Проект по Статистика и емпирични методи практикум

Иван Чучулски ф.н. 62167

1. Въведение

1.1. Цели на проекта

Целта на проекта е да се изследват зависимости между измервания на физически показатели на различни хора. Искаме да разберем как пола влияе върху показателите и дали има взаимодействие между стойностите на измерванията.

1.2. Описание на данните

Ще използваме данните survey от пакета MASS. Те представляват отговори на анкетно проучване, проведено сред студенти в университета в Аделаида, Австралия. Ние ще се разгледаме следните колони на data frame-а :

- пол на анкетираните, категорийна номинална променлива
- ръст на анкетираните, числова непрекъсната
- педя, т.е. дължина на дланта на ръката, числова непрекъсната

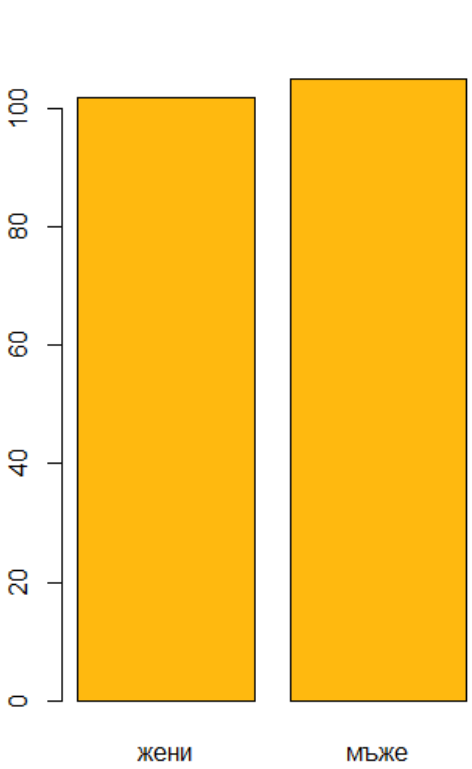
Първо изследваме как са разпределени данните поотделно, техните локация и разсейване. След това ще видим дали категорийната променлива може да е обясняваща за всяка от числовите и ще търсим дали има връзка между ръста на даден човек и дължината на неговата длан. Също така като ще проверим дали при разглеждане на наблюдения само върху мъже или жени има по-силна или слаба зависимост.

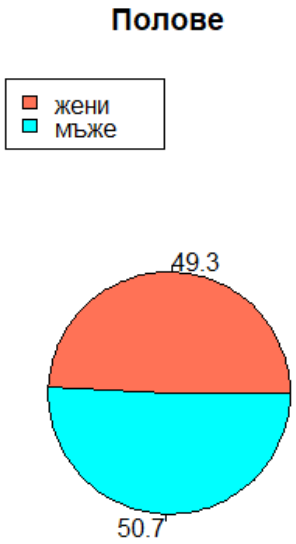
Като забележка може да се каже, че има редове в данните, където някоя от стойностите липсва, т.е. има NA. Преди да започнем анализа премахваме тези редове, където поне някоя от трите стойности липсва.

2. Изследване на променливите поотделно

2.1. пол, категорийна номинална

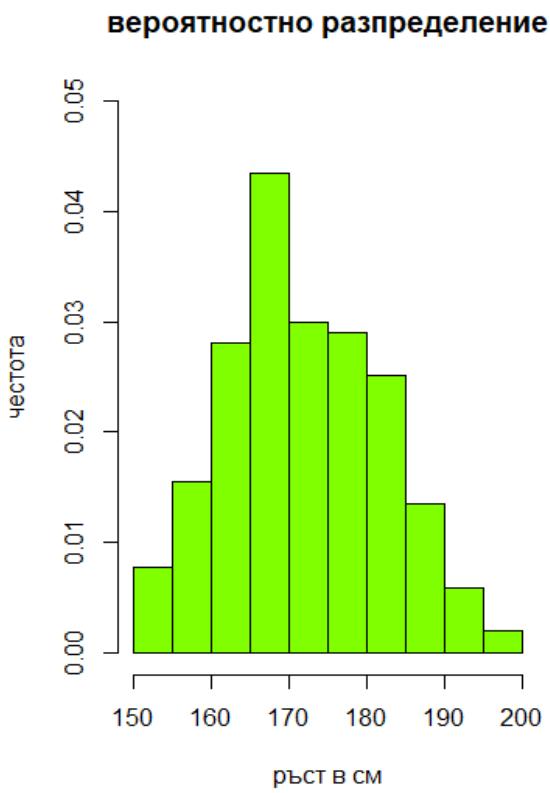
От графиката можем да видим, че в анкетираните имаме поравно мъже и жени.

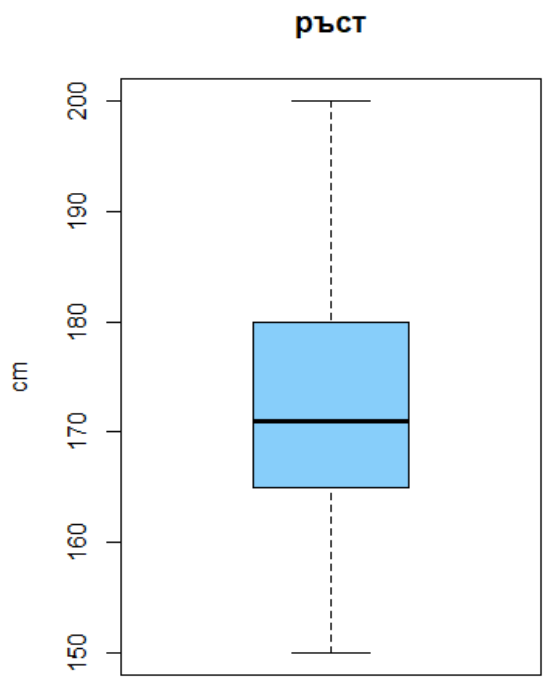


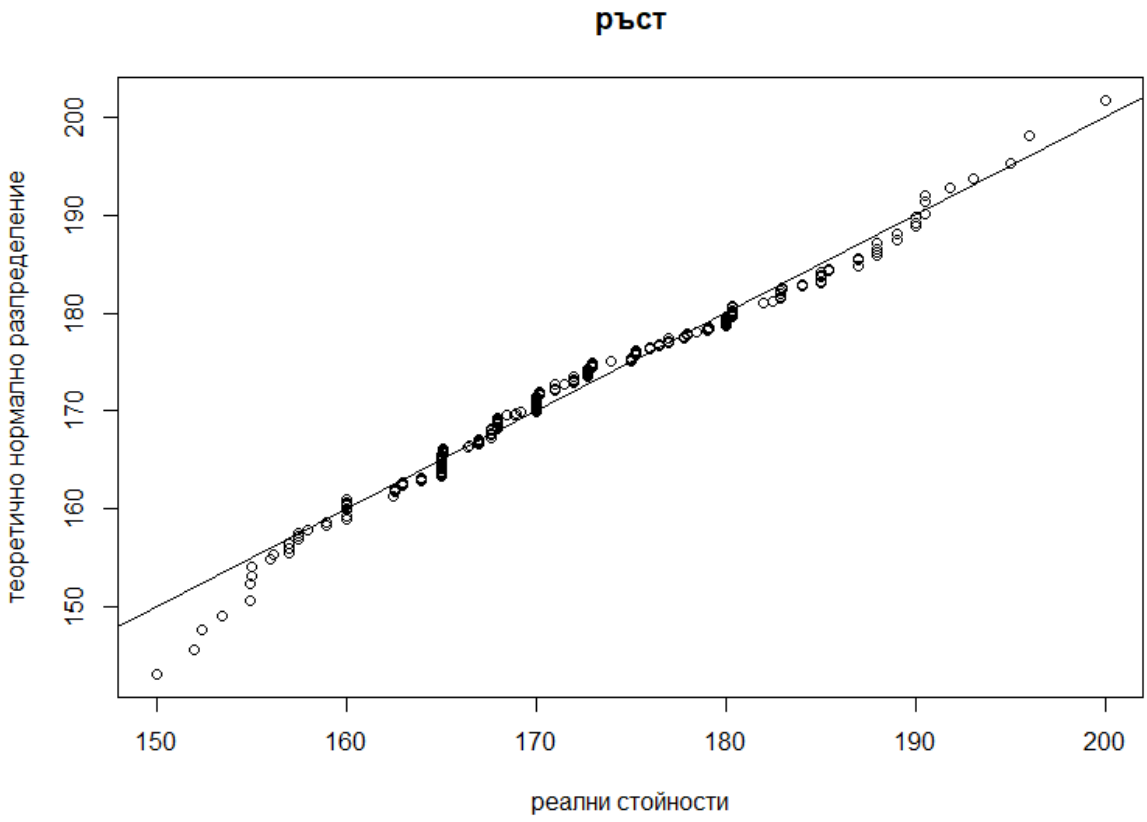


Това се потвърждава и от командата summary, където виждаме, че жените са 102, а мъжете са 105. Това е добре, защото ако единия пол имаше сериозен превес над другия, това може да окаже ефект и върху стойностите на другите измервания, например мъжете по принцип са по-високи и така ръстовете на жени щяха да са outlier-и в разпределението. Равният брой ще ни позволи да разглеждаме стойности на някоя от другите променливи върху мъже и жени и да направим изводи за взаимодействието между пола и променливата.

2.2. ръст, числова непрекъсната







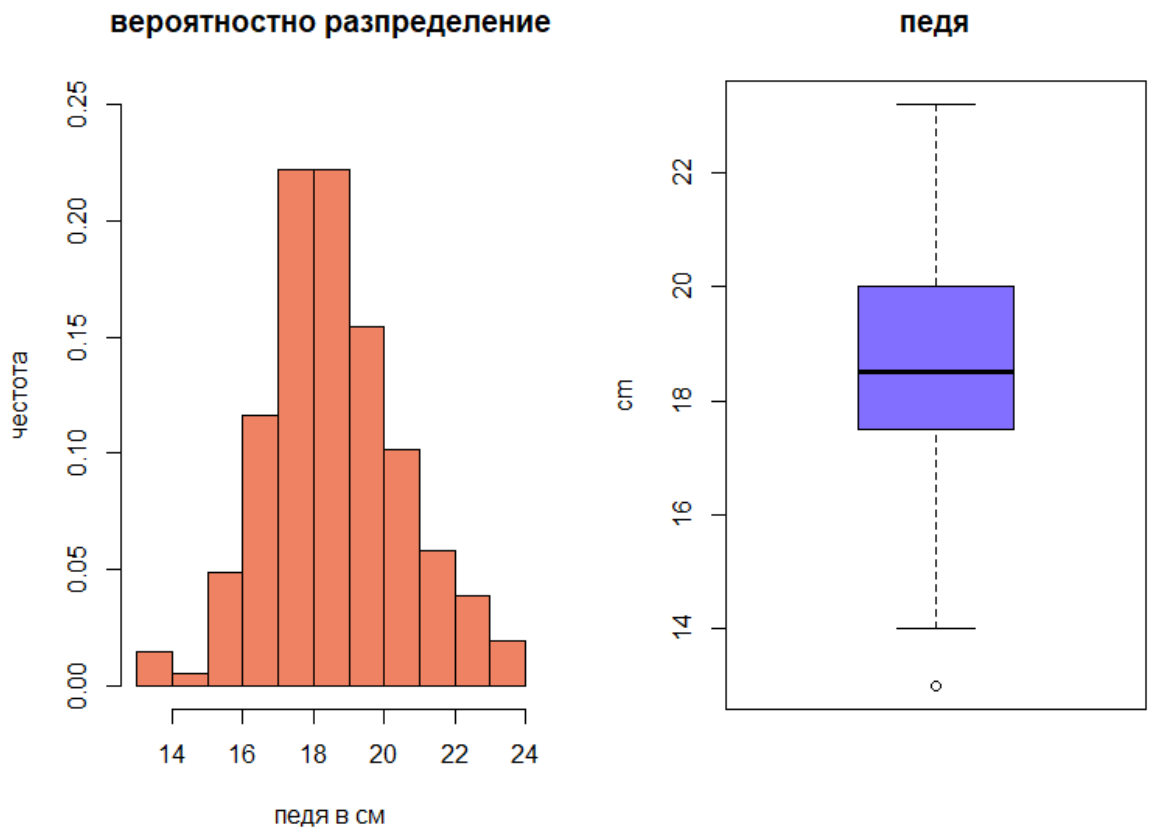
От приложените графики можем за кажем, че разпределението изглежда като нормално, не са налице outlier-и. Убеждаваме се в това и като направим тест за нормално разпределение на Shapiro-Wilk. Задаваме нивото на съгласие на 0.05. Резултата от теста за p-value е 0.08102 > 0.05. Следователно имаме нормално разпределение.

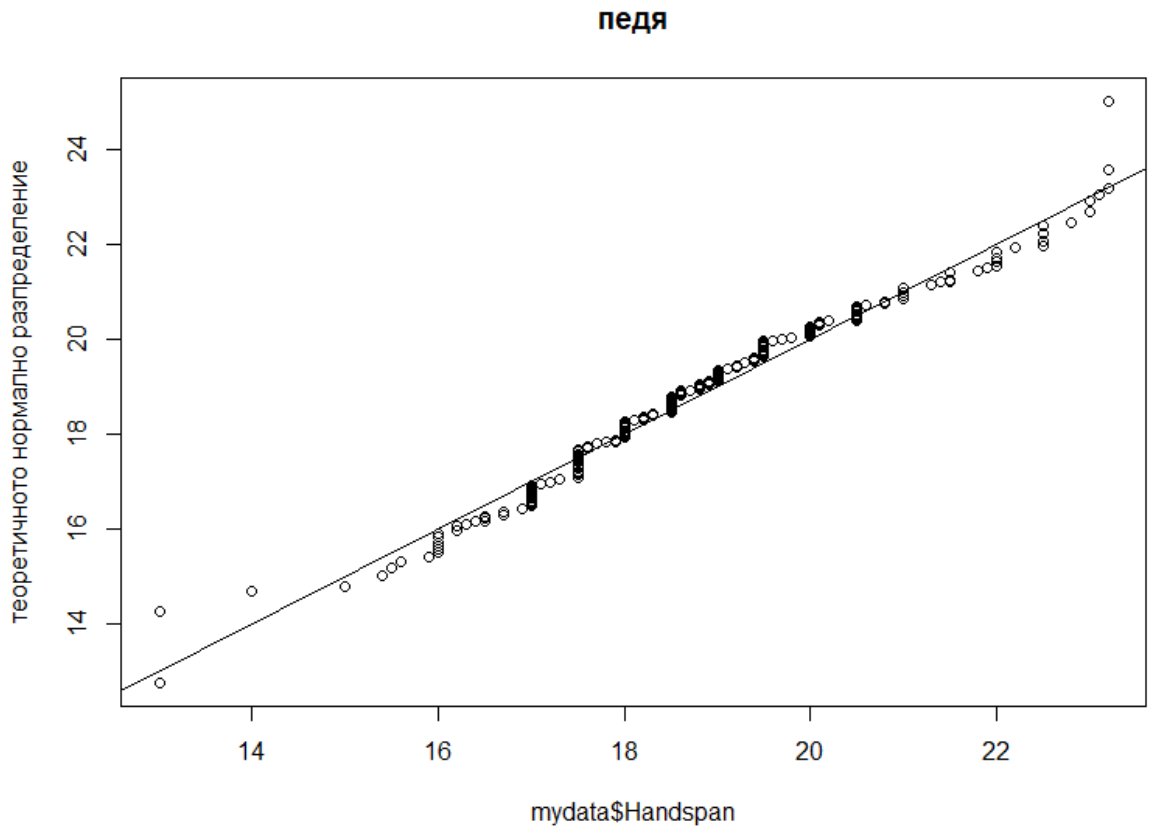
Използваме параметрични оценки за намиране на локацията със и дисперсията, т.е. средно аритметично и стандартно отклонение.

- локация = round(mean(mydata$Height), 3) = 172.385
- дисперсия = round(sd(mydata$Height), 3) = 9.895

Локацията е колкото средния ръст за човек, а от хистограмата и стойността на дисперсията показва, че данните са равномерно разпределени в интервал средната стойност +- дисперсията.

2.3. педя, числова непрекъсната





От хистограмата можем да видим, че разпределението се доближава до нормалното, но дължините на опашките в boxplot-а както и наличието на изкривявания в qqplot-а говорят, че може и да нямаме нормално разпределение.

При прилагане на тест за нормално разпределение Shapiro-Wilk отново с ниво на съгласие 0,05 обаче получаваме резултат за p-value = 0.003831 < 0.05. Следователно нямаме нормално разпределение.

За локацията и дисперсията използваме непараметрични оценки медиана и mean absolute deviation.

- локация = round(median(mydata$Handspan), 3)= 18.5
- дисперсия = round(mad(mydata$Handspan), 3)= 1.483

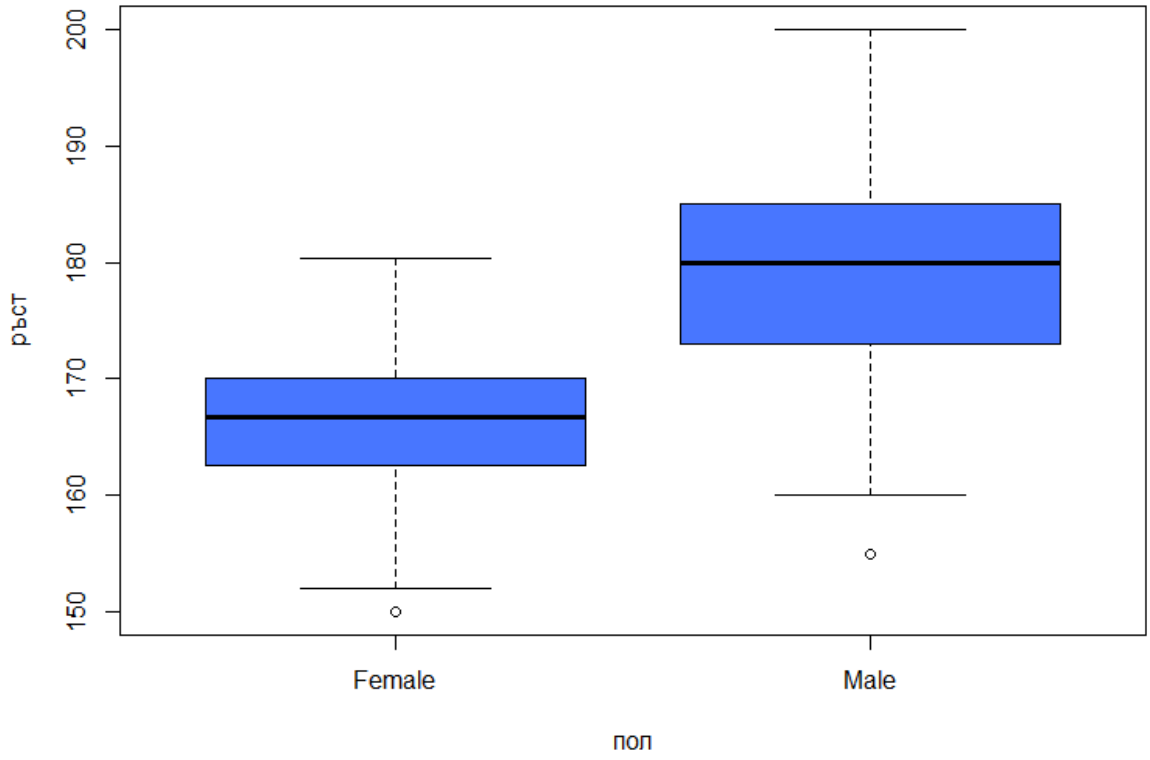
Локацията е нормална стойност за дължина на дланта, а неголямата стойност на дисперсията показва, че измерванията са струпани около медианата.

3. Изследване на взаимодействия между променливите

3.1. категорийни обясняващи и числови зависими

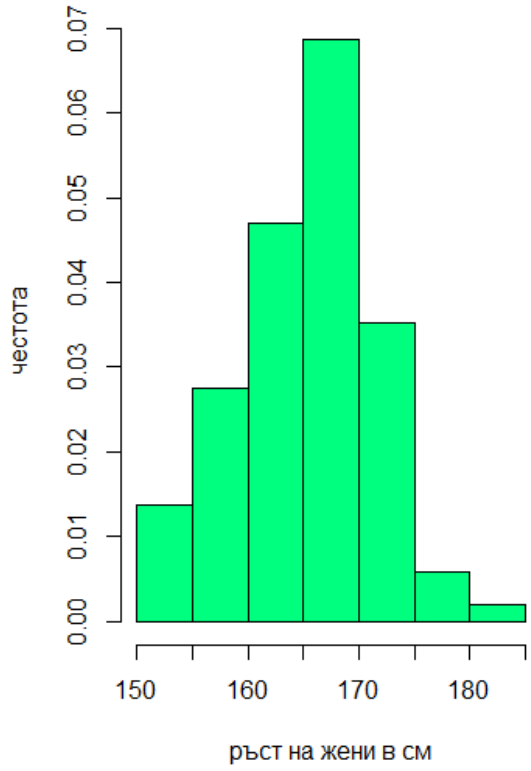
3.1.1. пол и ръст

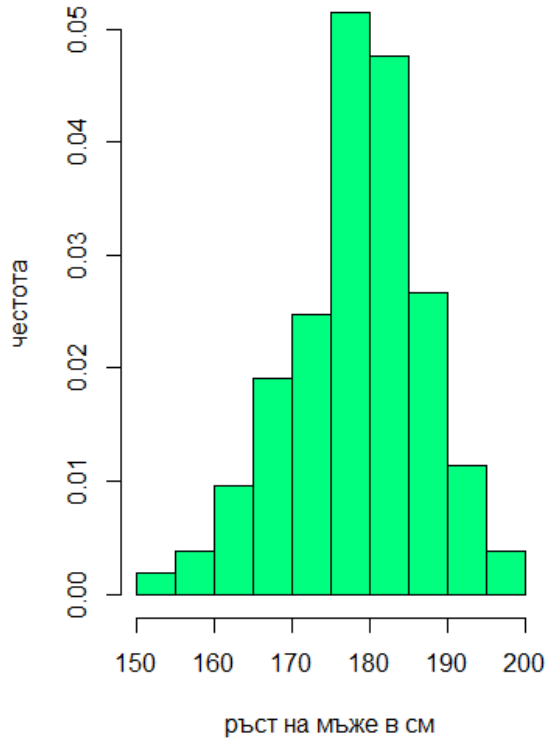
Първо да видим взаимодействието между пола и ръста.



От графиката виждаме, че по-голямата част от жените са по-ниски от мъжете, защото обхвата на първия и третия квартил на женския ръст строго под първия квартил на мъжките измервания. Също така можем да забележим, че най-високите стойности на женския ръст са равни или малко над медианата на мъжкия, а има и мъж, който е по-нисък от 50-те процента на ръста на жените. Както видяхме по-рано броят на наблюденията е поравно между половете, можем да кажем, че резултатите отговарят на действителността, а именно, че полът е обясняваща променлива за ръста.

Ако разгледаме данните, като разделим височината на жените и мъжете в отделни променливи от хистограмата изглежда, че те са с нормалното разпределение.



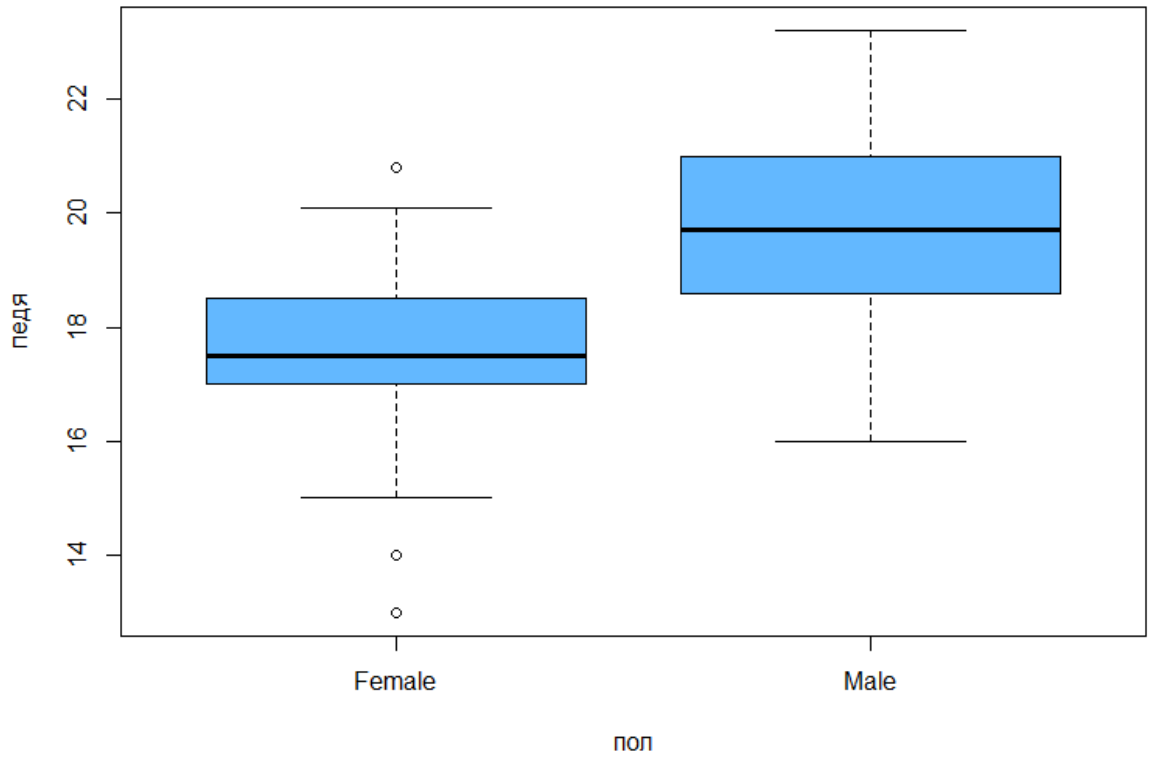


Като приложим тестове за нормално разпределение потвърждаваме хипотезата, че женския и мъжкия пол поотделно са нормално разпределени.

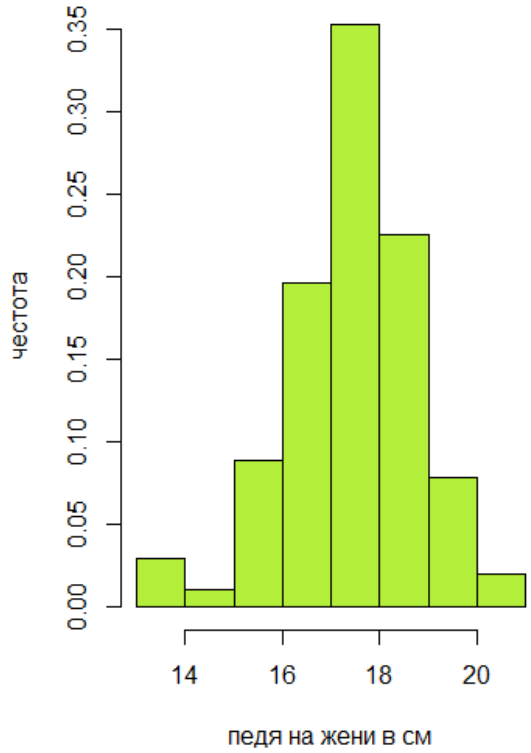
- shapiro.test(fem_heights)
  - p-value = 0.1313 > 0.05 = alpha
- shapiro.test(male_heights)
  - p-value = 0.7162 > 0.05 = alpha

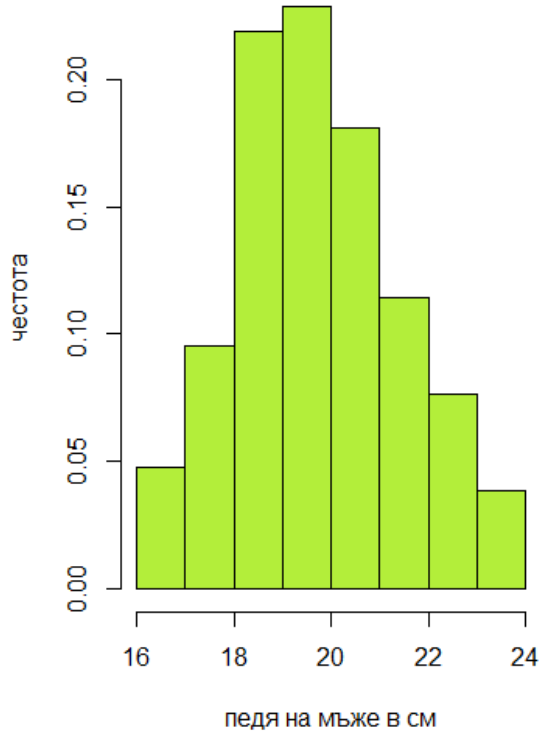
3.1.2. пол и дължина на дланта

Нека да видим взаимодействието между пол и педя



В графиката се наблюдава подобно нещо на това, което присъстваше и при сравнението на пола и ръста. По-голямата част от измерванията на педята на жените, т.е. обхватът между първи и трети квартил е с по-малка дължина от първия квартил на мъжките измервания. Това, което се различава е наличието на няколко по-високи стойности на педите при жените, които са над медианата на мъжките, както и няколко измервания под опашката от минимални стойности. Тези наблюдения и факта, че педята не беше с нормално разпределение, поставят под въпрос дали данните са нормално разпределени, ако разгледаме дължините на дланите само на жени и мъже.





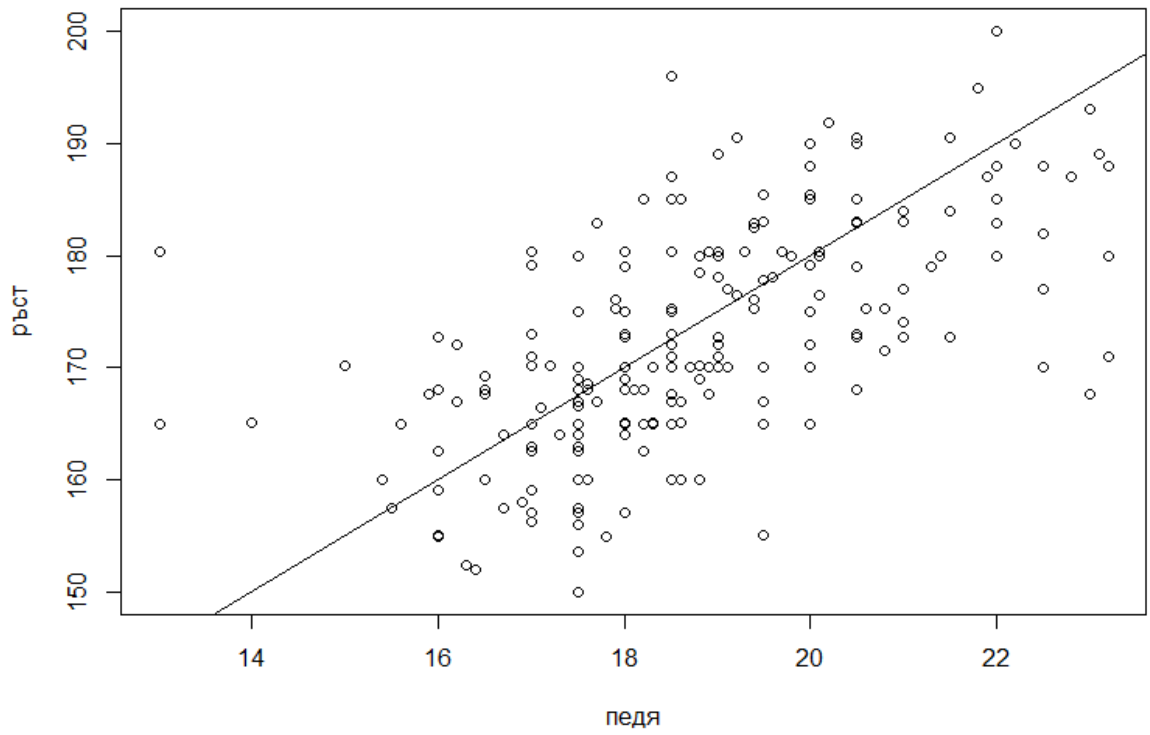
При хистограмата на мъжката педя разпределението наподобява нормалното, докато при жените имаме лява асиметрия. При прилагане на тест за нормално разпределение с ниво на съгласие 0,05 получаване следните резултати :

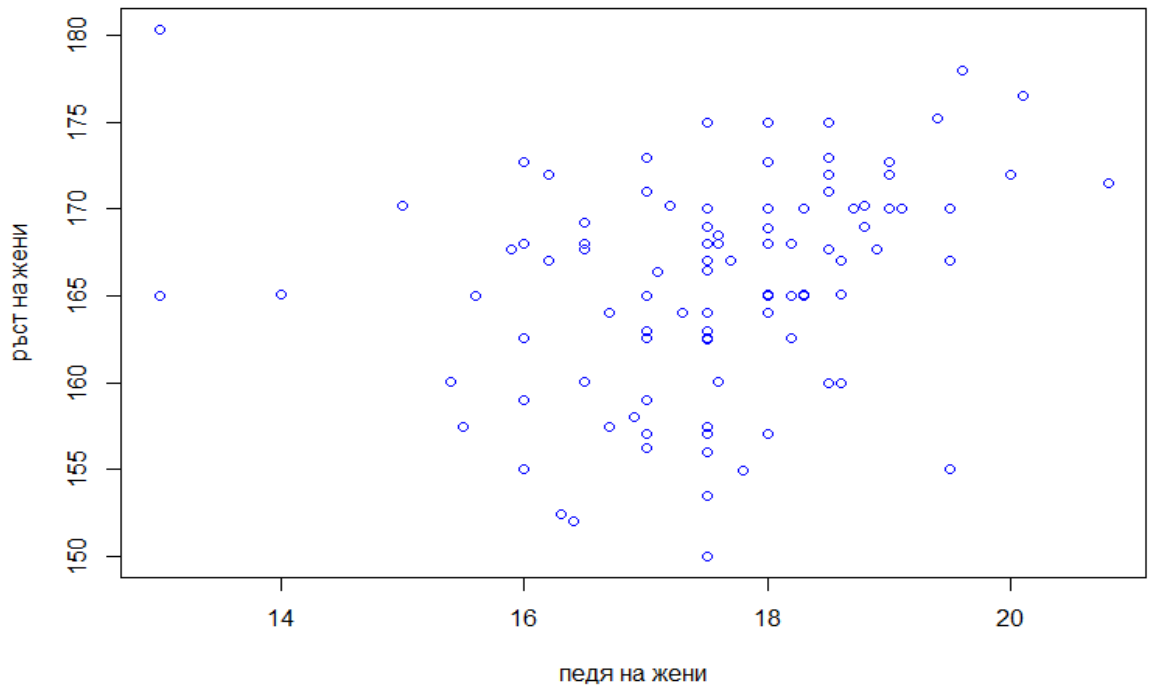
- shapiro.test(fem_handspan)
  - p-value = 0.002367 < 0.05 = alpha
- shapiro.test(male_handspan)
  - p-value = 0.06273 > 0.05 = alpha

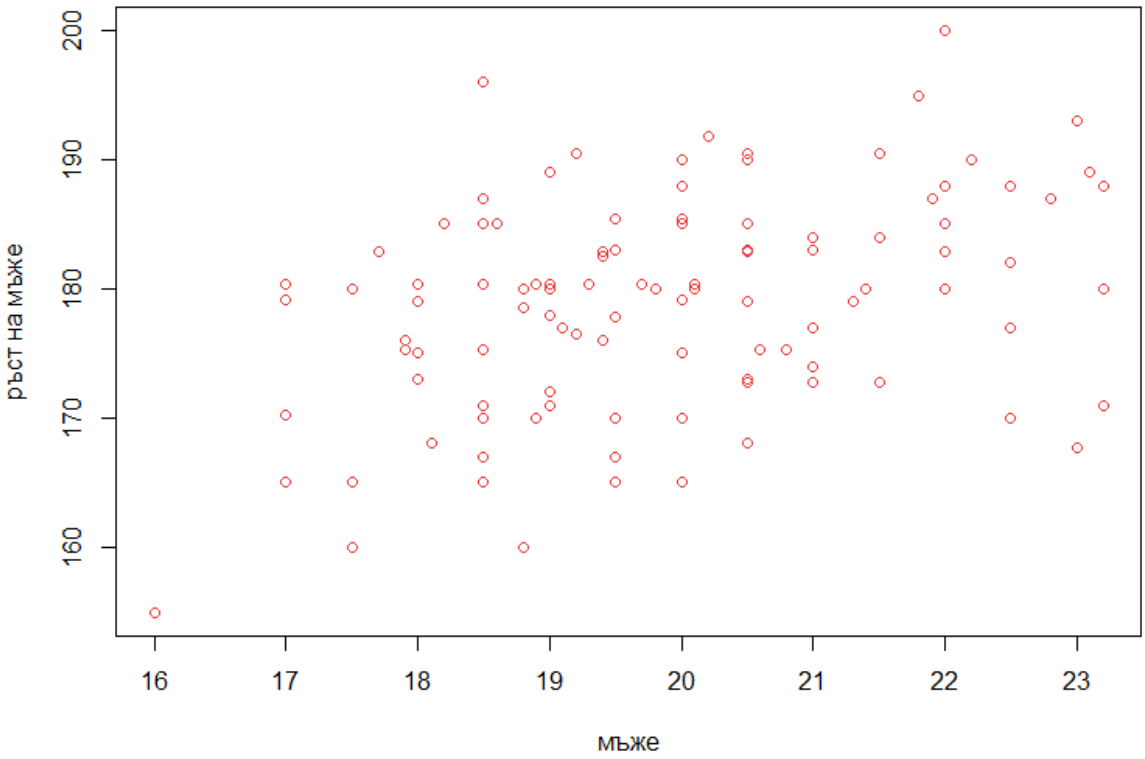
От тестовете следва, че мъжката педя е нормално разпределена, а женската не. Въпреки това можем да заключим, че полът е обясняваща променлива и за педята, тъй като се наблюдава изразена разлика в дължината при жените и мъжете както е и естественото ни очакване.

3.2. числови обясняващи и числови зависими

Ще търсим дали ръста може да обяснява педята.







Разглеждаме ръста и педята съвкупно, както и ако разделим наблюденията по полове. При съвкупното разглеждане можем да видим съвсем слаба позитивна линейна връзка. В останалите две графики, където се разглеждат само жени и мъже не можем да установим някаква явна връзка. Цялостно по-голяма педя имат по-високите хора, но има и някои отклонения.

Ако разгледаме корелацията при съвкупното разглеждане тя е 0.646, което говори за средна връзка. При пресмятането ѝ използваме непараметрична оценка метода на Spearman, защото ръстът е нормално разпределен, но педята не е.

При разглеждане само на жени корелацията е 0.341, което показва, че връзката е много слаба. При нейното пресмятане отново използваме непараметрична оценка метода на Spearman, тъй като педите не бяха с нормално разпределение.

При пресмятане на корелацията при мъже тя е 0.385, което отново означава че връзката е много слаба. При пресмятането ѝ използваме параметрична оценка метода на Pearson, тъй като и двете променливи бяха нормално разпределени.

4. Заключение

Предвид разгледаните визуализации и получените резултати можем да обобщим, че като цяло физическите фактори като ръст и дължина на дланта се влияят от пола на човека, като при жените измерванията са по-малки от тези на мъжете. Установихме, че при ръста имаше нормално разпределение съвкупно и при двата пола поотделно. При педята нямаше нормално разпределение съвкупно и при отделните полове, като разсейването беше по-малко и измерванията бяха струпани около медианата. Намерихме, че съществува средна връзка между двете числови променливи.

Като идеи за подобряване и получаване на по-точни резултати може да се направи проучване с по-голям размер на извадката, както и да се състави по-добра представителна извадка като се включат хора от по-широк етнически и социален статус.