ е коректен.

Един непараметричен аналог на one-way процедурата е критерият на Kruskal-Wallis (Kruskal-Wallis rank sum test), който в R се реализира с командата kruskal.test().

При предположение за нормалност със следващите команди на R може да се провери влиянието на фактора вид риба върху съдържанието на всички цветни метали, за които имаме данни.

```
alpha <- 0.05
for (i in 1:18)   {
print (i)
print (bartlett.result <- bartlett.test
              ( heavy.metal[,i] ~ species ))
if (bartlett.result$p.value < alpha)
```

```
{print (oneway.test (heavy.metal[,i] ~ species))
} else {
print ( anova ( lm (heavy.metal[,i] ~ species)))
}
}
```

Глава 8

Методи за прогнозиране

Прогнозирането е ключов елемент от процеса на вземане на управленски решения.

8.1 Регресионен анализ

Понятието регресия заема централно място в апарата на статистическото изследване на зависимостите между количествените променливи.

Според начина, по който се изследва закономерността на проявлението на връзките, моделите се разглеждат като линейни и нелинейни.

Линейният регресионен анализ е един от най-популярните и най-често използвани методи за изследване на зависимостите между явленията от действителността. Той най-често намира приложение за прогнозиране стойностите на количествени променливи, но освен това може да се използва за откриване и изследване на връзките между тях.

8.1.1 Задача

Пример 14 (Зехтинът и незаменимите мастни киселини)**.** *Зехтинът е растителна мазнина, извличана от плодовете на маслинови-*

те дървета (Olea europaea). Тя се използва предимно като здравословна храна, но също и в козметиката и сапуните. Доказано полезна е за организма. Зехтинът е чудесен пример за продукт, който съдържа някои незаменими или есенциални мастни киселини, които не се произвеждат в тялото, а сме длъжни да набавим чрез храната. Едни от тях са палматиновата и олеиновата киселина. Палматиновата киселина е омега-7 ненаситена мастна киселина. Олеиновата киселина е мононенаситена омега-9 мастна киселина, която освен в някои растения се съдържа и в организмите на някои животни.

Съгласно проведени научни изследвания, омега 9 мастните киселини спомагат за поддържане на доброто здравословно състояние, забавят развитието на атеросклероза и допринасят за правилното функциониране на мозъка и сърдечносъдовата система. Омега 7 мастните киселини са естествен компонент на кожата. Известно е, че полезна за организма е балансираната комбинация от омега мастни киселини.

Процентното съдържание на тези ненаситени мастни киселини, открити в 572 липидни фракции на италиански зехтин (% x100) е дадена на Интернет страницата http://www.ggobi.org/book/.

Нашата цел е да прогнозираме съдържанието на палматиновата киселина в зехтин, за които знаем съдържанието на олеиновата киселина. За да я постигнем ще построим адекватен модел, описващ съдържанието на палматиновата киселина в зехтина като функция на олеиновата киселина.

8.1.2 Модел

Регресия се нарича всяка комбинация между няколко линейно независими базисни функции от обясняващи променливи $X^{(i)}$ с неизвестни параметри β_i. Обясняващите променливи $X^{(i)}$ (??? дали е най-доброто обяснение? ???) и параметрите β_i не са случайни величини. Обясняемата променлива Y е случайна величина.

Регресионният анализ търси отговор на следните въпроси:

- Стойностите на отклика влияят ли се от предикторите?
- Каква е функционалната връзка между стойностите на променливите? *Или*: Може ли да се избере модел на зависимостта и да се оценят параметрите му?
- Адекватен ли е моделът? *Или*: Доколко получената връзка отговаря на действителност?
- Какво може да очакваме от отклика при зададени други стойности на предикторите? *Или*: Да прогнозираме!

Според броя на включените в анализа обясняващи променливи, които играят ролята на фактори, регресионните модели са еднофакторни и многофакторни. При еднофакторните се изследва връзката само между две явления Y и X. Най-често Y се приема като зависима променлива, а X като обясняваща променлива. На езика на регресионния анализ независимата променлива X се нарича предиктор, а зависимата променлива Y се нарича отклик.

Чрез многофакторните регресионни модели се изследва връзката между едно явление и други две или повече явления, в ролята на фактори. (??? po-dobre da go opredelq???) Общият вид на тези модели се представя с формулите:

- Еднофакторен: $Y_i = f(X_i, \epsilon_i)$
- Многофакторен: $Y_i = f(X_i^{(i)}, X_i^2, \ldots, X_i^k, \epsilon_i)$

След като построим модела (т.е. изберем стойности на β_i, които съгласно конвенцията означаваме с $\widehat{\beta_i}$) можем да правим предвиждане за стойностите на бъдещи експерименти, като построим оценка за Y, която ще означим с $\widehat{Y}$ и ще пресмятаме като $\widehat{Y} = \widehat{f}(X)$.

Разликите между реално наблюдаваните стойности за Y и съответните оценени от модела стойности на Y се наричат остатъци (residuals) . В модела те са случайно разпределени и независими помежду си. Колкото по-малки са стойностите на остатъците, толкова по-добре полученият регресионен модел описва наблюдаваните стойности.

В случая на проста линейна регресия моделът е еднофакторен и може да се представи с формулата:

$$Y_i = \beta_0 + \beta_i X_i + \epsilon_i$$

където:

- β_i са коефициенти на модела,
- ϵ_i е грешката. Тя е случаен компонент с нормално разпределение и $\mathbb{E}\epsilon_i = 0$, т.е. $\epsilon_i \in N(0, \sigma^2)$. Т.е. ще предполагаме, че грешките от наблюденията са независими, еднакво разпределени гаусови случайни величини с нулево очакване.
- Y_i е зависимата променлива, наречена *отклик*.
- X_i е независима променлива, наречена *предиктор*.

Тогава прогнозата може да се представи във вида

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1$$

където β_0, β_1 са оценки на коефициентите на модела.

Параметрите на простия линеен регресионен модел са коефициентите β_0, β_1 и дисперсията на грешката σ^2.

На фигура **??** е нарисувана апроксимиращата права при прост линеен модел $y = \beta_0 + \beta_1 x + \epsilon$.

Все още остава открит въпроса дали коефициентите са различни от нула, т.е. дали са значими. Доказва се, че оценките на параметрите са неизместени, гаусово разпределени и когато ??? , те стават независими. Това ни дава възможност лесно да строим доверителни интервали и да проверяваме хипотези за параметрите на модела.

Нека да въведем следните означения:

$$SS_x = \sum \left(X_i - \overline{X}\right)^2 = \sum X_i^2 - \frac{\left(\sum X_j\right)^2}{n}$$

$$SS_y = \sum \left(Y_i - \overline{Y}\right)^2 = \sum Y_i^2 - \frac{\left(\sum Y_j\right)^2}{n}$$

$$SS_{xy} = \sum \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) = \sum X_i Y_i - \frac{\left(\sum Y_j\right)\left(\sum X_j\right)}{n}$$

Тогава *извадковият корелационен коефициент* r се определя като

$$r = \frac{SS_{xy}}{\sqrt{SS_x}\sqrt{SS_y}}$$

По стойностите на корелационния коефициент се съди за наличието или отсъствието на корелация между две случайни величини. За тази цел се проверява хипотезата за коефициента на корелация ρ:

$H_0 : \rho = 0$ т.е. X и Y са корелационно независими ??? некорелирани

срещу алтернативата

$$H_1 : \rho \neq 0 \text{ т.е. между } X \text{ и } Y \text{ има корелация}$$

при ниво на значимост α.

За проверка на тези хипотези се прилага t-статистиката:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Критичната област е двустранна, ограничена от точките $\pm t_{\text{кр}}$, където $t_{\text{кр}}$ е квантил от разпределението на Стюдънт с $n-2$ степени на свобода и от порядък $q = 1 - \frac{\alpha}{2}$. Ако $|t| > t_{\text{кр}}$, се отхвърля основната хипотеза и се прави извод, че между величините съществува корелационна зависимост.

Коефициентът β_i показва каква е степента на влияние на X върху Y, т.е. с колко единици се променя Y при промяна на X с една единица.

$$\widehat{\beta_i} = \frac{SS_{xy}}{SS_y} = r\frac{S_y}{S_y}$$

$$\widehat{\beta_0} = \overline{y} - \widehat{\beta_1}\overline{x}$$

Частното

$$r^2 = \frac{(SS_{xy})^2}{\sqrt{SS_x}\sqrt{SS_y}}$$

се нарича *коефициент на детерминация*, който по определение е квадратът на извадковия корелационен коефициент.

Колкото r^2 е по близко до 1, толкова по-близко до линейна е зависимостта и толкова наблюдаваните стойности на Y са по-близки до стойности на $\widehat{Y}$ от регресионното уравнение, т.е. колкото коефициентът на детерминация е по-близко до единицата, толкова по-детерминиран е моделът.

Стандартната грешка се бележи с s_e и е мярка за отклоненията на пресметнатите по модела стойности на зависимата променлива от реално наблюдаваните ѝ стойности т.е.

$$s_e = \sqrt{\frac{\sum (y_i - \widehat{y})^2}{n-2}} \text{ е стандартната грешка.}$$

При построяването на линейния регресионен модел се проверяват следните хипотези:

$$H_0 : \beta_0 = 0 \text{ срещу } H_1 : \beta_0 \neq 0$$

$$H_0 : \beta_1 = 0 \text{ срещу } H_1 : \beta_1 \neq 0$$

Използва се статистиката

$$T_{\beta_i} = \frac{\widehat{\beta_i}}{s_{\beta_i}} \text{ за съответните } i = 1, 2$$

където

$$s_{\beta_0} = \sqrt{\sigma_2^2 \frac{\sum x_i^2}{n \sum (x_i - \overline{x})^2}}$$

$$s_{\beta_1} = \sqrt{\sigma_2^2 \frac{1}{n \sum (x_i - \overline{x})^2}}$$

$$\sigma_2^2 = \frac{\sum (y_i - \widehat{y})^2}{n-p}$$

n е обемът на извадката, а p е броят на параметрите на модела.

Тази статистика има разпределение на Стюдънт с $n-2$ степени на свобода. Критичната област и в двата случая е двустранна, ограничена от точките $\pm t_{\text{кр}}$. Ако $|T| > t_{\text{кр}}$, основната хипотеза се отхвърля в полза на алтернативната и се прави извод, че коефициентът е значим. Извод за значимост на коефициентите на модела може да се направи и като се използва p-стойността $Pr(> |t|)$. Коефициентът е значим, ако $Pr(> |t|) < \alpha$.

Полезна информация за адекватността на модела може да бъде получена от остатъците, които представляват разликите между наблюдаваните и оценените стойности на зависимата променлива y:

$e_1 = y_1 - \widehat{y}_1, e_2 = y_2 - \widehat{y}_2, ..., e_n = y_n - \widehat{y}_n$ Коефициентите на модела са оценени така, че сумата от квадратите на остатъците да бъде колкото се може по-малко число.

Един от най-популярните методи за проверка за адекватност на модела е *анализът на остатъците*. Обикновенно този анализ се извършва с графични средства. Две общоприети проверки се реализират със следните две графични представяния на модела: графиката на остатъците и предсказаните стойности, и нормалната графика.

При верен (адекватен) модел остатъците са независими и имат гаусово разпределение.

Ако остатъците са независими, върху графиката на остатъците и предсказаните стойности трябва да се визуализира облак от случайно разпръснати точки.

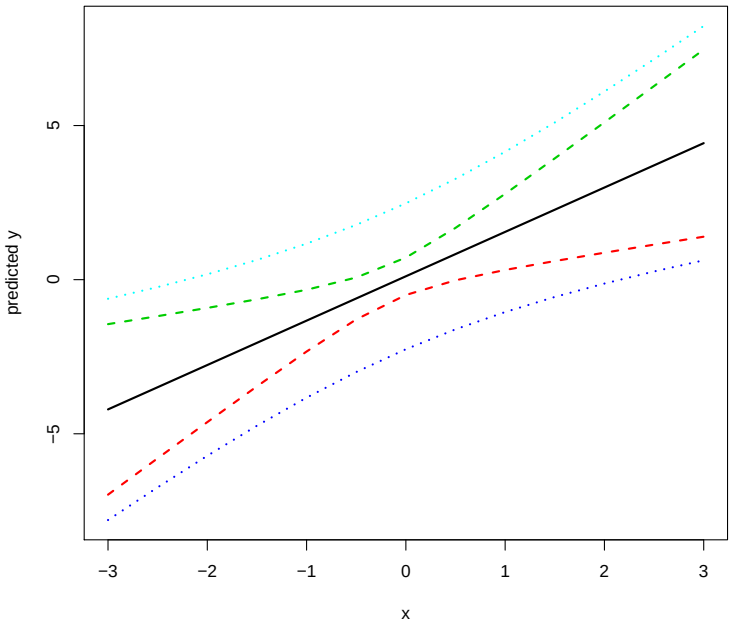
Най-възприетата и лесна проверка за нормалност на остатъците е нормалната графика. Ако остатъците са нормално разпределени, те трябва да лежат на една права. Тежките или прекалено леките опашки се визуализират в s-образно изменение от диагоналната права, която би се получила при нормално разпределение на данните.

С начертаване на хистограмата на остатъците при значителен брой наблюдения лесно се забелязват както отклонения от нормалността, така и наличие на големи остатъци. Може да се приложи и тест за нормалност.

Интерпретацията на тези графики не е лесна и изисква известен опит, който се натрупва в процеса на анализ на различни данни.

За произволни стойности X на предикторите от областта, за която е верен простият регресионен модел, случайната величина $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X$ е неизместена оценка на $\mathbb{E}Y$ и нейната дисперсия $\mathbb{D}\widehat{Y}$ може да се пресметне. По формулата за $\mathbb{D}\widehat{Y}$ се чертае коридорът за модела. Границите на коридора на модела и доверителните граници за наблюдаваната стойност са параболи. Това се вижда от фигура 8.1.

Това означава, че при силно отличаващи се стойности на X от наблюдаваните до момента реалните стойности на Y могат да се различават твърде много от предвижданията $\widehat{Y}$. Така се вижда, например, колко опасни, а понякога и безсмислени, могат да бъдат прогнози за далечното бъдеще, основани на тенденция, наблюдавана в малък интервал от време.



Фигура 8.1: Граници на коридора на линеен модел и доверителни граници за наблюдаваните стойности

8.1.3 Линейна регресия с R

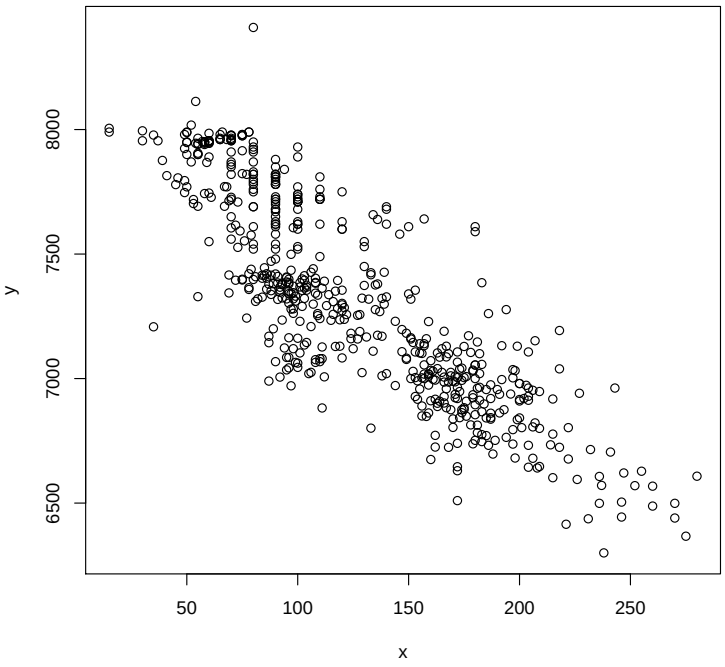
На променливата x ще присвоим данните за съдържанието на олеиновата киселина в липидна фракция на италиански зехтин, а с y - данните за съдържанието на палматиновата киселина.

```
> x <- oleic
> y <- palmitoleic
```

Данните, които биха могли да се опишат с регресионни модели, могат да се визуализират с функцията *plot*(). Получената графика може да ни окуражи в построяването на линеен модел, може да разкрие особености на данните.

Със следващата команда ще визуализираме данните в графичен прозорец:

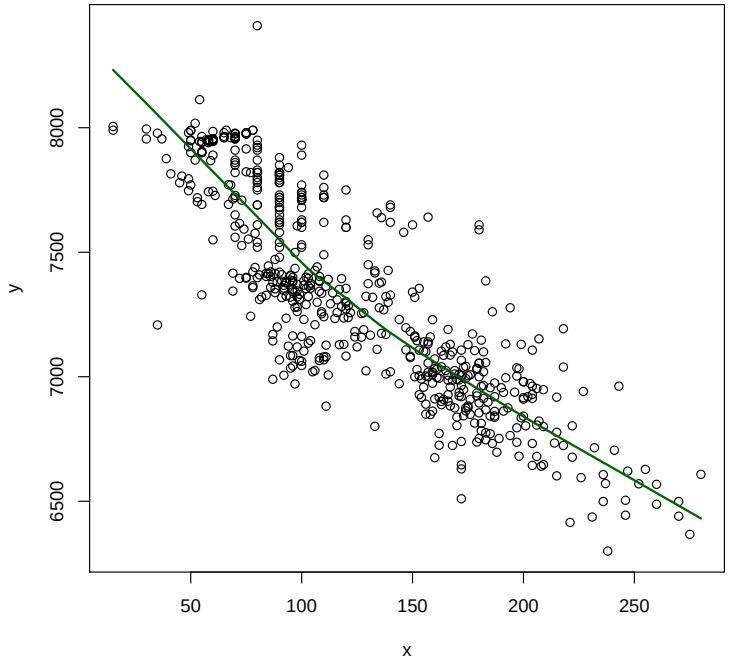
```
> plot(x,y)
```



Фигура 8.2: Данни

Обикновено променливата x се визуализира върху абцисната ос на координатната система и в модела играе роля на предиктор. Променливата y се визуализира върху ординатната ос, а в модела е в ролята на отклика, т.е. на зависимата променлива, чиито стойности искаме да прогнозираме. С крива ще визуализираме зависимостта между тези две променливи, чиито стойности сме наблюдавали. Кривата сред облака от данни във фигура 8.3 в действителност е една оценка на тази зависимост.

```
> plot(x,y); lines (lowess(x,y), col="dark green", lwd=2)
```



Фигура 8.3: Крива, оценяваща зависимостта между отклика и предиктора

От графиката на фигура 8.3 възникват следните въпроси: *Построената крива близка ли е до хипотетична права, която добре би описвала данните? Има ли смисъл да търсим линеен модел, с който успешно ще прогнозираме?*

За да си изясним по-добре ситуацията със съществуването на линеен модел, който описва данните, ще пресметнем извадковия корелационен коефициент :

```
> cor (x,y)
[1] -0.8524384
```

Абсолютната стойност на корелационния коефициент е близка до 1 и следователно може да се започне да се строи линеен модел.

Забележете, че ролята на двете киселини, палматиновата (palmitoleic) и олеиновата (oleic), представени съответно с променливите y и x, не е симетрична. Ние сме заинтересовани да предскажем стойностите на променливата y при фиксирана стойност на променливата x. Линейният регресионен модел ще има вида:

$$y = \alpha + \beta x + \epsilon$$

Компонентата $\alpha + \beta x$ е детерминистичната (не случайна) компонента на модела, а ϵ представя случайната компонента (грешката). Дадените наблюдения $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ можем да изразим като:

$$y_i = \alpha + \beta x_i + \epsilon_i$$

Обикновено ние предполагаме, че грешките ϵ_i са независими и еднакво разпределени нормални случайни величини със средно 0 и дисперсия σ^2, която е константа. Функцията $lm()$ намира оценките на свободните членове α и β, които оценки обикновенно означаваме с $\widehat{\alpha}$ и $\widehat{\beta}$, съответно.

```
> lm (y~x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    8142.69        -6.59
```

Получихме следния модел $y = 8142.69 - 6.59x$.

Ще проверим хипотезите $H_0 : \alpha = 0$ срещу $H_1 : \alpha \neq 0$ и $H_0 : \beta = 0$ срещу $H_1 : \beta \neq 0$ при ниво на значимост $\alpha = 0.05$.

```
> summary (lm (y~x))

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-704.04 -142.34  -13.65  158.57  794.50

Coefficients:
```

```
             Estimate   Std. Error  t value   Pr(>|t|)
(Intercept) 8142.6872    23.1195    352.20   <2e-16 ***
x             -6.5898     0.1693    -38.93   <2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 212.4 on 570 degrees of freedom
Multiple R-squared: 0.7267,     Adjusted R-squared: 0.7262
F-statistic:  1515 on 1 and 570 DF,  p-value: < 2.2e-16
```

Резултатът от тази команда съдържа:

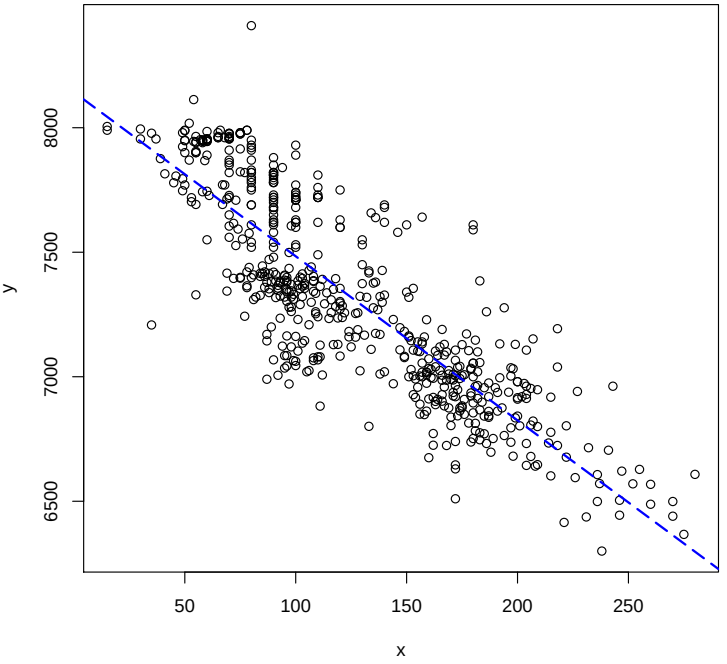
- таблица, наречена *Residuals*, съдържаща числови характеристики на остатъците,
- таблица $Coefficients$ съдържаща оценките на коефициентите на модела и техните стандартни грешки, в която са включени стойностите на t статистиките и p стойностите $Pr(>|t|)$ за проверка на хипотезите за стойностите на коефициентите на модела,
- стандартното отклонение на грешката, означена като стандартна грешка на остатъците ($Residualstandarderror$), която в случая е 212.4,
- други коефициенти за оценка на модела. Коефициентът на детерминация е 0.73.

Основната хипотеза H_0 се отхвърля и в двата случая в полза на алтернативната хипотеза, при което се прави извод, че коефициентите са значими. Сега вече върху графиката можем да построим правата $\widehat{y} = 8142.69 - 6.59x$ през облака от данни.

```
> plot(x,y); abline (lm (y~x), col="blue", lty=23, lwd=2)
```

На фигура 8.4 начертахме правата, която най-точно описва линейната зависимост между x и y. Прогнозираните стойности $\widehat{y}_1, \widehat{y}_2, ..., \widehat{y}_n$ се пресмятат по следния начин: $\widehat{y}_1 = 8142.69 - 6.59x_1, \widehat{y}_2 = 8142.69 - 6.59x_2, ..., \widehat{y}_n = 8142.69 - 6.59x_n$. Тези стойности лежат върху построената права.

Полезна информация за адекватността на модела може да бъде получена от остатъците, които представляват разликите между



Фигура 8.4: Данните и регресионната права

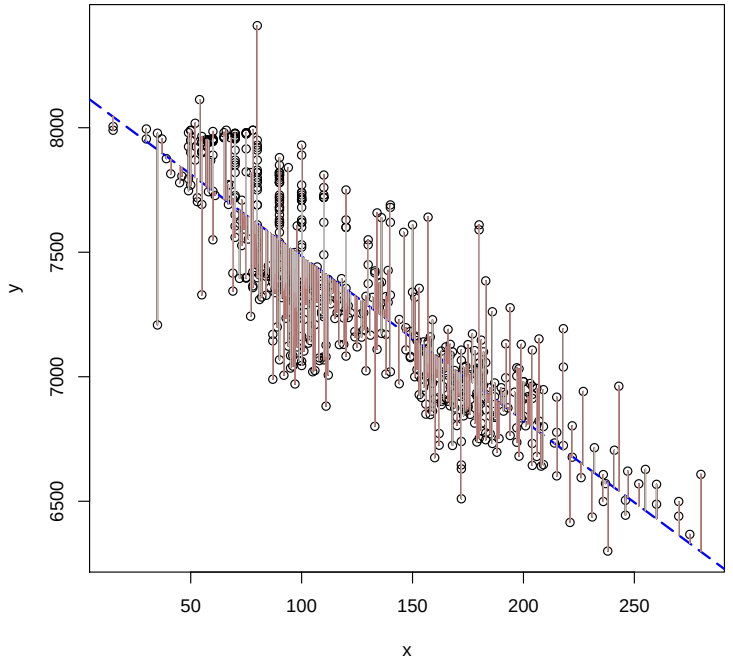
наблюдаваните и оценените стойности на зависимата променлива y: $e_1 = y_1 - \widehat{y}_1, e_2 = y_2 - \widehat{y}_2, ..., e_n = y_n - \widehat{y}_n$

```
> segments (x, fitted (lm(y~x)),x,y, col=8)
```

Ще начертаем остатъците $e_i, i = 1, 2, ..., n$ върху графиката на фигура 8.5.

Всички предположения на линейния регресионен модел трябва да се проверят веднага щом това стане възможно. Две общоприети проверки се реализират със следните две графични представяния на модела: графика на остатъците и предсказаните стойности, и нормална графика.

```
> op <- par(mfrow = c(1,2)); plot(lm(y~x), which=1);
> plot(lm(y~x), which=2); par (op)
```

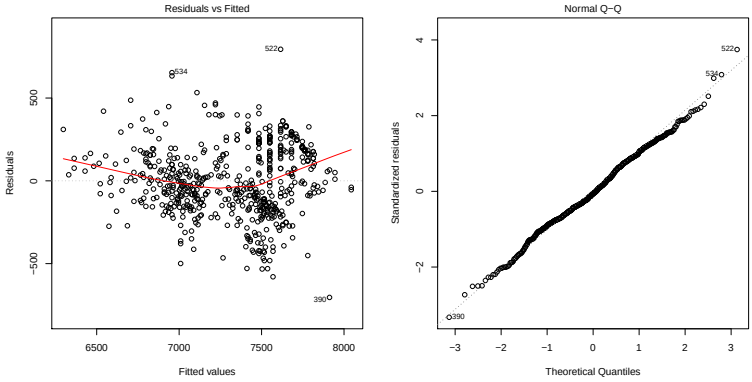


Фигура 8.5: Графика на регресионната права и остатъците

Върху графиката на остатъците и предсказаните стойности на фигура 8.6, остатъците се визуализират като облак от случайно разпръснати точки. В нормалната графика остатъците лежат на една права. Изводът от графичното представяне на остатъците, който в този случай може да направим, е: *Предположенията на модела са изпълнени.*

Сега ще използване модела, който сме построили, за да прогнозираме. Интересуваме се какво е процентното съдържание на палматиновата киселина в зехтин, в който измереното съдържание на олеиновата киселина е 0.8%. Мерните единици, в които са представени всички данни са % $x100$.

```
> newdata = data.frame(x=80)
> predict(lm(y ~ x), newdata, interval="predict")
```



Фигура 8.6: Графика на остатъците и предсказаните стойности, и нормална графика.

```
       fit       lwr       upr
1 7615.502 7197.763 8033.241
```

Прогнозираното процентно съдържание на палматиновата киселина в този зехтин е 76.155%.

Предсказания 95% интервал за процентното съдържание на палматиновата киселина в този зехтин е (71.98%, 80.33%).

Ние сме 95% сигурни, че процентното съдържание на палматиновата киселина в този зехтин е по-голямо от 71.98% и по-малко от 80.33%.

8.2 Дискриминантен анализ

Дискриминантният анализ е многомерен статистически метод за класифициране на обекти в две или повече предварително дефинирани и взаимно изключващи се групи, на основата на множество променливи, характеризиращи тези обекти. Групите се определят от стойностите на една зависима променлива, наречена категорийна, а характеристиките на обектите се представят с променливи, наречени предиктори. Като се използва информацията, която осигуряват предикторите, отделните обекти се класифицират към една от предва-

рително дефинираните групи. Подобни задачи възникват в случаите, когато трябва да се прогнозира принадлежността на даден обект към предварително определена категория. Логическата и познавателната същност на дискриминантния анализ, представен за първи път от Роналд Фишер (Sir Ronald Fisher), се основава на предположението, че обектите от една и съща група притежават близки характеристики по множество показатели.

8.2.1 Задача

Пример 15 (Приложение на дискриминантния анализ в хранително-вкусовата промишленост ???)**.** *Зехтинът и виното са две от най-първите земеделски култури, познати на човека. Зехтинът е растителна мазнина с отличен вкус и несравними здравословни качества. Днес маслини се отглеждат в много държави по света, но родината на този плод е Средиземноморието. Там плодовете поемат силното средиземноморско слънце и узряват перфектно.*

Зехтинът е неразделна част от популярната средиземноморска диета, на която хората от региона дължат доказано по-доброто си здраве в сравнение с народите, обитаващи по-северните краища на Европа. Едно от най-големите предимства на зехтина, освен разбира се превъзходния му вкус, е абсолютната липса на холестерол.

Зехтинът е един от най-добрите източници на ненаситени мастни киселини. Омега мастните киселини се наричат незаменими, понеже те играят важни роли в множество биохимични процеси, а не винаги човешкото тяло може да ги синтезира от други вещества и затова често организмът ги усвоява директно от храната. Полезността на зехтина ще стане разбираема, ако не търсим обяснение изключително в състава на зехтина, а вземем под внимание взаимодействието на зехтина и организма. Установено е, че тялото синтезира омега 6 киселините по-бързо, ако на мястото на синтеза протича успореден синтез на омега 3 мастни киселини, в противен случай за организма е по-лесно да усвои омега 6 киселини от храната. Омега 9 мастните киселини са незаменими само в известна степен. Те могат да бъдат произведени от омега 3 и омега 6 мастни киселини. Когато нивото на послед-

ните в организма е ниско, може да се говори за незаменимост на омега 9 киселините.

Класическите производства и днес са в Северна и Южна Италия и остров Сардиня. Тези 3 области в Италия днес произвеждат около 25% от световното производство. Зехтинът, произведен във всеки район, си има свой вкус и букет от полезни ненаситени мастни киселини. Трудно е днес на пазара да се ориентираш в морето от зехтин.

Можем ли да определим автентичността на зехтина по район, в зависимост от съдържанието на мастните киселини в него?

Къде е произведен зехтин, съдържащ 13% линолова киселина, 0,66% линоленовата киселина и 0.44% ейкозановата киселина? А зехтин, съдържащ 12.53% линолова киселина, 0,22% линоленовата киселина и 0.01% ейкозановата киселина?

Данните се състоят от процентното съдържание на мастни киселини, открити в състава на 572 проби от италиански зехтини.

8.2.2 Модел

Въпреки, че съществуват много модели за класификация, вероятностният модел, основан на Бейсовия подход, успешно се прилага при решаването много задачи за класификация в различни приложни области.

Характеристиките на обектите са представени като вектор X. Априорните вероятности на класовете са означени с $\mathbb{P}(\theta_l)$, $l = 1, \ldots, L$. Приемаме, че обектите от фиксиран клас θ_l , $l = 1, \ldots, L$, са разпределени съгласно функцията на плътността за класа $p(x|\theta_l)$. Вероятността на вектора от предикторите X е дадена като безусловна функция на плътността $p(x) = \sum_{i=1}^{n} \mathbb{P}(\theta_i)p(x|\theta_i)$.

Построяваме правило, основаващо се само на предварителната (априорна) информация. По–точно, за всеки обект имаме данни за параметъра θ_i (т.е. имаме априорна информация), получаваме нови данни D и тогава допълваме знанията си за класа, използвайки

знаменитата формула на Бейс

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)} = \frac{p(\theta, D)}{\int p(\theta, D)d\theta}$$

и получаваме апостериорна информация за класа. Във формулата $p(\theta|D)$ е апостеорното разпределение, $p(\theta)$ е априорното разпределение, $p(D|\theta)$ е условно разпределение.

Предполагаме, че искаме да класифицираме вектор $X = (x_1, \dots x_m)^T$, към клас $c_i, c_i \in \{1, \dots, L\}$. Разполагаме с данни за класификация D:

$$D = \{(x_1, c_1), ..., (x_n, c_n)\}, c_i \in \{1, ..., L\}$$

Необходимо е да зададем форма за класификационната функция:

$$\widehat{c}(X) = arg\ m\ ax_{c\in\{1,...,L\}}\mathbb{P}(c|x)$$

т.е. да намерим оценка на групиращата променлива c като използваме характеристиките на самия обект X.

Класифицираме обекта с характеристики X в класа l с максимална апостериорна вероятност

$$\mathbb{P}(c = l|X) = \frac{p(X|c = l)\mathbb{P}(c = l)}{p(x)} = \frac{p(x|l)\mathbb{P}(c = l)}{\sum_{\lambda=1}^{L}\mathbb{P}(c = \lambda)p(x|\lambda)}$$

Нека разгледаме частния случай когато имаме два класа. Нека двете условни плътности $p(x|c = 0)$ и $p(x|c = 1)$ са нормални (гаусови).

За да можем да класифицираме данните, в действителност не се нуждаем от пресмятане на апостериорната вероятност. Всичко, от което се нуждаем е отношението на числителите, защото знаменателите са едни и същи.

Отношението

$$\frac{p(x|c = 0)\mathbb{P}(c = 0)}{p(x|c = 1)\mathbb{P}(c = 1)}$$

сравняваме с единица.

Ако това отношение е по–голямо от 1, то обектът е от клас 0. В противен случай от клас 1. Достатъчно е да намерим

$$\log \frac{p(x|c=0)\mathbb{P}(c=0)}{p(x|c=1)\mathbb{P}(c=1)}$$

С други думи, граничните стойности се намират, когато е изпълнено

$$\log \frac{p(x|c=0)\mathbb{P}(c=0)}{p(x|c=1)\mathbb{P}(c=1)} = 0$$

Имаме нормални плътности $p(x|c=0)$ и $p(x|c=1)$, и априорни вероятности $\mathbb{P}(c=0) = q_1$ и $\mathbb{P}(c=1) = q_2$. Т.е.

$$p(x|c=0) = \frac{1}{\sqrt{(2\pi)^m|\Sigma_1|}} exp(\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1))$$

$$p(x|c=1) = \frac{1}{\sqrt{(2\pi)^m|\Sigma_2|}} exp(\frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2))$$

Тогава

$$\log\ p(x|c=0)\mathbb{P}(c=0) = -\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + \log\ q_1 - \frac{1}{2}\log(2\pi)^m|\Sigma_1| =$$

$$= -\frac{1}{2}x^T\Sigma_1^{-1}x + \mu_1^T\Sigma_1^{-1}x - \frac{1}{2}\mu_1^T\Sigma_1^{-1}\mu_1 + \log q_1 - \frac{1}{2}\log(2\pi)^m|\Sigma_1|$$

$$\log\ p(x|c=1)\mathbb{P}(c=1) = -\frac{1}{2}x^T\Sigma_2^{-1}x + \mu_2^T\Sigma_2^{-1}x - \frac{1}{2}\mu_2^T\Sigma_2^{-1}\mu_2 + \log\ q_2 -$$

$$-\frac{1}{2}\log(2\pi)^m|\Sigma_2|$$

Когато двете нормални плътности са с различни дисперсиии получаваме квадратна функция за логаритъма на отношението

$$\log \frac{p(x|c=0)\mathbb{P}(c=0)}{p(x|c=1)\mathbb{P}(c=1)}$$

Т.е.

$$\log \frac{p(x|c=0)\mathbb{P}(c=0)}{p(x|c=1)\mathbb{P}(c=1)} = x^TWx + w^Tx + w_0$$

Очевидно когато популационните разпределения имат равни дисперсии, границата между всеки два класа е линия.

Извод: Разглеждаме проблема за класифицирането на обекта към една от групите, като използваме характеристиките на самия обект.

На основата на наблюденията прогнозираме принадлежността на обекта към една от две предварително определени популации. Решаващото правило се състои в разделянето на множеството от възможните стойности на характеристиките на обекта на две взаимно изключващи се области. Когато имаме три популации, една дискриминантна функция разделя първата популация от втората и третата, и друга дискриминантна функция разделя втората съвкупност от третата. Целта е да се разработят правила за класифициране на отделните обекти към една от групите. Границата между всеки два класа може да е линейна или квадратична. Ако данните във всеки клас са с гаусова плътност с различни дисперсии, то границата между всеки два класа е квадратна функция на x. Ако данните във всеки клас са с гаусова плътност с еднакви дисперсии, то границата между всеки два класа е линейна функция на x.

8.2.3 Процедура на дискриминантния анализ

Дискриминантният анализ се използува, когато се нуждаем от прогнозиране стойностите на групиращата променлива. Тази задача е известна още като класификация или разпознаване на образи. Популацията, която изучаваме, е нееднородна или с други думи, се състои от няколко групи (класове) обекти с различни характеристики. Целта ни е за дадено ново наблюдение да определим принадлежността му към групата (класа), от която произлиза.

Сравнителната характеристика на групите, формирани от зависимата категорийна променлива, може да се разглежда като предварителен етап в процеса на прилагане на дискриминантния анализ. Този етап е насочен към формулиране на заключения относно характера и степента на различията между групите по отношение на средните и разсейването на отделните предиктори и ковариацията между тях. Получените резултати позволяват да се направи предварителна оценка на възможностите за приложение на дискриминантния анализ и надеждността на очакваните резултати при съществуващите данни. Естествено, колкото по–големи са различията между отделните групи, толкова по–добри ще бъдат резултатите от анализа по отношение на възможностите за редуциране на вероятността за погрешна класификация на обектите.

Много важно значение в това отношение има правилният подбор

на предикторите, т.е. на независимите променливи, които се използват в анализа. Те трябва да осигурят в достатъчна степен възможности за дискриминация (разделяне) на обектите по отношение на групите.

Процедурата на дискриминантния анализ се състои от две основни стъпки, известни като *фаза на обучение* и *фаза на тестване*. В първата си фаза, процедурата на дискриминантния анализ обработва информацията с цел да я кондензира в така наречените *решаващи правила*. Когато тези правила са получени, естествено е те да бъдат изпробвани върху обектите от обучаващата извадка или върху други обекти с известен клас, наречена тестваща извадка.

При положение, че тези обекти (или поне достатъчно голям процент от тях) бъдат класифицирани правилно, можем да очакваме, че разпознаващите правила са добри и коректно ще работят и за обекти от неизвестен клас.

8.2.4 Дискриминантен анализ с R

В тази глава ще разгледаме задачата, поставена в пример 8.2.1, и ще търсим отговор на въпроса: *Може ли да се предскаже произхода на зехтина (т.е. областта, в която е произведен) като се използва процентното съдържание на омега 3, омега 6 и омега 9 ненаситени мастни киселини?*

Всички мастни киселини, които се съдържат в италиянския зехтин, са дадени в таблица 8.1. [1]

За да решим тази задача ще използваме подходящ метод на дискриминантния анализ, като първо изберем променливите, които ще участват в модела на дискриминантния анализ.

Омега мастните киселини се наричат незаменими, понеже те играят важни роли в множество биохимични процеси, а не винаги човешкото тяло може да ги синтезира от други вещества и затова често организмът ги усвоява директно от храната. Незаменимостта им става разбираема, когато не търсим обяснение само в състава на зехтина, а вземем под внимание взаимодействието на зехтина и организма. Установено е, че тялото синтезира омега 6 киселините по-бързо, ако

[1] Данните достигат до нас благодарение на Интернет страницата http://www.ggobi.org/book/.

Променлива	Пояснение
palmitic	Палмитиновата киселина (хексадеканова киселина) е една от най-често срещаните наситени мастни киселини при животните и растенията. Среща се и в палмовото масло, маслото, сиренето, млякото и месото.
palmitoleic	Омега-7 ненаситена мастна киселина
stearic	Стеариновата киселина, наричана още октадеканова киселина, е една от полезните наситени мастни киселини, които идват от животински и растителни мазнини и масла.
oleic	Олеиновата киселина е мононенаситена omega-9 мастна киселина, която се съдържа в организмите на някои животни и в някои растения.
linoleic	Линоловата киселина е незаменима мастна киселина, необходима за нормалната жизнена дейност на човека. В организма на човека и животните тази киселина постъпва с храната. Тя се съдържа в соята, памука, слънчогледа, лена, конопа, китовата мас и другаде.
linolenic	Линоленовата киселина е незаменима мастна киселина. Съдържа се в много растителни масла, като лененото, конопеното, соевото и други.
arachidic	Фъстъчената киселина е наситена мастна киселина. Съдържа се във фъстъченото олио и в кравето масло.
eicosenoic	Ейкозановата киселина е омега-9 ненаситена мастна киселина, която се съдържа в много видове растително олио.

Таблица 8.1: Разяснения за действието на мастните киселини, съдържащи се в зехтина

на мястото на синтеза протича успореден синтез на омега 3 мастни киселини, в противен случай за организма е по-лесно да усвои омега 6 киселини от храната. Омега 9 мастните киселини са незаменими в само в известна степен. Те могат да бъдат произведени от омега 3 и омега 6 мастни киселини. Когато нивото на последните в организма е ниско, може да се говори за незаменимост на омега 9 киселините.

За предиктори в модела на дискриминантния анализ ще вземем процентното съдържание на ейкозановата киселина, линоловата киселина и линоленовата киселина. Линоленовата киселина е omega-3 незаменима мастна киселина. Линоловата киселина е omega-6 незаменима мастна киселина. Ейкозановата киселина е omega-9 ненаситена мастна киселина. Ако не получим резултат с удоволетворяваща ни точност, ще добавим нови променливи.

За да използваме разгледаните модели на дискриминантния анализ, данните трябва да отговарят на следните две условия:

- предикторите да са нормално разпределени, и
- отделните групи да имат различни средни.

Процентното съдържание на всяка от мастните киселини в зехтина зависи от много фактори и всеки един от тях сам по себе си има малко влияние върху съдържанието на тази мастна киселина в зехтина. Затова можем да предположим, че вероятностното разпределение на съдържанието на всяка мастна киселина в зехтина е нормално с неизвестни параметри.

Освен това трябва да изясним равни или различни са ковариационни матрици за отделните групи. За целта ще проверим хипотези за хомогенност на дисперсиите в групите за всяка количествена променлива. Ще прочетем всичките данни, представени в таблица с 572 реда и 11 колони, и ще генерираме нова таблица, съдържаща само променливите, които са ни необходими, а именно колонките за съдържанието на трите мастни киселини и колонката за област.

```
> data.olive <- read.csv ("F:\\olive.csv")
> d.olive.sub <- subset (data.olive, select = c(region,
+ c(linoleic, linolenic, eicosenoic)))
```

За да се убедим, че разполагаме с данни за три количествени променливи и една категорийна променлива ще визуализираме първите

Променлива	Пояснение
region	Категорийна променлива с три нива за района (1 - Южна Италия, 2 - остров Сардиня, 3 - Северна Италия).
linoleic	Количествена променлива за процентното съдържание на линолевата киселина (x100%).
linolenic	Количествена променлива за процентното съдържание на линоленовата киселина (x100%).
eicosenoic	Количествена променлива за процентното съдържание на ейкозановата киселина (x100%).

Таблица 8.2: Предиктори

шест реда от новата таблица.

```
> head(d.olive.sub)
  region linoleic linolenic eicosenoic
1      1      672        36         29
2      1      781        31         29
3      1      549        31         29
4      1      619        50         35
5      1      672        50         46
6      1      678        51         44
```

В предварителния анализ трябва да дадем отговор на въпроса: *Кои мастни киселини имат разделящ ефект върху множеството от данни?*

Броят на количественните променливи е три. Всяка моделира процентното съдържание на една киселина. Важно е от всички количественни променливи да се изберат за предиктори само тези, които са полезни за разделянето. Проверката на хипотези за груповите средни допринася за този избор. Естествено, добро разделяне може да се очаква когато хипотезата за равенство на средните се отхвърля. Кой критерий за проверка на хипотези за средни ще използваме, зависи от това дали дисперсията е хомогенна. От това дали дисперсията е хомогенна зависи и дали ще използваме линеен или квадратичен модел на дискриминантния анализ.

Затова за всяка количествена променлива проверяваме хипотеза за хомогенност на дисперсията.

```
> for (i in 2:4) {
+ print (bartlett.test (d.olive.sub [,i] ~ d.olive.sub$region))
+ }

        Bartlett test of homogeneity of variances

data:  d.olive.sub[, i] by d.olive.sub$region
Bartlett's K-squared = 69.0636, df = 2, p-value = 1.007e-15


        Bartlett test of homogeneity of variances

data:  d.olive.sub[, i] by d.olive.sub$region
Bartlett's K-squared = 193.1119, df = 2, p-value < 2.2e-16


        Bartlett test of homogeneity of variances

data:  d.olive.sub[, i] by d.olive.sub$region
Bartlett's K-squared = 879.5224, df = 2, p-value < 2.2e-16
```

Извод: Отхвърляме всички основни хипотези за равенство на дисперсиите в полза на съответните алтернативи.

Ще проверим три хипотези за съвпадане на груповите средни и ще изберем за предиктори само тези променливи, за които съответната нулева хипотеза се отхвърля. Тестването на хипотези за средни разгледахме в **??** на стр. **??**. Тъй като в този случай нямаме равенство на дисперсиите на групите, не можем да използваме по-силния метод *anova*(). Затова използваме тест, който не изисква равенство на дисперсиите, именно *oneway.test*().

```
> for (i in 2:4) {
+ print (oneway.test (d.olive.sub [,i] ~ d.olive.sub$region))
+ }

   One-way analysis of means (not assuming equal variances)

data:  d.olive.sub[, i] and d.olive.sub$region
F=442.1535, num df=2.00, denom df=307.176, p-value<2.2e-16


   One-way analysis of means (not assuming equal variances)
```

```
data:  d.olive.sub[, i] and d.olive.sub$region
F = 157.0475, num df = 2.000, denom df = 253.809, p-value < 2.2e-16


    One-way analysis of means (not assuming equal variances)

data:  d.olive.sub[, i] and d.olive.sub$region
F = 1457.806, num df = 2.00, denom df = 332.75, p-value < 2.2e-16
```

Във всяка от направените проверки отхвърлихме основната хипотеза. Следователно има значима разлика в средното съдържание на тези киселини в зехтина, произведен в Северна Италия, Южна Италия и остров Сардиня.

Ще пресметнем извадковите корелационни коефициенти за да се убедим, че количествените променливи са некорелирани.

```
> cor (d.olive.sub[,-1])
              linoleic   linolenic eicosenoic
linoleic     1.00000000 -0.05743858 0.08904499
linolenic   -0.05743858  1.00000000 0.57831851
eicosenoic   0.08904499  0.57831851 1.00000000
```

От предположението за нормалност следва, че променливите са и независими. Избрали сме добри променливи за предиктори.

Зареждаме библиотеката $MASS$, която реализира дискриминантния анализ.

```
> library (MASS)
```

Тъй като видяхме, че условието за хомогенност на дисперсията е нарушено ще приложим функцията $qda()$, реализираща процедурата на квадратичния дискриминантен анализ.

Да напомним, че основната цел на дискриминантния анализ е да се получи правило за причисляване на обект към даден клас. За този обект може да съществува априорна информация за неговата възможна принадлежност към класовете. Прието е такава информация да бъде формулирана в термини на априорни вероятности и да бъде въведена чрез съответните параметри на функцията. Когато такава информация не съществува априорните вероятности за класовете могат да се приемат за равни. По подразбиране функцията $qda()$ приема априорните вероятности за пропорционални на обема

на класовете в обучаващата извадка. Априорните вероятности са необходими за определяне на оптимални класификационни правила.

```
> olive.qda <- qda(region~linoleic+linolenic+eicosenoic,
+ d.olive.sub)
```

Функцията *qda*() обработи наличната информация за обектите от извадката, т.е. стойностите на характеристиките на обектите и класовете към които обектите принадлежат, и я кондензира в обучаващи правила. Да припомним, че в квадратичния дискриминантен анализ се строят квадратични дискриминантни функции от предикторите.

Сега ще изпробваме тези обучаващи правила върху същите обекти от извадката.

```
> predicted.region <- predict(olive.qda, d.olive.sub)$class
```

Визуализираме с таблица доколко прогнозите ни са съвпаднали с реалността.

```
> table (d.olive.sub$region, predicted.region)
   predicted.region
      1   2   3
  1 323   0   0
  2   0  98   0
  3   0   0 151
```

В тази таблица номерът на реда съвпада с действителния клас. Номерът на колонката съвпада с прогнозирания клас. Резултатът, който получихме, е изключително добър. Елементите извън главния диагонал на получената матрица са 0, т.е. всяка проба от италианския зехтин е класифицирана правилно.

При положение, че тези обекти са класифицирани правилно, можем да очакваме, че разпознаващите правила са добри и коректно ще работят и за обекти с неизвестен клас.

Като отговор на основния въпрос можем да кажем, че можем да определим автентичността на италианския зехтин по област като използваме комбинация от процентното съдържание на три ненаситени мастни киселини.

Въпрос: Може ли да се намали броят на предикторите?

Любознателният читател може да се запознае с концепция за избор на подходящ набор от количествени променливи, с които да

построи модел, от статията [**?**]

Сега целта ни е за новите две наблюдения да определим принадлежността им към съответната група (клас), от която произлизат.

```
> new.data.1 <- list ( linoleic=1300, linolenic=66, eicosenoic=44)
> predict (olive.qda, new.data.1)$class
[1] 1
Levels: 1 2 3
```

Да видим и апостериорните вероятности.

```
> predict(olive.qda, new.data.1)$posterior
  1 2 3
1 1 0 0
```

Този зехтин, съдържащ 13% линолова киселина, 0,66% линоленовата киселина и 0.44% ейкозановата киселина, е произведен в Южна Италия.

```
> new.data.2 <- list ( linoleic=1253, linolenic=22, eicosenoic=1)
> predict (olive.qda, new.data.2)$class
[1] 2
Levels: 1 2 3
> predict(olive.qda, new.data.2)$posterior
          1           2            3
1 0.001996842    0.997991 1.209381e-05
```

Зехтинът, съдържащ 12.53% линолова киселина, 0,22% линоленовата киселина и 0.01% ейкозановата киселина, най-вероятно е произведен на остров Сардиня.

Код на R: (???)

```
data.olive <- read.csv ("F:\\olive.csv")
d.olive.sub <- subset(data.olive, select=c(region,c(linoleic,linolenic
head (d.olive.sub)
for (i in 2:4) {
print (bartlett.test (d.olive.sub [,i] ~ d.olive.sub$region))
}
for (i in 2:4) {
print (oneway.test (d.olive.sub [,i] ~ d.olive.sub$region))
}
cor (d.olive.sub[,-1])
library (MASS)
olive.qda <- qda (region ~ linoleic + linolenic + eicosenoic, d.olive
```

```
predicted.region <- predict(olive.qda, d.olive.sub)$class
table (d.olive.sub$region, predicted.region)
new.data.1 <- list ( linoleic=1300, linolenic=66, eicosenoic=44)
predict (olive.qda, new.data.1)$class
predict(olive.qda, new.data.1)$posterior
new.data.2 <- list ( linoleic=1253, linolenic=22, eicosenoic=1)
predict (olive.qda, new.data.2)$class
predict(olive.qda, new.data.2)$posterior
```

Други известни приложения на дискриминантния анализ:

- Стоково кредитиране:

  Имате нужда от потребителски кредит за покупки на стоки на изплащане. Специализирана финансова компания, лидер в областта на потребителското кредитиране в България, в продължение на години е събирала данни за своите клиенти, които включват големина на кредита, срок за погасяване на кредита, история на предишния кредит, трудов стаж, и други. Тя използва тези данни за да моделира кредитния риск на своите потребители в България. Всеки бъдещ неин клиент с този модел е класифициран като надежден или ненадежден за получаване на кредит.

Глава 9

Заключение

Целта на тази книга беше да ви внуши едно послание. Вече сме стигнали до онзи етап в развитието на бизнеса, в който извличането на знания от данните е стил на работа и ние сме изправени пред необходимостта да го овладеем.

Никой от нас - бил той ръководител или служител - не може да си позволи да стои безпомощно пред данните. Всички ние трябва да се научим да извличаме информация и знания от данните с лекотата, с която изпълняваме обичайните си служебни задължения.

Дойде момента да затворите тази книга и да пристъпите обратно в ежедневието си, изпълнено с вземане на решения. Пожелаваме ви успешно бъдеще!

Приложение А

Различни често използвани разпределения

Хипергеометрично разпределение

Казваме, че целочислената случайна величина xi има хипергеометрично разпределение, ако $P(\xi = k) = \frac{C_M^m C_{N-M}^{n-m}}{C_N^n}$. 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 случайната величина "брой дефектни изделия в извадката" е хипергеометрично.

Разпределение на Поасон

Поасоновото разпределение се определя лесно като граница на биномни разпределения, когато $n \to \infty$, така че $np \to \lambda$ (виж Теорема 2.10). Поасоново разпределената случайна величина ξ може да приема всякакви целочислени стойности с вероятност $P(\xi = k) = \frac{\lambda^k}{k!} e^{-\lambda}$. Означаваме $\xi \in Po(\lambda)$. Поасоновото разпределение 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.

Математическото очакване и дисперсията на поасоновото разпределение съвпадат: $E\xi = D\xi = \lambda$. Това най-лесно се вижда от пораждащата функция на поасоновото разпределение, която се пресмята

директно (виж ???).

Интерпретация на поасоновото разпределение Разглеждаме появите на дадено събитие в рамките на експеримент, подчиняващ се на следните три условия:

- Броят на сбъдванията на събитието в произволен интервал от време не зависи от броя на появяванията му в произволно избран друг интервал.
- Очакваният брой реализации на събитието е правопропорционален на големината на наблюдавания интервал.
- Във всеки произволно малък интервал събитието може да се реализира не повече от един път.

При тези условия случайната величина брой на сбъдванията на събитието в интервал с големина 1 е поасоново разпределена. Интерпретацията на параметъра на разпределението λ е очакван брой реализации на събитието в интервал с големина единица.

Приложение Б

Списък със статистическите термини и техните еквиваленти на английски

R $\mathbf{R}$ R R $\mathbb{R}$ R

Азбучен указател