***Въведение в Статистиката с R – Мони – II тема***

След краткото въведение в основните концепции в R ще покажем как можем да работим с конкретните типове данни. В тази част ще демонстрираме основните възможности на R за обработка на данни като използваме ***data frame*** с име tires. Той съдържа резултатите от проучване на мнението на потребителите на автомобилни гуми. Може да бъде зареден чрез командата

*name =* ***read.csv*** *(file, header = TRUE, sep = ",", quote = "\"",dec = ".", fill = TRUE, comment.char = "", ...)*

В нашия случай тя ще бъде

tires = read.csv(file = "C:\\Users\\User\\Desktop\\MoniStat\\tires.csv", header = TRUE, sep = ";", dec = ",")

Нейните компоненти можем да разгледаме с командата ***ls***

ls(tires)

[1] "N" "X1" "X10" "X11\_1" "X11\_2" "X11\_3" "X12\_1" "X12\_2" "X12\_3" "X2" "X3" [12] "X4" "X5" "X6" "X7" "X8" "X9"

По точно с горните вектори са моделирани

N – Номер на анкетирания

X1 – Вид гума, за която ще бъде попълнена анкетата

X2 - Производител

X3 - Доставчик

X4 - Цена на закупуване, актуализирана към днешна дата

X5 - Пробег

X6 - Продължителност на живот в дни

X7 - Вид на гумите според сезона

X8 – Място на гумата според колата

X9 - Диаметър на джантата в цолове

X10 – Ниво на удовлетвореност на клиента

X11\_1 - Първа любима марка гуми на клиента

X11\_2 - Втора любима марка гуми на клиента

X11\_3 - Трета любима марка гуми на клиента

X12\_1 - Първа марки гуми, които клиента не би купил

X12\_2 - Втора марки гуми, които клиента не би купил.

X12\_3 – Трета марки гуми, които клиента не би купил.

Данните са хипотетични, поради които значенията на признака са кодирани със символи. Първите няколко реда на таблицата могат да бъдат разгледани с функцията head

head(tires)

N X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11\_1 X11\_2 X11\_3 X12\_1 X12\_2 X12\_3

1 T1 P1 S1 93.98 51656.45 1196 Зимни 1л 14 1 T1 T2 T3 T4 T5 T6

2 T2 P2 S2 94.76 40885.81 760 Зимни 1л 15 2 T2 T4 T3 T21 T14 T13

3 T3 P3 S3 104.28 25534.63 732 Зимни 1л 16 3 T1 T5 T2 T11 T3 T21

4 T4 P4 S4 104.18 33402.56 1163 Зимни 1л 16 4 T1 T2 T3 T11 T21 T5

5 T5 P5 S5 98.50 28624.92 865 Зимни 1д 16 1 T2 T4 T3 T21 T6 T7

6 T6 P6 S1 111.22 44554.31 727 Зимни 1д 17 1 T1 T5 T2 T11 T7 T8

1. **Категорийни, качествени данни –**

Едни данни се наричат категорийни или още качествени, ако техните възможни значения са словестни, описателни. Тези възможни значения трябва да бъдат описани като непресичащи се класове. Например признакът пол има възможни значения мъж или жена, признакът „Най-висока степен на завършено образование“ обикновено се изчерпва с няколко възможни значения „Без образование“, „Основно “, „Cредно“, „Бакалавър“, „Магистър“ или „Доктор“. От таблицата с данни tires категорийни променливи са

X1, X2, X3, X7, X8, X10, X11\_1, X11\_2, X11\_3, X12\_1, X12\_2, X12\_3

Сега ще демонстрираме основните възможности на R за обработка на категорийни данни като използваме таблицата с данни tires.

**Групировка в абсолютни честоти**

В резултат от групировката на данните според наблюдаваните при тях значения на признака се получават така наречените честотни разпределения. Те съдържат броят на наблюденията, които попадат в непресичащите се групи, определени от различните регистрирани значения на признака при наблюдаваните статистически единици.

#### Например

в таблицата с данни tires честотното разпределение по признака Х1 се получава след групировка според вида на гумата, за която ще бъде попълнена анкетата.

До векторите, които съдържат резултатите от наблюденията можем да достигнем чрез функцията

> attach(tires)

или чрез оператора “$” като преди него напишем името на таблицата с данни. Т.к. ние ше работим изцяло с тази таблица с данни ще използваме функцията attach.

За да намерим броевете на наблюденията в отделните групи можем да използваме функцията table. Резултатът от тази функция е таблица и нека присвоим този резултат на променливата X1.freq.

> X1.freq = table(X1)

X1

T1 T2 T3 T4 T5 T6 T7

22 24 23 26 28 63 64

За да разменим местата на колоните и редовете можем да използваме функцията cbind. Тя отпечатва резултата във формат на колони.

> cbind(X1.freq)

X1.freq

T1 22

T2 24

T3 23

T4 26

T5 28

T6 63

T7 64

За да заменим кратките етикети с истинските имена на колите можем да използваме функцията ***levels***.

> table(X1)

X1

T1 T2 T3 T4 T5 T6 T7

22 24 23 26 28 63 64

> levels(X1) = c("Audi", "BMW", "Honda", "Mercedes", "Nissan", "Peugeot", "Citroen")

> table(X1)

X1

Audi BMW Honda Mercedes Nissan Peugeot Citroen

22 24 23 26 28 63 64

*Задачи за упражнение.* Направете групировки по признаците

X2 - Производител

X3 - Доставчик

и представете това разпределение в формат на редове и във формат на колони.

На кои производители най-често се срешат гуми в извадката ни? На кои доставчици най-често се срещат гуми в извадката ни?

**Групировка в относителни честоти**

Групировка по относителни честоти по дадена променлива наричаме определянето на относителните честоти в непресичащите се групи, определени от различните регистрирани значения на признака при наблюдаваните статистически единици.

Връзката между абсолютна *fA* и относителна честота *pA* се дава с формулата pА = , където n е обемът на извадката.

Например

В данните tires относителната честота на използвалите гуми Т1 е броят на Т1 в колоната Х1 върху броят на всички наблюдавани 250, което е *pТ1* = 22/250 = 0.088.

I начин За да направим групировка по признака „Вид на използваната гума“, т.е. по променливата Х1 намираме

> X1.relfreq = X1.freq/length(X1)

> X1.relfreq

X1

T1 T2 T3 T4 T5 T6 T7

0.088 0.096 0.092 0.104 0.112 0.252 0.256

С помощта на функцията ***round*** можем да определим до колко знака след десетичната запетая да се закръгля резултата

> X1.relfreq = round(X1.freq/length(X1),2)

> X1.relfreq

X1

T1 T2 T3 T4 T5 T6 T7

0.09 0.10 0.09 0.10 0.11 0.25 0.26

Ако искаме да я превърнем в проценти, умножаваме по 100. Т.е. 8,8% от всички наблюдавани са използвали гуми от вида Т1.

> X1.relfreq = X1.freq/length(X1)\*100

> X1.relfreq

X1

T1 T2 T3 T4 T5 T6 T7

8.8 9.6 9.2 10.4 11.2 25.2 25.6

II начин Можем да направим групировка по признака „Вид на използваната гума“, т.е. по променливата Х1 и с помощта на функцията ***prop.table*** като можем и да закръглим резултата до където ни харесва с помощта на функцията ***round***

> round(prop.table(table(X1)), 2)

X1

T1 T2 T3 T4 T5 T6 T7

0.09 0.10 0.09 0.10 0.11 0.25 0.26

*Задачи за упражнение.* Направете групировки в относителни честоти по признаците

X2 - Производител

X3 - Доставчик

В относителни дялове и в проценти и представете това разпределение в формат на редове и във формат на колони.

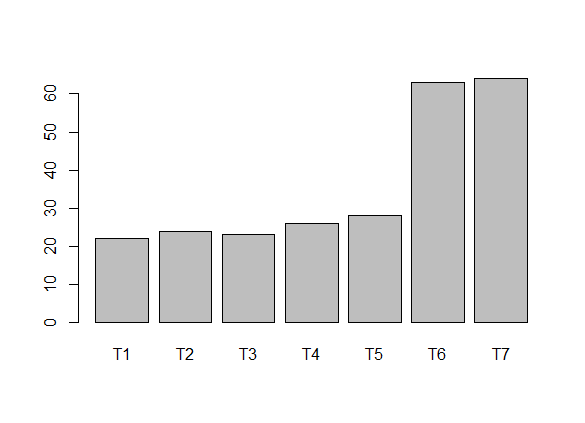
**Правоъгълна плоскостна диаграма (Bar Graph)**

Правоъгълната плоскостна диаграма се използва при качествени данни. Състои се от успоредни правоъгълници, чиито вертикални страни са равни на абсолютната (относителната) честота на наблюденията в групите. Служи за онагледяване на съответното честотно разпределение на статистическите единици според значенията на наблюдавания признак.

#### Например

в таблицата с данни tires честотното разпределение по признака Х1 се получава след групировка според вида на гумата, за която ще бъде попълнена анкетата. Онагледяването на това разпределение може да стане с помощта на функцията ***barplot***. В абсолютни честоти това ще бъде

> barplot(table(X1))

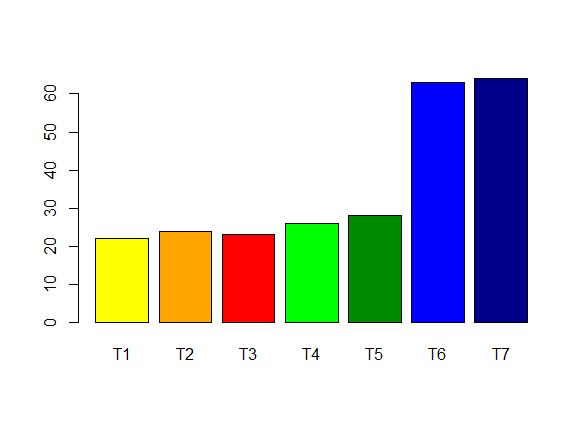


Ако искаме да добавим цветове първо трябва да си ги зададем

> colors = c("yellow", "orange", "red", "green", "green4", "blue", "blue4")

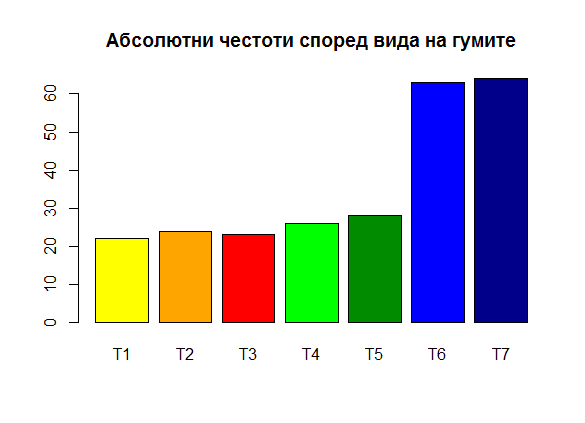
И след това да ги използваме

> barplot(table(X1), col=colors)



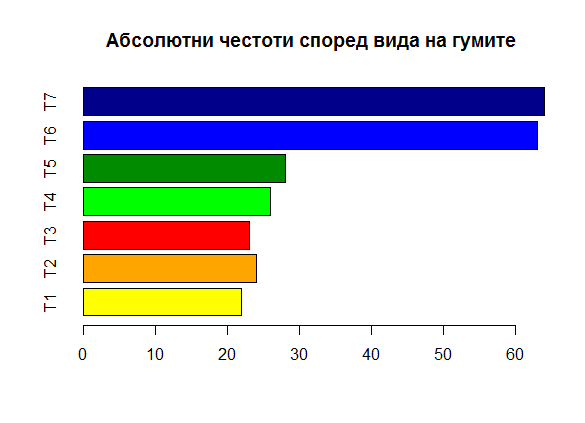
Можем да домавим заглавие

> barplot(table(X1), col=colors, main = "Абсолютни честоти според вида на гумите")



Можем да обърнем осите

> barplot(table(X1), col=colors, main = "Абсолютни честоти според вида на гумите", horiz = TRUE)



Да сменим имената като е добре да ги зададем като отделен вектор

names = c("Audi", "BMW","Honda", "Mercedes", "Nissan","Peugeot", "Citroen")

barplot(table(X1), col=colors, main = "Абсолютни честоти според вида на гумите", horiz = TRUE, names.arg = names)

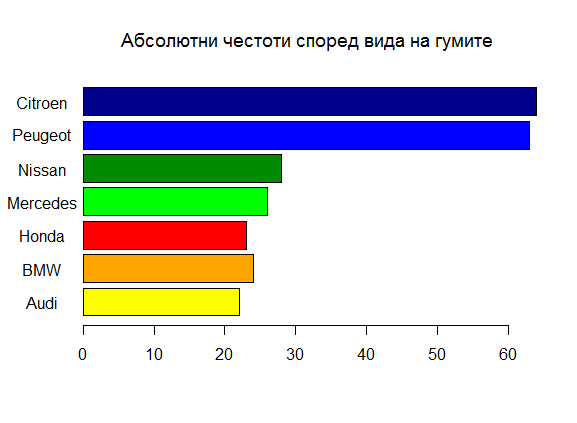
Да сменим големината на шрифта на имената

barplot(table(X1), col=colors, main = "Абсолютни честоти според вида на гумите", horiz = TRUE, names.arg = names, cex.names = 0.65)

Да завъртим имената за да се съберат. Всъщност те първо се изтриват и после се добавят хоризонтално

> BP = barplot(table(X1), col=colors, main = "Абсолютни честоти според вида на гумите", horiz = TRUE, names.arg = FALSE)

> text(cex=1, x=colMeans(BP)-10, y=c(0.7,1.9,3.1,4.3,5.5,6.7,7.9), names, xpd=TRUE, srt=0)



Можем да сменим шрифта на графиките, който е по подразбиране

> windowsFonts(times = "Times new roman")

> par(family = "times", font = 1, font.axis=2, font.main = 3)

> BP = barplot(table(X1), col=colors, main = "Абсолютни честоти според вида на гумите", horiz = TRUE, names.arg = FALSE)

> text(cex=1, x=colMeans(BP)-10, y=c(0.7,1.9,3.1,4.3,5.5,6.7,7.9), names, xpd=TRUE, srt=0)



За повече промени в параметрите може да се разгледа помощната документация на R.

*Задачи за упражнение.* Направете групировки в относителни честоти по признаците

X2 - Производител

X3 - Доставчик

В относителни дялове и в проценти и представете това разпределение чрез правоъгълна плоскостна диаграма.

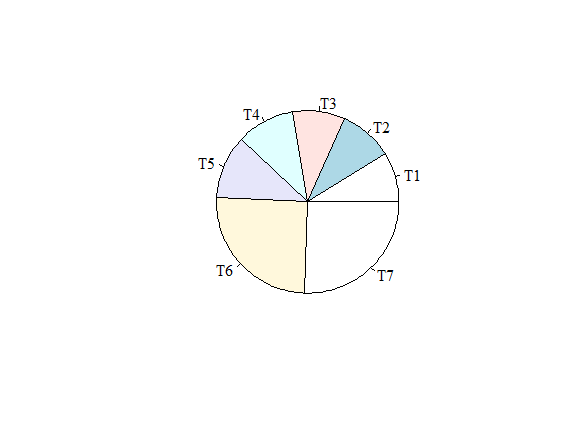
**Кръгова диаграма - Pie Chart**

Кръговите диаграми се използват за представяне на структурата на съвкупността по наблюдавания признак. Общото правило при построяването им е, че лицето на частта, която представя дадена подсъвкупност към лицето на целия кръг се отнася както броят на единиците в подсъвкупността към обема на наблюдаваната съвкупност.

*Например*

По честотното разпределение по признака Х1, което се получава след групировка според вида на гумата, в данните от таблицата tires онагледяването на разпределението може да стане с помощта на функцията ***pie***.

> pie(table(X1))



Отново ако искаме да добавим цветове трябва да си ги зададем.

> colors = c("yellow", "orange", "red", "green", "green4", "blue", "blue4")

И след това да ги използваме

> pie(table(X1), col=colors)

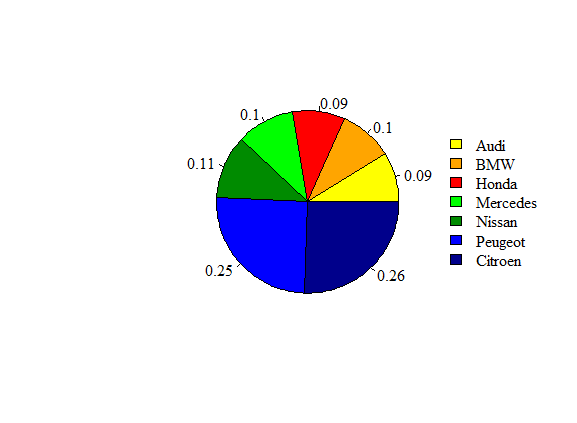
#### 

Можем да махнем имената и да добавим легенда към графиката. Мястото на легендата се определя най-добре с кликване на мишката върху нея.

> percentlabels = round(prop.table(table(X1)),2)

> pie(table(X1), col = colors, labels = percentlabels)

> legend(locator(1), names, cex=1, fill=colors,bty="n")



При работа с категорийни променливи често пъти е удачно да се разгледат и сравнят различни статистики в подгрупите като се вземат данните на друг метриран признак на същите статистически единици. Това може да стане като поотделно сметнем въпросната статистика във всяка от подгрупите. Например средната цена на закупуване, актуализирана към днешна дата (X4) за колите от вид Т1 може да бъде определена с помощта на функциите ***mean*** и ***which*** и логическия оператор „==“

> mean(X4[which(X1=="T1")])

[1] 107.5523

Ако искаме да направим това едновременно за всички подгрупи можем да използваме функцията ***tapply***, която прилага функцията, посочена като трети параметър, поотделно към векторите, които се получават от вектора, посочен на първо място, които имат еднакви значения по признака, посочен на второ място.

*Например*

За да пресметнем средната цена на закупуване на гумите, актуализирана към днешна дата (X4) за колите от всеки вид поотделно можем да използваме

> tapply(X4, X1, mean)

T1 T2 T3 T4 T5 T6 T7

107.5523 104.9625 105.2165 104.0677 104.1918 107.8871 105.6030

Добре е да закръглим резултата, например до втория знак след десетичната точка например с помощта на функцията ***round***

> round(tapply(X4, X1, mean) ,2)

|  |  |
| --- | --- |
| T1 T2 T3 T4 T5 T6 T7  107.55 104.96 105.22 104.07 104.19 107.89 105.60  В случая наблюдаваме, че средните цени на всички видове наблюдавани гуми са почти равни.   1. **Количествени данни**   Количествените данни се подразделят на дискретни и непрекъснати. Те могат да бъдат обработвани чрез използване на различни аритметични операции. Ще обясним как може да бъде направен анализ на такива данни посредством няколко примера. Всички те ще са свързани с обработка на данни от таблицата tires. Този път ще работим с метрираните признаци, а те са:  X4 - Цена на закупуване, актуализирана към днешна дата  X5 - Пробег  X6 - Продължителност на живот в дни  X9 - Диаметър на джантата в цолове.  **Честотни разпределения на количествени данни**  Тези разпределения се получават след групировка на данните и представляват представяне на броевете на наблюденията, които попадат в отделните подинтервали.  *Например*  За да определим честотното разпределение на гумите според тяхната цена на закупуване, актуализирана към днешна дата, първо е добре да определим краищата на интервалите, които ще разграничават групите. Освен това групите трябва да са така направени, че да са с равна ширина, всяко наблюдение да попада в точно една от групите, най-малкото наблюдение да е ляв край на първия интервал и най-голямото наблюдение да е десен край на най-последния интервал. Това не трябва да се абсолютизира. Може обхвата на данните да се увеличи (по възможност симетрично), така че краищата на интервалите да са цели числа или дробни числа с най-много една цифра след десетичната точка.  Минималната и максималната измерена стойност, определят обхвата на наблюденията   > range(Х4)  [1] 82.13 167.00  Т.е. в нашата извадка минималната наблюдавана цена е 82.13 лв, а максималната 167.00 лв.  Увеличаваме интервала до (82; 167). Получаваме краищата на интервалите например с  > breaks = seq(82, 167, by=17)  > breaks  [1] 82 99 116 133 150 167  С функцията ***cut*** превръщаме количествения признак в категориен, като на мястото на всяко число поставим интервала, в който то принадлежи. За да осигурим всяко наблюдение да бъде в точно една група интервалите може да са например затворени отляво и отворени от дясно. Това се постига, когато на параметъра right зададем стойност FALSE.  > X4\_cut = cut(X4, breaks, right=FALSE)  Първите 5 наблюдения можем да разгледаме с функцията ***head***.  > head(X4\_cut)  [1] [82,99) [82,99) [99,116) [99,116) [82,99) [99,116)  Levels: [82,99) [99,116) [116,133) [133,150) [150,167)  Сега, както при категорийни признаци можем да построим честотното разпределение по признака Х4.  *В абсолютни честоти*  > X4\_cut\_freq = table(X4\_cut)  > X4\_cut\_freq  X4\_cut  [82, 99) [99, 116) [116, 133) [133, 150) [150, 167)  94 110 25 11 9  Най-много от гумите имат цена между 99 и 116 лв.  Отново можем да завъртим редовете и колоните.  > cbind(X4\_cut\_freq )  X4\_cut\_freq  [82,99) 94  [99,116) 110  [116,133) 25  [133,150) 11  [150,167) 9  *В относителни честоти*  Можем да превърнем честотните разпределения в разпределения по относителни честоти. Т.е. вместо брой единици групите да имаме относителните честоти в групите:  > breaks = seq(82, 167, by=17)  > breaks  [1] 82 99 116 133 150 167  > X4\_cut = cut(X4, breaks, right=FALSE)  > X4\_cut\_relfreq = table(X4\_cut)/length(X4)  > X4\_cut\_relfreq  X4\_cut  [82, 99) [99, 116) [116, 133) [133, 150) [150, 167)  0.376 0.440 0.100 0.044 0.036  Може да закръглим числата  > X4\_cut\_relfreq = round(table(X4\_cut)/length(X4),2)  > X4\_cut\_relfreq  X4\_cut  [82, 99) [99, 116) [116, 133) [133, 150) [150, 167)  0.38 0.44 0.10 0.04 0.04  Често е по-добре редовете и колоните на таблиците да бъдат сменени  > X4\_cut\_relfreq = cbind(round(table(X4\_cut)/length(X4),2))  > X4\_cut\_relfreq  [,1]  [82,99) 0.38  [99,116) 0.44  [116,133) 0.10  [133,150) 0.04  [150,167) 0.04  Можем да използваме функцията cbind  и да отпечатаме едно до друго разпределенията в абсолютни и  в относителни числа.  > cbind(X4\_cut\_freq, X4\_cut\_relfreq)   X4\_cut\_freq  [82,99) 94 0.38  [99,116) 110 0.44  [116,133) 25 0.10  [133,150) 11 0.04  [150,167) 9 0.04  *Задача за упражнение*  Групирайте данните по признаците  X5 - Пробег  X6 - Продължителност на живот в дни  X9 - Диаметър на джантата в цолове.  И постройте техните честотни разпределения в брой и в относителни честоти.  **Хистограма**  Хистограмата се състои от правоъгълници с успоредни страни, които показват графично честотното разпределение на наблюдаваната количествена величина. Вертикалните страни на правоъгълниците може да са равни на абсолютните честоти в съответните групи. Тогава говорим за хистограма на абсолютните честоти. Когато вертикалните страни на правоъгълниците са равни на относителните честоти в съответните групи говорим за хистограма на относителните честоти.  *Например*  За да начертаем хистограмата на абсолютните честоти на разпределението на гумите според тяхната цена на закупуване, актуализирана към днешна дата можем да използваме функцията ***hist***. Отново за да осигурим всяко наблюдение да бъде в точно една група образуваме интервалите така, че да са затворени отляво и отворени от дясно. Да припомним, че това се постига, когато на параметъра right зададем стойност FALSE.  > hist(X4, right = FALSE)    Както при кръговата диаграма, така и тук можем да оцветим хистограмата, да сменим надписите по осите или заглавието.  > hist(X4, right = FALSE, main = "Хистограма на разпределението според цената в лв. ", xlab = "лв.", ylab = "Броя", col = c("green4", "green","yellow","orange","red", "red4","purple","blue","blue4"))    Ако искаме да направим хистограма на относителните честоти трябва да добавим в параметрите на функцията ***probability=TRUE.*** Получаваме  hist(X4, probability=TRUE, right = FALSE, main = "Хистограма на разпределението според цената в лв. ", xlab = "лв.", ylab = "Броя", col = c("green4", "green","yellow","orange","red", "red4","purple","blue","blue4"))    Лицето на цялата хистограма на относителните честоти е 1. По тази причина тя върши по-добра работа по-късно при сравняване със съответните теоретични вероятности.  Едно добро допълнение към хистограмата са отметките, които отбелязват къде са точно измерените значения в групите и ни помагат да наблюдаваме концентрацията на наблюдения в отделните групи.  Те се поставят с помощта на  > rug(jitter(X4))    *Задача за упражнение*  Постройте хистограми на разпределението на данните по признаците  X5 - Пробег  X6 - Продължителност на живот в дни  X9 - Диаметър на джантата в цолове  по абсолютни и по относителни честоти.  **Честоти с натрупване (Cumulative Frequency Distribution)**  Кумулативните честоти са честотите с натрупване. Все едно при групировката групите са сформирани само като "До …".  *Например*  За да определим честотите с натрупване на разпределението на гумите според тяхната цена на закупуване, актуализирана към днешна дата можем да използваме функцията ***cumsum,*** като първо намираме честотното разпределение  > breaks = seq(82, 167, by=17)  > breaks  [1] 82 99 116 133 150 167  > X4\_cut = cut(X4, breaks, right=FALSE)  > X4\_cut\_relfreq = table(X4\_cut)  > X4\_cut\_relfreq  X4\_cut  [82, 99) [99, 116) [116, 133) [133, 150) [150, 167)  94 110 25 11 9  След това прилагаме функцията ***cumsum***.  > cumsum(X4\_cut\_relfreq)  [82, 99) [99, 116) [116, 133) [133, 150) [150, 167)  94 204 229 240 249  Сега да сменим местата на редовете и колоните  > cbind(round(cumsum(X4\_cut\_relfreq), 2))  [,1]  [82,99) 94  [99,116) 204  [116,133) 229  [133,150) 240  [150,167) 249  Относителните честоти с натрупване, когато данните не са групиране се наричат емпирична функция на разпределение. Точната дефиниция е следната  Стойностите на емпиричната функция на разпределение могат да бъдат намерени с помощта на функцията ***ecdf***.  > ecdf(X4)  Empirical CDF  Call: ecdf(X4)  x[1:240] = 82.13, 83.82, 84.03, ..., 163.64, 167  Нейната графика може да бъде изчертана с помощта на функцията ***plot***  plot(ecdf(X4), verticals = FALSE, col="darkblue", do.points = FALSE, lwd = 1, main = "Функция на разпределение на X4")    *Задача за упражнение*  Постройте честотите с натрупване на данните по признаците  X5 - Пробег  X6 - Продължителност на живот в дни  X9 - Диаметър на джантата в цолове.  Изчертайте функциите на разпределение на тези величини.  **Stem-and-Leaf Plot**  Графиката **stem-and-leaf** е текстова графика, която показва разпределението по наблюдавания признак. Например по-долу цената е закръглена до цяло число и последната цифра на това число е след отвесната черта.  > stem(X4)  The decimal point is 1 digit(s) to the right of the |  8 | 244  8 | 56889999  9 | 0000000000001111111122222223333333333334444444444444  9 | 555555566777777788888888889999999999  10 | 00000111111111111111222222222222222333333333344444444444  10 | 555555556666666777778889999  11 | 0000111122223334444  11 | 55566667888  12 | 123333444  12 | 6778  13 | 0001344  13 | 56789  14 | 4  14 | 59  15 | 02  15 | 567788  16 | 4  16 | 7  *Задача за упражнение*  Постройте **stem-and-leaf** графиките на наблюденията на признаците  X5 - Пробег  X6 - Продължителност на живот в дни  X9 - Диаметър на джантата в цолове.  Изчертайте функциите на разпределение на тези величини.  **Корелационно поле(Scatter Plot)**  Нека изследваме зависимостта между променливите Х и Y. Обикновено се започва с изчертаване на корелационното поле на данните върху Декартова координатна система. От неговия графичен образ получаваме първична представа за формата на зависимост между наблюдаваните величини. По абсцисната ос, обикновено на равномерна скала се нанасят измерените значения на факторпризнака X, а по ординатната, измерените значения на резултативната величина Y. Да предположим, че разполагаме с n на брой двойки от наблюдения *(X1,Y1), (X2,Y2), … ,(Xn,Yn).* За да построим тяхното корелационното поле на всяка двойка съпоставяме точка със съответните координати.  В R това може да бъде направено с помощта на функцията ***plot***.  *Например*  В таблицата с данни tires корелационното поле, чрез което може да се онагледи формата на зависимост между променливите X4 - Цена на закупуване, актуализирана към днешна дата и X9 - Диаметър на джантата в цолове  > head(cbind(X4, X9))  X4 X9  [1,] 93.98 14  [2,] 94.76 15  [3,] 104.28 16  [4,] 104.18 16  [5,] 98.50 16  [6,] 111.22 17  ………………………………  може да бъде построено по следния начин  > plot(X9, X4, xlab=" Диаметър на джантата в цолове ", ylab=" Цена на закупуване ",  + col="blue", cex =0.5)  Получаваме    Наблюдава се струпване на точките около възходяща права. Т.е. имаме правопропорционална зависимост между наблюдаваните признаци. При това имаме по-голямо разсейване при по-големите размери на гумите.  Можем да генерираме графика на правата, която най-добре минава между точките. Т.е. да построим линията на регресия. За алгоритъма на построяване, както и за нейните свойства ще учим по-късно. Тази права се построява с помощта на функцията ***abline***. Моделът се построява с помощта на функцията ***lm***, а уравнението на регресия се задава с оператора ***~***, като от ляво се поставя резултативната величина X4, а от дясно независимата променлива X9.  > abline(lm(X4 ~ X9))  По тази права можем да определим очакваните значения на резултативната величина, в случая – цената на закупуване при зададени значения на фактор признака X9 - Диаметър на джантата в цолове    Колкото точките от корелационното поле са по-силно струпани около линията на регресия, толкова корелационната зависимост между наблюдаваните признаци е по-силна. Колкото точките са по-разпръснати, толкова корелационната зависимост е по-слаба. Да отбележим, че корелационната зависимост не е причино-следствена зависимост. Тя моме да бъде породена от трети невключени в модела фактори.  *Задача за упражнение*  Постройте корелационните полета на зависимостите между всеки два от признаците  X4 - Цена на закупуване, актуализирана към днешна дата  X5 - Пробег  X6 - Продължителност на живот в дни  X9 - Диаметър на джантата в цолове.  Изчертайте линиите на регресия и ги анализирайте.  **Числови характеристики**  В тази част ще разгледаме основните числови характеристики на метрираните признаци и по-точно: средно аритметично, мода(без), медиана, квантили, квартили, персентили, обхват, междуквартилно разстояние, кутия с мустачки, вариация, стандартно отклонение, ковариация, корелационен коефициент, централни моменти, асиметрия, ексцес.  *Средно аритметично(Mean)*  Средното аритметично е числова характеристика на центъра на разпределението на наблюдаваните единици според на наблюдаваната величина. То представлява сумата от данните, разделена на броя им. Ще разграничаваме две средни:   * Извадково средно (sample mean)   1-∑n x¯= n    xi       i=1  където n е обемът на извадката и   * Популационно средно (population mean)   1 ∑N μ = --   xi     N i=1  където N е обемът на цялата популация, за която ще правим извод.  *Например*  Намерете средната цена на закупуване на гумите в извадката, актуализирана към днешна дата. Това може да стане с помощта на функцията ***mean***.  > mean(X4)  [1] 105.9456  Можем да я закръглим до втория знак след десетичната запетая.  > round(mean(X4), 2)  [1] 105.95  Т.е. средната цена на гумите в извадката е 105.95 лв.  Можем да изчисляваме и средни аритметични само в подгрупи, в които желаем ние. Например средната цена само на колите от вид Т5 може да бъде намерена с  > mean(X4[X1 == "T5"])  [1] 104.1918  *Задача за упражнение*  Определете средните аритметични на признаците  X5 - Пробег  X6 - Продължителност на живот в дни  X9 - Диаметър на джантата в цолове.  и ги анализирайте. Намерете средната цена на колите с най-малки джанти.  *Медиана(Median)*  Медианата е онова значение на признака, за което можем кажем, че половината от наблюденията са по-големи или равни на него и половината от наблюденията са по-малки или равни на него.  При определянето й първо се подреждат данните по големина, обикновено възходящо, и след това при нечетен брой наблюдения медианата е наблюдението, което се намира в средата на подредения ред от данни. При четен брой наблюдения, медианата е средното аритметично на двете значения на признака, които се намират в средата на подредения ред от данни. Тя също характерзира центъра на разпределението на наблюдаваните единици според на наблюдаваната величина. Основното й предимство е, че за разлика от средното аритметично тя е по-малко чувствителна на грешки в наблюденията, защото при определянето й участва техния брой и значенията само на едно или най-много две от наблюденията. В R един от начините за определяне на медианата е с помощта на функцията median.  > median(X4)  [1] 101.895  Можем да я закръглим до втория знак след десетичната запетая.  > round(median(X4), 2)  [1] 101.89  Т.е. половината от цените на наблюдаваните гуми са по-малко или равни на 101.89 лв. и половината от цените са по-големи от тази цена.  *Задача за упражнение*  Определете медианите на признаците  X5 - Пробег  X6 - Продължителност на живот в дни  X9 - Диаметър на джантата в цолове.  и ги анализирайте.  *Квартили(Quartile) и квантили*  Квартилите разделят статистическия ред на 4 равни части. Първо подреждаме данните възходящо. Тогава 25% от наблюденията са по-малки или равни на първия квартил, който се бележи с Q1 и се нарича още долен квартил (QL). 25% от наблюденията са по-големи или равни на първия квартил и по-малки или равни на втория квартил, който съвпада с медианата. 25% от наблюденията са по-големи или равни на медианата и по-малки или равни на третия квартил (Q3), който се нарича още горен квартил и се бележи с QU. 25% от наблюденията са по-големи или равни на третия квартил. Тъй като квартилите се явяват съответно 0.25, 0.50 и 0.75 квантили в R един от начините за определяне на квартилите е с помощта на функцията ***quantile***.  quantile(x, probs, ...)  Параметрите 0, 0.25, 0.50, 0.75 и 1 са зададени по подразбиране за стойности на probs. Да определим квартилите на променливата Х4.  > quantile (X4)  0% 25% 50% 75% 100%  82.130 94.450 101.895 110.840 167.000  Можем да ги закръглим например до втория знак след десетичната запетая.  > round(quantile (X4), 2)  0% 25% 50% 75% 100%  82.13 94.45 101.89 110.84 167.00  Т.е. минималната цена на наблюдаваните гуми е 82.13 лв. Една четвърт от наблюденията имат цена по-малка или равна на 94.45 лв. Една четвърт от наблюденията имат цена в интервала [94.45; 101.89]. Една четвърт от наблюденията имат цена в интервала [101.89; 110.84]. И една четвърт от цените на наблюдаваните гуми са по-големи от 110.84 лв. Максималната наблюдавана цена е 167 лв.  Има няколко начина за определяне на квартилите. Те могат да бъдат прочетени в документацията на R посредством  ***help(quantile).***  Нека p ∈ [0, 1]. *Емпиричен р – квантил* се нарича онова число, за което можем да кажем, че р.100%от наблюденията са по-малки или равни от него и (1 - p)100% от наблюденията са по-големи или равни на него. От тази дефиниция е ясно, че емпиричните квантили не са определени еднозначно. Има няколко начина за тяхното определяне, които са реализирани R. По подразбиране емпиричният р – квантил се определя по следния начин. Първо де определя съответният интервал, в който се намира квантилът. Той е между k-тото и k + 1 – вото наблюдение, в подредения възходящо ред от данни, където k e най-голямото цяло число такова, че  Т.е. k = [p(n-1) + 1], където с квадратни скоби сме означили цялата част на числото в тях.  След като се определи този интервал се построява права линия, която свързва точките с координати (X(n, k); ) и (X(n, k+1); ), където X(n, k) е k-тото по големина наблюдение в подредения възходящо ред от данни, а X(n, k+1) е k+1-вото по големина наблюдение в подредения възходящо ред от данни. В уравнението на получената права заместваме ординатата с р и получаваме съответния квантил като абсциса на тази точка. Т.е.  Трябва да отбележим, че често пъти за намирането р – квантил се постъпва по следния начин. Първо се построява права линия, която свързва точките с координати (X(n, [(n+1)p]); ) и (X(n, [(n+1)p] + 1); ), където с [а] е означена цялата част на числото а. В уравнението на получената права заместваме ординатата с р и получаваме съответния квантил като абсциса на тази точка. Т.е.  Различните начини за определянето на емпиричните квантили не дават един и същ резултат. Тук е мястото на изследователя за прецени кой от начините да избере.  Да определим 0.3 квартилите на променливата Х4.  > quantile(X4, 0.3)  30%  96.671  Можем да ги закръглим например до втория знак след десетичната запетая.  > round(quantile (X4, 0.3), 2)  30%  96.67  Т.е. 30% от гумите в извадката имат цени по-малки или равни на 96.67 лв. и 70% от гумите имат по-големи или равни на тази цена.  *Задача за упражнение*  Определете квартилите, 0.9 и 0.1 квантилите на признаците  X5 - Пробег  X6 - Продължителност на живот в дни  X9 - Диаметър на джантата в цолове.  и ги анализирайте.  *Персентили(Percentile)*  Персентилите разделят статистическия ред на 100 равни части. Първо подреждаме данните възходящо. Тогава 1% от наблюденията са по-малки или равни на първия персентил. Той съвпада с 0.01 квантилът. 1% от наблюденията са по-големи или равни на първия персентил и по-малки или равни на втория персентил. Той съвпада с 0.02 квантилът и т.н. 50-тия персентил съвпада с медианата и 0.5 квантилът. 1% от наблюденията са по-големи или равни на стотния персентил, който съвпада с 0.99 квантилът. Един от начините за определяне на персентилите в R е с помощта на функцията ***quantile***.  quantile(x, probs, ...)  Да определим персентилите на променливата Х4.  > quantile (X4, probs =seq(0.01,1,0.01) )  1% 2% 3% 4% 5% 6% 7% 8% 9% 10%  84.6474 88.3466 89.0628 89.4072 89.6625 89.7488 90.0075 90.1792 90.4792 90.8680  …………………………………………………………………………………………………….  91% 92% 93% 94% 95% 96% 97% 98% 99% 100%  130.2967 133.4484 134.8675 136.8262 141.9770 148.9160 153.4809 156.8334 157.9348 167.0000  Можем да ги закръглим например до втория знак след десетичната запетая.  > round(quantile (X4, probs =seq(0.01,1,0.01) ), 2)  1% 2% 3% 4% 5% 6% 7% 8% 9% 10% 11% 12%  84.65 88.35 89.06 89.41 89.66 89.75 90.01 90.18 90.48 90.87 91.42 91.48  ……………………………………………………………………  85% 86% 87% 88% 89% 90% 91% 92% 93% 94% 95% 96%  120.13 122.98 123.15 123.75 126.59 128.24 130.30 133.45 134.87 136.83 141.98 148.92  97% 98% 99% 100%  153.48 156.83 157.93 167.00  *Задача за упражнение*  Определете персентилите на признаците  X5 - Пробег  X6 - Продължителност на живот в дни  X9 - Диаметър на джантата в цолове.  и ги анализирайте.  *Функцията summary*  Подобно на функцията ***quantile***. функцията ***summary*** определя минимум, долен квартил, медиана, средно аритметично, горен квартил и максимум.  Да определим минимум, долен квартил, медиана, средно аритметично, горен квартил и максимум на променливата Х4.  > summary(X4)  Min. 1st Qu. Median Mean 3rd Qu. Max.  82.13 94.45 101.90 105.90 110.80 167.00  *Задача за упражнение*  Определете една вградена функция минимум, долен квартил, медиана, средно аритметично, горен квартил и максимум поотделно на всеки от признаците  X5 - Пробег  X6 - Продължителност на живот в дни  X9 - Диаметър на джантата в цолове.  и ги анализирайте.  *Хинджове*  Както вече казахме квартилите разделят статистическия ред на 4 равни части. Първо подреждаме данните възходящо. Тогава 25% от наблюденията са по-малки или равни на първия квартил, който се бележи с Q1 и се нарича още долен квартил (QL). 25% от наблюденията са по-големи или равни на първия квартил и по-малки или равни на втория квартил, който съвпада с медианата. 25% от наблюденията са по-големи или равни на медианата и по-малки или равни на третия квартил (Q3), който се нарича още горен квартил и се бележи с QU. 25% от наблюденията са по-големи или равни на третия квартил. Тъй като квартилите се явяват съответно 0.25, 0.50 и 0.75 квантили в R един от начините за определяне на квартилите е с помощта на функцията ***quantile***.  quantile(x, probs, ...)  Често пъти вместо Q1 и Q3 се използват, техни приближения – така наречените хинджове. Първо се определя медианата на разпределението. Долният хиндж е приближение на първия квартил и той е медианата на по малките от медианата данни, като медианата на изходните данни не се включва в разсъждението. Горният хиндж е приближение на третия квартил и той е медианата на данните, които са по-големи от медианата на изходните данни. Отново медианата на изходните данни не се включва в разсъждението. Един от начините за определяне на хинджовете в R е с помощта на функцията ***fivenum***.  fivenum(x, na.rm = TRUE)  Подобно на функцията ***quantile***. функцията ***fivenum*** определя минимум, долен хиндж, медиана, горен хиндж и максимум.  Да определим квартилите на променливата Х4.  > fivenum(X4)  [1] 82.130 94.400 101.895 110.850 167.000  *Задача за упражнение*  Определете минимум, долен хиндж, медиана, горен хиндж и максимум на признаците  X5 - Пробег  X6 - Продължителност на живот в дни  X9 - Диаметър на джантата в цолове.  и ги анализирайте.  *Обхват(Range)*  Обхват на данните наричаме разликата между най-голямата и най-малкото наблюдение. Например  > max(X4)-min(X4)  [1] 84.87  *Междуквартилно разстояние (Interquartile Range)*  Междуквартилно разстояние наричаме разликата между третия и първия квартил. Тя може да бъде намерена с помощта на функцията IQR в R.  Например  > IQR(X4)  [1] 16.39  *Задача за упражнение*  Определете обхвата и междуквартилното разстояние на признаците  X5 - Пробег  X6 - Продължителност на живот в дни  X9 - Диаметър на джантата в цолове.  **Кутия с мустачки(Box Plot)**  Кутията с местачки е графика, която онагледява разпределението на единиците по наблюдавания признак като изобразява последователно: минимум, Q1, медиана, Q3 и максимум. Ако в извадката ни има силно отличаващи се наблюдения, които се наричат още аутлайъри (outliers), те се изобразяват като точки извън кутията с мустачки и се прави графика с мустачки на останалите наблюдения. Аутлайърите в R са тези наблюдения, които са извън интервала [].  За да видим влиянието на аутлайърите нека първо начертаем кутиите с мустачки на променливите X4, X5, X6, X9. За целта първо ще отворим ред от 4 графични прозореца с функцията ***par***. После ще използваме функцията ***boxplot***.  > par(mfrow = c(1, 4))  > boxplot(X4)  > boxplot(X5)  > boxplot(X6)  > boxplot(X9)    Наблюдаваме, че само променливата X4 има силно отличаващи се стойности. Можем да ги изключим по следния начин  X4\_without = X4 [X4 > quantile(X4, .25) - 1.5\*IQR(X4) & X4 < quantile(X4, .75) + 1.5\*IQR(X4)]  Можем да извлечем само тези силно отличаващи се стойности чрез  X4\_out = X4 [X4 < quantile(X4, .25) - 1.5\*IQR(X4) | X4 > quantile(X4, .75) + 1.5\*IQR(X4)]  X4\_out  [1] 167.00 136.05 138.49 144.29 150.02 156.40 145.33 139.15 151.61 156.83 148.87 136.72  [13] 157.69 157.00 163.64 155.14 158.17  Сега с функцията ***boxplot*** да изчертаем графиките на кутиите с мустачки на променливата X4 и същата променлива, като изключим изключим силно отличаващите се стойности. При това, за по-голяма прегледност да ги направим на два реда, да сменим шрифта и да обърнем графиките хоризонтално  > windowsFonts(times = "Times new roman")  > par(family = "times", font = 1, font.axis=2, font.main = 3)  > par(mfrow = c(2, 1))  > boxplot(X4, horizontal = TRUE, main="X4")  > boxplot(X4\_without, horizontal = TRUE, main="X4\_without")    Виждаме, че отново имаме силно отличаващи се стойности и повтаряме процедурата  > X4\_without\_without = X4\_without [X4\_without > quantile(X4\_without, .25) - 1.5\*IQR(X4\_without) &  X4\_without < quantile(X4\_without, .75) + 1.5\*IQR(X4\_without)]  > par(mfrow = c(3, 1))  > boxplot(X4, horizontal = TRUE, main="X4")  > boxplot(X4\_without, horizontal = TRUE, main="X4\_without")  > boxplot(X4\_without\_without, horizontal = TRUE, main="X4\_without\_without")    По-лесно е да се използва функцията ***rm.outlier*** за премахване на силно отличаващи се стойности.  Първо инсталираме пакета  > instal.packages(outliers)  После го зареждаме  > require("outliers")  Използваме функцията  > X4\_short= rm.outlier(X4\_without\_without)  И изчертаваме всичките графики с мустачки за да ги сравним  > par(mfrow = c(1, 4))  > boxplot(X4, main="X4")  > boxplot(X4\_without, main="X4\_without")  > boxplot(X4\_without\_without, main="X4\_without\_without")  > boxplot(X4\_short, main="X4\_short")    В много случаи, ако нямаме основание да смятаме, че силно отличаващите се наблюдения са грешни, махането им не е препоръчително, т.к. ако в много случаи те съдържат ценна информация за екстремумите на разпределението на наблюдаваната величина.  А сега да разгледаме една употреба на функцията ***boxplot***, която е изключително полезна при изследване на влиянието на качествена променлива върху количествена. Отново функцията ***boxplot*** се използва едновременно с оператора “~” за задаване на функция. От лявата страна на този оператор стои зависимата променлива, по-долу това е цената на гумите - Х4, а от дясно независимата променлива. В примера по-долу това е видът на гумите Х1. Т.е. с помощта на ***boxplot***, ще наблюдаваме дали вида на гумите влияе на тяхната цена.  > boxplot(X4~X1)  От графиките с мустачки на следващата картина наблюдаваме, че за всеки от видовете гуми средните цени са почти едни и същи.     |  | | --- | | Преди да изчертаем тези графики може, с функциите ***sort*** и ***tapply*** да подредим групите в X1 според медианите в Х4. За целта виждаме реда на подгрупите в Х1 според медианите по признака Х4 в тях  > sort(tapply(X4, X1, median))  T2 T3 T4 T1 T7 T6 T5  99.080 100.950 101.105 101.685 101.945 102.770 103.080  и дефинираме нова променлива о, която е таблица с групировката по Х1, но нивата са подредени според реда, който сме задали.  > o = ordered(X1, levels = c("T2","T3","T4","T1","T7","T6","T5"))  Изчертаваме графики с мустачки на зависимостта на Х4 от новата, подредена по наше желание променлива.  > boxplot(X4~o) |   За да видим едновременно хистограми и кутии с мустачки, например за разпределението на цените поотделно в можем да използваме функцията  ***simple.hist.and.boxplot***  от библиотеката ***UsingR***.  В случая получаваме  > library(UsingR)  > simple.hist.and.boxplot(X4)    *Задача за упражнение*  Начертайте boxplots на признаците  X5 - Пробег  X6 - Продължителност на живот в дни  X9 - Диаметър на джантата в цолове  поотделно за всеки от видовете коли.  **Честотни полигони**  Можем да заменим хистограмата с честотен полигон. Това е многоъгълникът, ограден от точките с координати (Xmin, 0), следват средите на хоризонталните отсечки в хистограмата, (Xmax, 0), и накрая се връщаме в (Xmin, 0). Преди да зададем изчертаването на полигона трябва да зададем координатите на точките, от които той е ограден. Това става най-лесно като първо изчертаем хистограмата. След това вземем от нея краищата на интервалите с компонентата ***breaks***, средите на интервалите с компонентата ***mids*** и височините с компонентата ***counts***. В нашия случай това става с  > tmp = hist(X4)  > lines(c(min(tmp$breaks), tmp$mids, max(tmp$breaks)), c(0, tmp$counts, 0), type="l")    Ако това ни е трудно можем да използваме вградената функция ***simple.freqpoly*** от библиотеката ***UsingR***. Т.е. можем да построим горната графика с:  > library(UsingR)  > simple.freqpoly(X4)  Полигонът се използва обикновено при непрекъснати метрирани признаци за да се свърже по-добре емпиричното разпределение на извадката с теоретичната плътност на наблюдаваната величина.  Подобна, но гладка крива може да бъде получена с помощта на функциите ***lines*** и ***density***, като преди това трябва да сме използвали функцията ***hist***. При нея се оценява плътността на наблюдаваната величина, като се използва осредняване на оценките по подинтервали. Ширините на тези подинтервали се задават в параметъра bw, чието съкращение идва от bandwidth. Колкото ширината на подинтервалите е по-голяма, толкова приближаващата плътността крива е по-гладка. Самият алгоритъм за построяването на гладката крива, понякога е доста сложен. В нашия пример тази крива може да бъде построена с  > hist(X4,15,prob = T)  > lines(density(X4), col='red')  или  > hist(X4,15,prob = T)  > lines(density(X4,bw="SJ"), col='red')    Aко искаме да намалим назъбеността й увеличаваме bw  > hist(X4,15,prob = T)  > lines(density(X4,bw=10), col='red')  Aко искаме да увеличим назъбеността й намаляваме bw  > hist(X4,15,prob = T)  > lines(density(X4,bw=0.8), col='red')  Толкова назъбени плътности, обаче почти не се използват на практика. *Медианно абсолютно отклонение (Median absolute deviation - MAD)* Медианната абсолютно отклонение e медианата на абсолютните стойности на отклоненията на измерените значения на признака от тяхната медиана. За да се използва като състоятелна оценка за стандартното отклонение, MAD трябва да се нормира, т.е. да се умножи по константа. Нормировката зависи от типа на разпределението. Тази нормировка произлиза от разсъжденията: ако Х е нормално разпределена с параметри μ и σ, то  = P(|X - Me| < MAD) = P( < ) = P(|Z| < MAD),  където Z e нормално разпределена случайна величина със средно μ - Me и дисперсия σ2.  Тогава  = P( < < ) =  = Φ() – Φ( - )) = 2Φ() - 1  = Φ()  Т.е. е равно на 0,75 квантилът на стандартното нормално разпределение. От тук една състоятелна оценка за стандартното отклонение е    = 1.4826(MAD – Me + μ) = 1.4826 MAD  където в последното равенство сме използвали симетрията на нормалното разпределение.  Един от начините за определяне на медианното абсолютно отклонение в R е с помощта на функцията ***mad***.  Да определим медианното абсолютно отклонение на променливата Х4.  > mad(X4)  [1] 11.69771  Друг начин е да го пресметнем ръчно.  > median(abs(X4 - median(X4)))  [1] 7.89  Често пъти тома разстояние се нормира така, че да се сравнява с нормалното  > median(abs(X4 - median(X4))) \* 1.4826  [1] 11.69771  *Задача за упражнение*  Определете медианното абсолютно отклонение на признаците  X5 - Пробег  X6 - Продължителност на живот в дни  X9 - Диаметър на джантата в цолове.  *Дисперсия (вариация, Variance)*  Дисперсията на наблюдаваната величина в цялата популация (генерална съвкупност) е средното аритметично на квадратите на отклоненията на наблюдаваните значения на признака от тяхната средна. По-точно тя се дефинира чрез формулата:  2   1-∑N       2 σ  = N    (xi - μ)        i=1  където *μ* е популационното средно, а *N* обемът на цялата популация.  По подобен начин се дефинира извадковата дисперсия, но тя е сабо близка до средното аритметично на квадратите на отклоненията на наблюдаваните значения на признака от тяхната средна аритметична. Дефинира се чрез формулата  n s2 =--1--∑  (x - ¯x)2     n - 1i=1  i  С помощта на R, вариацията на даден вектор може да се определи чрез функцията ***var***. Например вариацията на цената на гумите в извадката е  > var(X4)  [1] 262.8253  Вариациите на цените в отделните подгрупи, според вида са  > tapply(X4, X1, var)    T1 T2 T3 T4 T5 T6 T7  248.2837 409.8885 178.4393 193.2625 202.4837 377.9845 200.5164  *Задача за упражнение*  Определете вариациите на признаците  X5 - Пробег  X6 - Продължителност на живот в дни  X9 - Диаметър на джантата в цолове  поотделно за всеки от видовете коли.  *Стандартно отклонение*  Стандартното отклонение е корен квадратен от съответната вариация. Т.е. то е много близко до средната квадратична на отклоненията на измерените значения на признака от тяхната средна.  С помощта на R, стандартното отклонение на даден вектор може да се определи чрез функцията ***sd***. Например стандартното отклонение на цената на гумите в извадката е  > sd(X4)  [1] 16.21189  Стандартните отклонения на цените в отделните подгрупи, според вида са  > tapply(X4, X1, sd)    T1 T2 T3 T4 T5 T6 T7  15.75702 20.24570 13.35812 13.90189 14.22968 19.44182 14.16038  *Задача за упражнение*  Определете стандартните отклонения на признаците  X5 - Пробег  X6 - Продължителност на живот в дни  X9 - Диаметър на джантата в цолове  поотделно за всеки от видовете коли.  *Ковариация*  Ковариацията е измерител за силата на две праволинейно свързани величини. Положителната ковариация е признак за правопропорционалност, а отрицателната ковариация за обратна пропорционалност. Извадковата ковариация се определя по формулата  n s  = --1--∑  (x  − ¯x)(y − ¯y) xy   n − 1 i=1 i     i  Тя е близка до средното аритметично на произведенията на отклоненията на наблюдаваните величини от тяхната средна аритметична.  По аналогичен начин ковариацията на признаци, наблюдавани в цялата популация е  -1 N∑ σxy = N   (xi − μx)(yi − μy)         i=1  където *μx* и *μy* са съответно популационните средни на признаците Х и Y.  Ковариацията между два количествени признака може да бъде намерена с помощта на функцията ***cov*** в R. Например в извадката, описана в таблицата tires ковариацията между цената на гумата и пробега й се намира с  > cov(X4, X5)  [1] 4474.801  *Задача за упражнение*  Определете ковариациите на всеки две двойки от признаците  X4 - Цена на закупуване, актуализирана към днешна дата  X5 - Пробег  X6 - Продължителност на живот в дни  X9 - Диаметър на джантата в цолове.  *Корелация*  Корелацията е измерител за силата на две праволинейно свързани величини. За разлика от ковариацията, корелацията е нормирана, т.е. най-голямата й стойност може да бъде 1 и тогава говорим за силна правопропорционална зависимост. В този случай точките от корелационното поле са силно концентрирани върху възходяща права. Най-малката стойност на ковариацията е -1 и тована говорим за силна обратно пропорционална зависимост. При нея точките от корелационното поле са силно концентрирани върху низходяща права. Положителната корелация е признак за правопропорционалност, а отрицателната корелация за обратна пропорционалност. Колкото по-близо до нула е корелационния коефициент, толкова повече можем до кажем, че между наблюдаваните признаци отсъства праволинейна зависимост. Точките от корелационното поле не са групирани около права. При наличие на корелация трябва да се отбележи, че това не означава непременно причино-следствена връзка. Възможно е тя да е породена от трета невключена в разсъжденията величина.  Извадковата корелация се определя по формулата  s rxy =--xy      sxsy  където *sx* и *sy* са извадковите стандартни отклонения, а *sxy* е извадковата ковариация.  По аналогичен начин се определя корелацията между два признака, наблюдавани в цялата популация  ρ  = -σxy-  xy  σxσy  където *σx* и *σy* са съответно стандартните отклонения, съответно на признаците X и Y в цялата популация, а *σxy* е популационната ковариация.  Корелацията между два количествени признака може да бъде намерена с помощта на функцията ***cor*** в R. Например в извадката, описана в таблицата tires корелацията между цената на гумата и пробега й се намира с  > cor(X4, X5)  [1] 0.03172892  Тъй като тя е близко до нула, можем да кажем, че между наблюдаваните признаци няма праволинейна зависимост.  *Задача за упражнение*  Определете корелациите на всеки две двойки от признаците  X4 - Цена на закупуване, актуализирана към днешна дата  X5 - Пробег  X6 - Продължителност на живот в дни  X9 - Диаметър на джантата в цолове.  *Central Moment*  k – ти централен момент на цялата популация наричаме  1-∑N       k μk = N    (xi - μ)        i=1  k – ти централен момент на извадката наричаме  1-∑n       k mk = n    (xi - ¯x)        i=1  Тези моменти могат да бъдат намерени с помощта на функцията moment в R, която се намира в библиотеката e1071.  Първият централен момент е нула, а вторият централен момент е равен на вариацията.  Например третият централен момент на признака Цена, в таблицата tires e  > library(e1071)  > moment(X4, order=3, center=TRUE)  [1] 6862.969  *Задача за упражнение*  Определете четвъртите централни моменти на признаците  X5 - Пробег  X6 - Продължителност на живот в дни  X9 - Диаметър на джантата в цолове  поотделно за всеки от видовете коли.  *Асименрия (Skewness)*  Отклоненията в хоризонтална посока от симетричното нормално разпределение се наричат асиметрия. Количествено можем да я измерим с моментния коефициент на асиметрия  γ1 = μ3∕μ3∕22  където *μ*2 и *μ*3 са съответно втория и третия централен момент. Да припомним, че вторият централен момент съвпада с дисперсията. Тогава знаменателя в горната формула представлява третата степен на стандартното отклонение.    Отрицателната асиметрия (- .. -) е белег, че средното на данните е по-малко от медианата. В този случай говорим за лява асиметрия. Имаме по-голямо струпване на статистически единици около по-големите значения на наблюдавания признак.  Положителната асиметрия (…) е белег, че средното на данните е по-голямо от медианата. В този случай говорим за дясна асиметрия. Имаме по-голямо струпване на статистически единици около по-малките значения на наблюдавания признак.  Моментният коефициент на асиметрия може да бъде намерен с помощта на функцията ***skewness*** в R.  Например моментният коефициент на асиметрия на признака Цена, в таблицата tires e  > skewness (X4)  [1] 1.610687  *Задача за упражнение*  Определете моментите коефициенти на асиметрия на признаците  X5 - Пробег  X6 - Продължителност на живот в дни  X9 - Диаметър на джантата в цолове.  *Ексцес(Kurtosis)*  Eксцес значи още изостреност. Т.к. стандартният нормален ексцес е три, за да сравняваме с нула в следващата формула имаме минус три. Коефициентът на ексцес се намира по формулата  γ2 = μ4∕μ22 - 3  където *μ*2 и *μ*4 са съответно вторият и четвъртият централен момент.  Когато коефициентът на ексцес е по-голям от нула, разпределението на наблюдаваната величина е по-високо … от нормалното. В такива случаи говорим за **leptokurtic** разпределения. При тях наблюдаваните значения на признака са силно концентрирани около тяхната средна аритметична.  Когато коефициентът на ексцес е по-малък от нула, разпределението на наблюдаваната величина е по-ниско (--) от нормалното. В такива случаи говорим за **platykurtic** разпределения. При тях имаме по-големи разсейвания на наблюдаваните значения на признака около тяхната средна аритметична.    При стандартното нормално разпределение ексцесът е нула. За такива разпределения казваме, че са **mesokurtic***.*  Коефициент на ексцес може да бъде намерен с помощта на функцията ***kurtosis*** в R.  Например коефициент на ексцес на признака Цена, в таблицата tires e  > kurtosis (X4)  [1] 2.520156  *Задача за упражнение*  Определете коефициентите на ексцес на признаците  X5 - Пробег  X6 - Продължителност на живот в дни  X9 - Диаметър на джантата в цолове. |
|  |
| |  | | --- | |  | |