

Статистически анализ на данни от кръвта на атлети

проект по
Вероятности и статистика

Изготвил: Калоян Стоилов, ф.н. 81609,
спец. КН, курс 3, група 5

*СОФИЙСКИ УНИВЕРСИТЕТ
"СВ. КЛИМЕНТ ОХРИДСКИ"*



ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА

Съдържание

I	Информация за данните	1
II	Цели на проекта	2
III	Статистики	2
IV	Хистограми	3
IV.1	Хистограми на LBM	4
IV.2	Хистограми на концентрации на бели кръвни телца	5
V	Изследване на разпределенията	6
V.1	Нормална разпределеност	6
V.2	Сравнение на LBM при двата пола атлети	7
V.3	Сравнение на концентрацията на бели кръвни телца при двата пола атлети	7
VI	Зависимости между LBM и концентрацията на бели кръвни телца при атлетите	8
VI.1	Корелационни коефициенти	8
VI.2	Диаграми на разсейване и регресия	8
VII	В заключение	10

I Информация за данните

Данните са взети от Rdatasets, Package DAAG, Item ais. Данните са относно австралийски атлети от различни дисциплини и са направени измервания на различни характеристики (концентрация червени/бели кръвни телца, маса, пол, спорт и т.н.). За целите на поставената задача се ограничаваме до изследване на номиналната категорийната променлива пол (*sex*) с възможни стойности **мъж**(m) и **жена**(f), както и на двете непрекъснати числови променливи: маса без масата на мастната тъкан (още известна като LBM) в **кг**(*lbm*); концентрация на бели кръвни телца в **10⁹ за литър**(*wcc*).

Ще отбележим, че няма повторни измервания, тъй като при тях има ограничения на множеството използвани статистически тестове.

II Цели на проекта

В проекта ще направим таблично и/или графично представяне на основните статистики и данните. Опитваме да отговорим на следните въпроси:

- Нормално разпределени ли са LBM и концентрацията на бели кръвни телца при атлетите (въобще, при мъжете и при жените)?
- Има ли статистически значима разлика между концентрацията на бели кръвни телца при мъжете и жените атлети?
- Има ли статистически значима разлика между LBM при мъжете и жените атлети?
- Има ли корелация между концентрацията на бели кръвни телца и LBM при атлетите (въобще, при мъжете и при жените)?
- Има ли регресионна права за предвиждане на концентрацията на белите кръвни телца чрез LBM?

III Статистики

За получаванете им са използвани вградените в R функции `mean`, `median`, `sd` и `var`. Тъй като в R не е вградена функция за мода на извадка, е използвана предоставената във упражнение функция `modeFunction` с малка промяна, за да връща реално число.

```
1 modeFunction <- function(x) {  
2     tt <- table(x)  
3     return(as.double(names(tt)[tt==max(tt)]))  
4 }
```

За бройките съответно на всички атлети и отделните полове използваме вградените функции `length` и `table`. Атлетите са 202, като 100 от тях са жени, а 102 - мъже.

За взимане само на записите за отделен пол използваме:

```
1 a1sf=ais[ais$sex == 'f',]  
2 a1sm=ais[ais$sex == 'm',]
```

Така достигнахме до следната таблица с дескриптивни статистики:

Извадка	Средна	Медиана	Мода ¹	Дисперсия	Стандартно отклонение
LBM-всички	64.87371	63.035	78	170.8301	13.0702
Конц. бели кръвни телца-всички	7.108911	6.85	6.4	3.241214	1.800337
LBM-мъже	74.656863	74.5	78	97.752378	9.886980
Конц. бели кръвни телца-мъже	7.221569	7.1	7.5 и 8.9	3.606857	1.899173
LBM-жени	54.8949	54.92	53.11, 53.2 и 56.05	47.916748	6.922192
Конц. бели кръвни телца-жени	6.994000	6.7	6.4	2.874509	1.695438

Таблица 1: Таблица с дескриптивни статистики

IV Хистограми

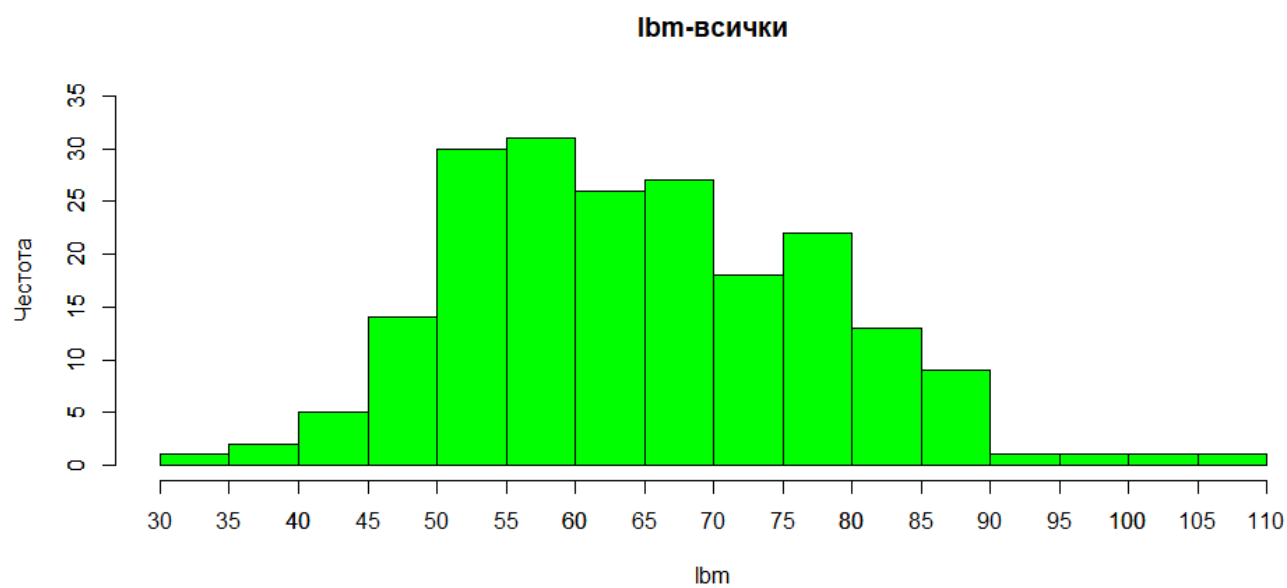
Хистограмите са изпечатани с написана функция `hister`(картината долу), използваща вградените `hist` и `axis`. Причината е, че базово интервалите на x координатата зависят от предоставената информация. Така е и с краищата ѝ. Това обаче може да даде грешна представа за разположението на подмножествата от данните за двата пола (едно спрямо друго, както и спрямо данните от всички атлети). Поради желанието за еднакви краища и интервали навсякъде, за всяка от числовите променливи се решава какви да са те, в зависимост от минималните и максималните стойности при всички.

```

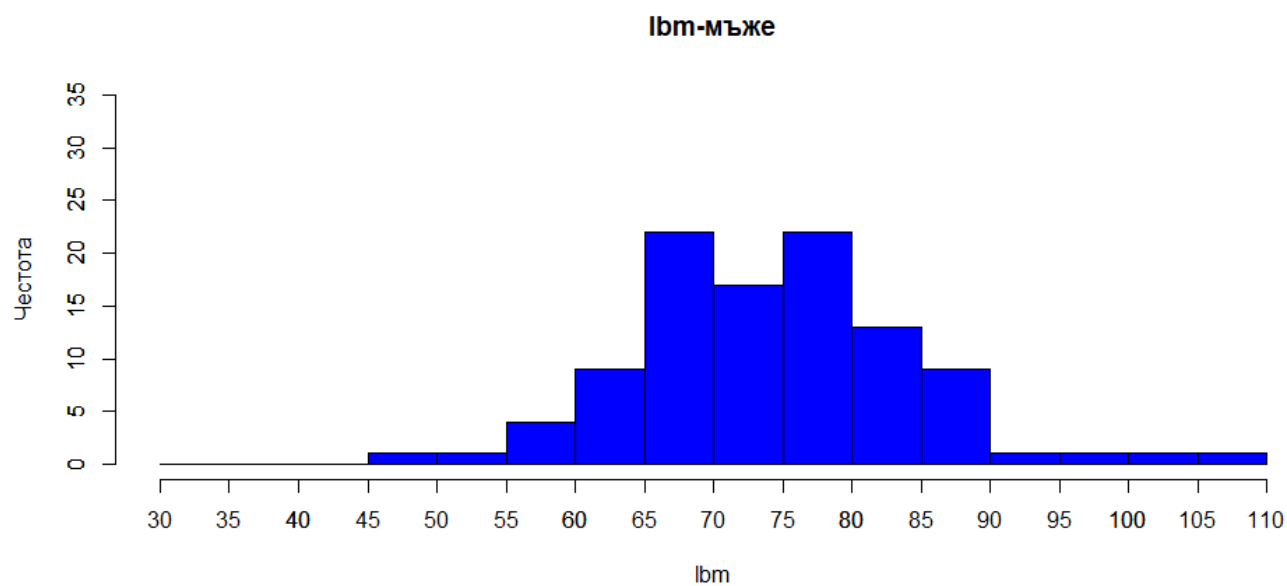
1 hister<-function(info, mname, xname, yname, hcol, xl, yl, xm, ym, xint, yint, brnum){
2   hist(info, main=mname, xlab=xname, ylab=yname, col=hcol,
3       breaks = brnum, xlim=c(xl,xm), ylim=c(yl,ym))
4   axis(side=1,at=seq(xl,xm,xint))
5   axis(side=2,at=seq(yl,ym,yint))
6 }
```

¹Зачитаем, че при повече от една най-често срещана стойност, и тя е мода.

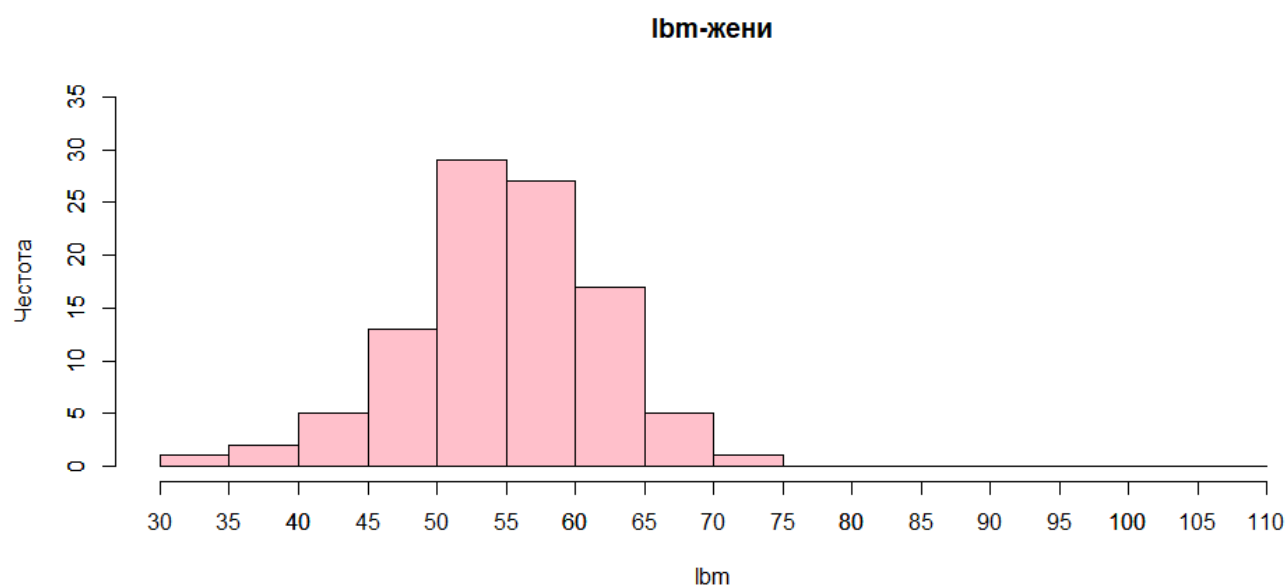
IV.1 Хистограми на LBM



Фигура 1: Хистограма на маса без масната тъкан за всички атлети



Фигура 2: Хистограма на маса без масната тъкан за мъжете атлети



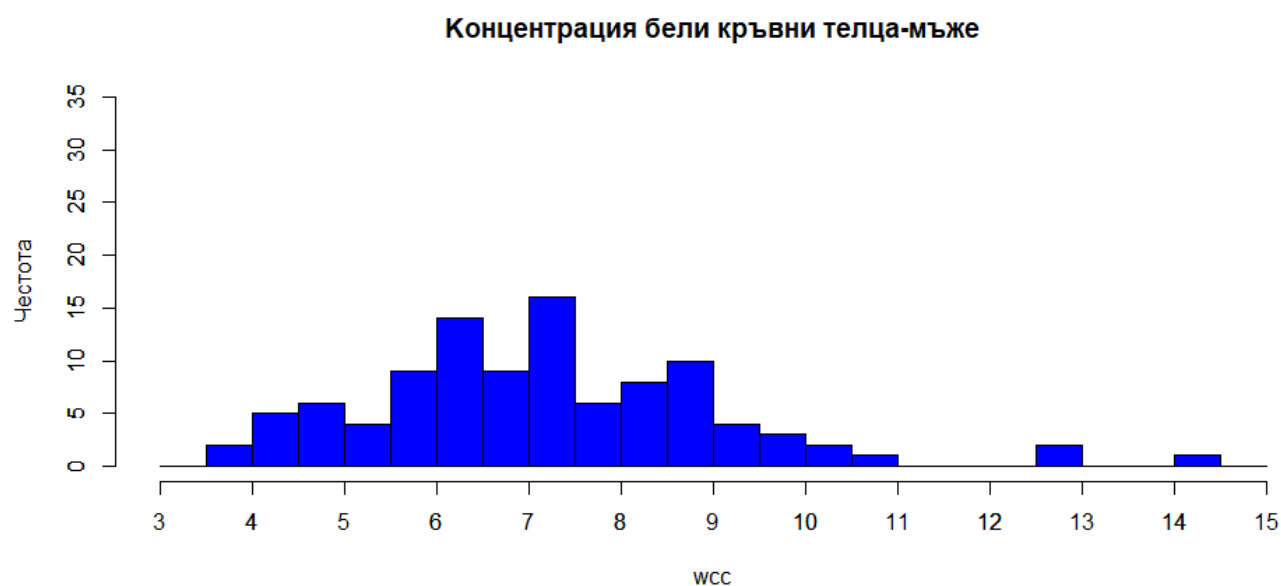
Фигура 3: Хистограма на маса без масната тъкан за жените атлети

IV.2 Хистограми на концентрации на бели кръвни телца²



Фигура 4: Хистограма на концентрацията на бели кръвни телца за всички атлети

²Поради възникнали проблеми с визуализацията на хоризонталната ос при разделители реални числа, тя е разграфена през 1, а интервалите за хистограмата са с дължина 0.5.



Фигура 5: Хистограма на концентрацията на бели кръвни телца за мъжете атлети



Фигура 6: Хистограма на концентрацията на бели кръвни телца за жените атлети

V Изследване на разпределенията

V.1 Нормална разпределеност

Тук изследваме дали извадките са нормално разпределени, чрез тестът на Shapiro-Wilcoxon (с $\alpha = 0.05$). Използвана е вградената функция `shapiro.test`. Ре-

зультатите са представени в следната таблица:

Извадка	p	Нормално разпределена
LBM-всички	0.1286	Не
Конц. бели кръвни телца-всички	0.00002591	Не
LBM-мъже	0.2175	Да
Конц. бели кръвни телца-мъже	0.01237	Не
LBM-жени	0.468	Да
Конц. бели кръвни телца-жени	0.001129	Не

Таблица 2: Таблица с корелационни коефициенти

V.2 Сравнение на LBM при двата пола атлети

Тъй като LBM при мъжете и жените са нормално разпределени, може да приложим t-тест на Welch за тяхно сравнение. От хистограмата изглежда, че може би LBM при мъжете е с по-голяма средна. За това решаваме тестът да е с:

H_0 Мъжете и жените атлети имат еднакви средни стойности LBM.

H_1 Средната стойност на LBM при мъжете атлети е по-голяма от тази на жените атлети.

Тоест правим едностранен t-тест, като нека $\alpha = 0.05$. Използвайки вградената в R функция `t.test`, достигаме до резултат $\mathbf{p} < 2.2 \times 10^{-16}$. Нулевата хипотеза се отхвърля. Достигаме до заключението, че има статистически значима разлика между LBM при мъжете и жените атлети, като средностатистически мъжете атлети имат по-голям LBM.

V.3 Сравнение на концентрацията на бели кръвни телца при двата пола атлети

Видяхме, че концентрацията на бели кръвни телца при двата пола не са разпределени нормално. Поради това не може да използваме t-тест за тяхното сравнение. Ще се наложи да използваме някой непараметричен тест. Тъй като броят на тествани мъже е различен от този на жените, ще приложим тестът Mann-Whitney U/Wilcoxon rank sum:

H_0 Няма разлика между концентрациите на бели кръвни телца при мъжете и жените атлети.

H_1 Налице е разлика между концентрациите на бели кръвни телца при мъжете и жените атлети.

Тоест правим двустранен U тест, като нека $\alpha = 0.05$. Използвайки вградената в R функция `wilcox.test`, достигаме до резултат $p = 0.3853$. Нулевата хипотеза не се отхвърля. Достигаме до заключението, че няма статистически значима разлика между концентрациите на бели кръвни телца при мъжете и жените атлети.

VI Зависимости между LBM и концентрацията на бели кръвни телца при атлетите

VI.1 Корелационни коефициенти

Използваме вградената функция `cor`, за да получим корелацията между LBM и концентрацията на белите кръвни телца при всички атлети, както и само при мъжете и само при жените. Резултатите са представени в следната таблица:

Извадка	Корелационен коефициент	Интерпретация
Всички	0.1026625	Много слаба положителна корелация
Мъже	0.1067	Много слаба положителна корелация
Жени	0.04830104	Няма корелация

Таблица 3: Таблица с корелационни коефициенти

VI.2 Диаграми на разсейване и регресия

За да представим зависимостта, ще използваме написана от нас функция `scatterer`, показана по-долу, като функциите, използвани в дефиницията ѝ са само вградени. Тя рисува диаграма на разсейването с `plot`. След това намира линейна регресия чрез `lm`. Ако регресията е статистически значима, чертае графиката ѝ, а иначе изпечатва противното в конзолата.

```

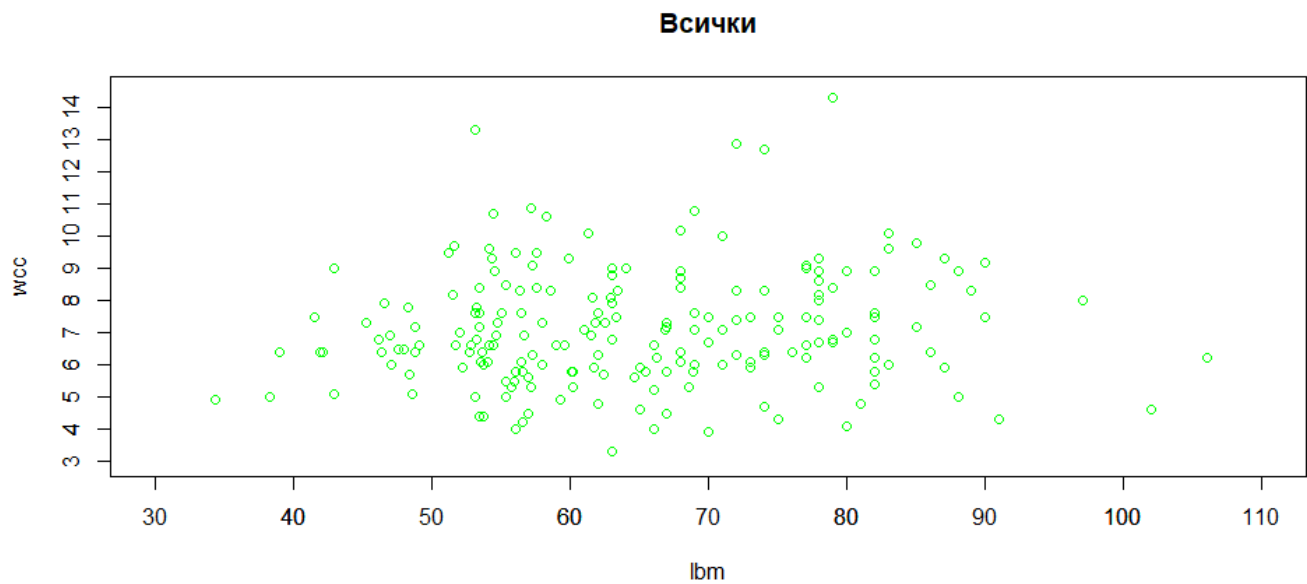
1 scatterer<-function(info1, info2, mname, xname, yname, scol,
2                       xl, yl, xm, ym, xint, yint, alpha){
3   plot(info1, info2, main=mname, xlab=xname, ylab=yname, col=scol,
4         xlim=c(xl, xm),ylim=c(yl, ym))
5   axis(side=1, at=seq(xl, xm, 2*xint))
6   axis(side=2, at=seq(yl, ym, 2*yint))
7
8   regression=lm(info2~info1)
9   regsum=summary.lm(regression)
10  pvalue=regsum$coefficients[2, 4]
11

```

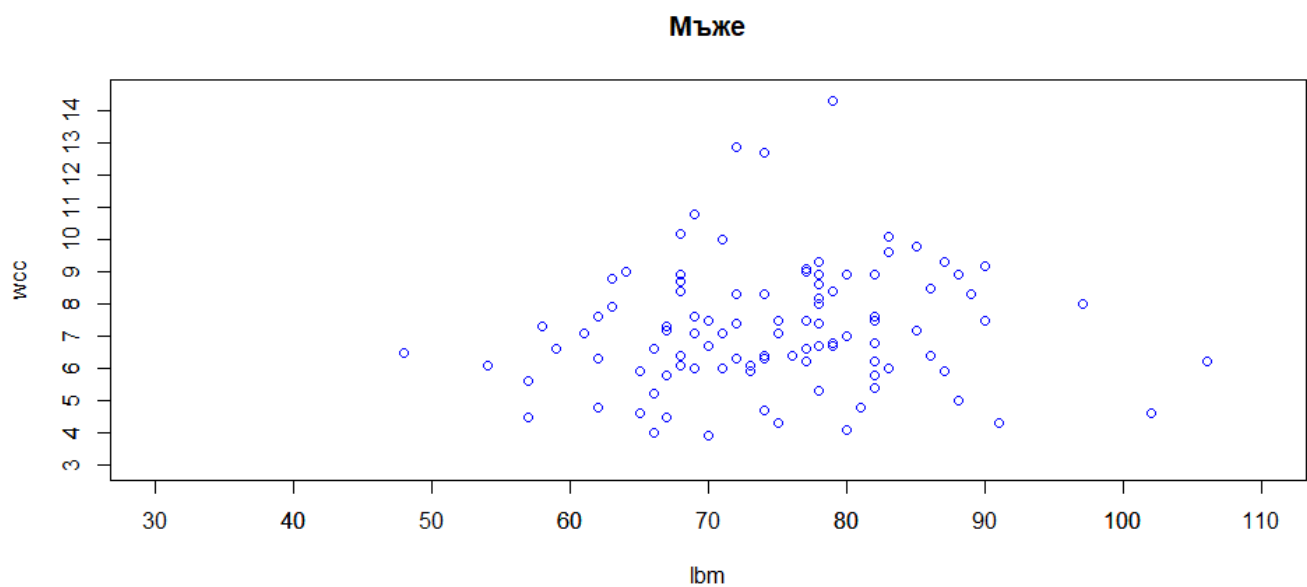
```

12 if(pvalue<=alpha) abline(regression,col=scol)
13 else print("Linear regression is not significant")
14 }

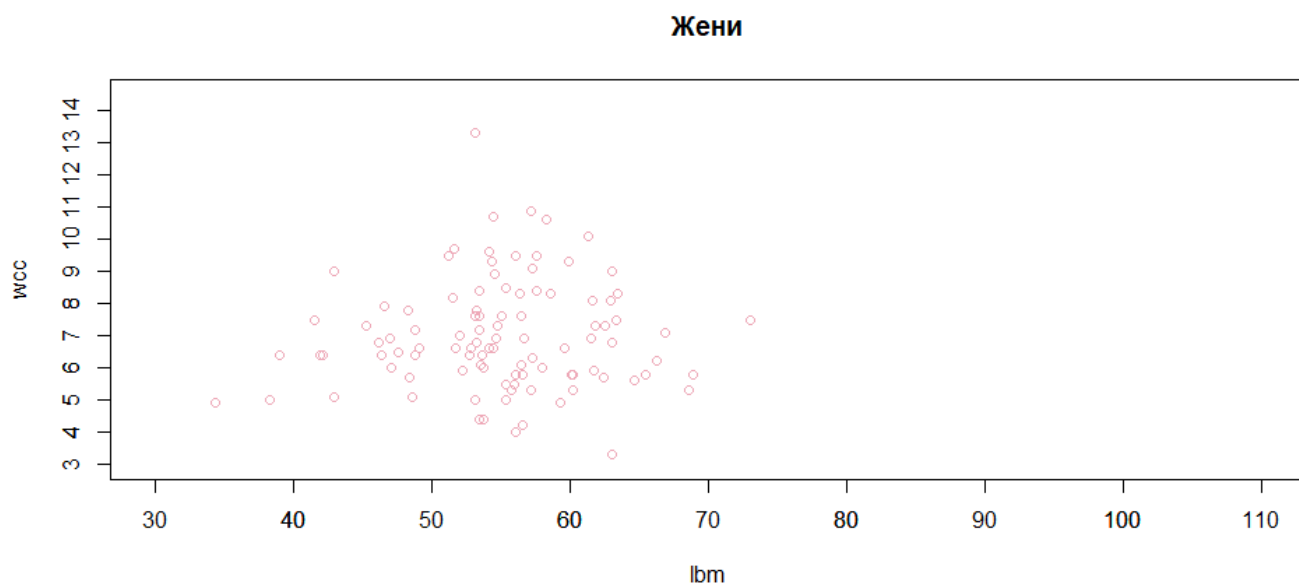
```



Фигура 7: Диаграма на разсейването за всички атлети



Фигура 8: Диаграма на разсейването за мъжете атлети



Фигура 9: Диаграма на разсейването за жените атлети

Диаграмите показват, че между LBM и концентрацията на бели кръвни телца няма статистически значима линейна зависимост (при ниво на значимост $\alpha = 0.05$).

Съдейки по формата на диаграмите на разсейване, по-скоро няма функционална зависимост от какъвто и да било тип.

VII Заключение

Следните заключения правим при разумното предположение, че няма някаква голяма разлика във физическото устройство между австралийските атлети и атлетите от другите страни.

На база на представените данни може да твърдим ³, че:

1. LBM при мъжете и жените атлети отделно е нормално разпределен, но не е, ако ги разглеждаме заедно, като средностатистически атлет от мъжки пол има по-голям LBM.
2. Концентрацията на белите кръвни телца не е нормално разпределена при атлетите, дори и разглеждайки половете поотделно. Няма разлика в концентрацията на белите кръвни телца между двата пола.

³Хистограми на концентрации на бели кръвни телца

3. Между LBM и концентрацията на белите кръвни телца при атлетите няма забележима корелация. Регресионна права за изразяване на концентрацията на белите кръвни телца чрез LBM няма.