**Проект по Статистика и емпирични методи практикум**

**Иван Чучулски ф.н. 62167**

1. **Въведение**
   1. **Цели на проекта**

Целта на проекта е да се изследват зависимости между измервания на физически показатели на различни хора. Искаме да разберем дали при наличие на данни за едните показатели можем да направим изводи за стойностите на други показатели.

* 1. **Описание на данните**

Ще използваме данните survey от пакета MASS. Те представляват отговори на анкетно проучване, проведено сред 237 студенти в университета в Аделаида, Австралия. Ние ще се разгледаме следните колони на data frame-а :

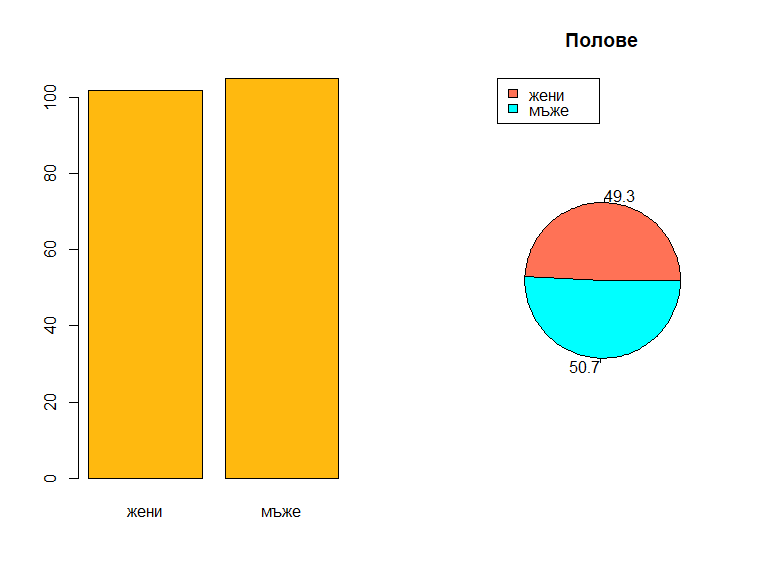
* пол на анкетираните, категорийна номинална променлива
* ръст на анкетираните, числова непрекъсната
* педя, т.е. дължина на дланта на ръката, числова непрекъсната

Първо изследваме как са разпределени данните поотделно, техните локации и разсейвания. 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.

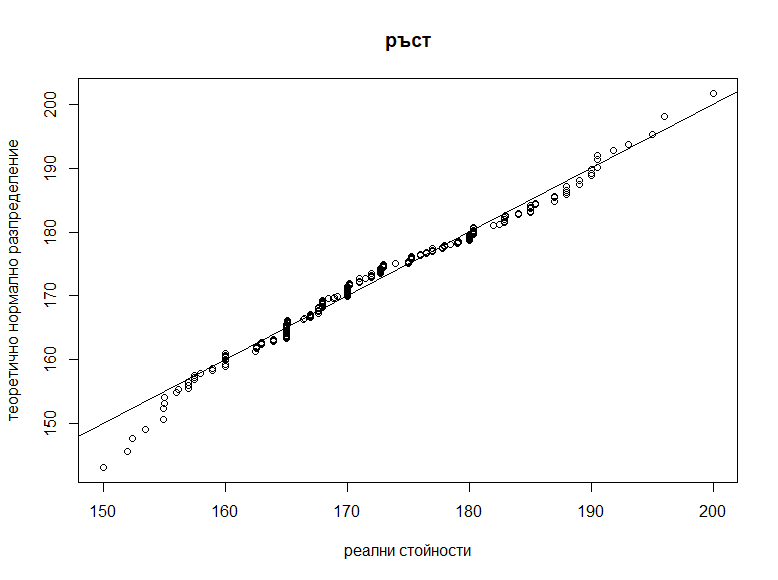
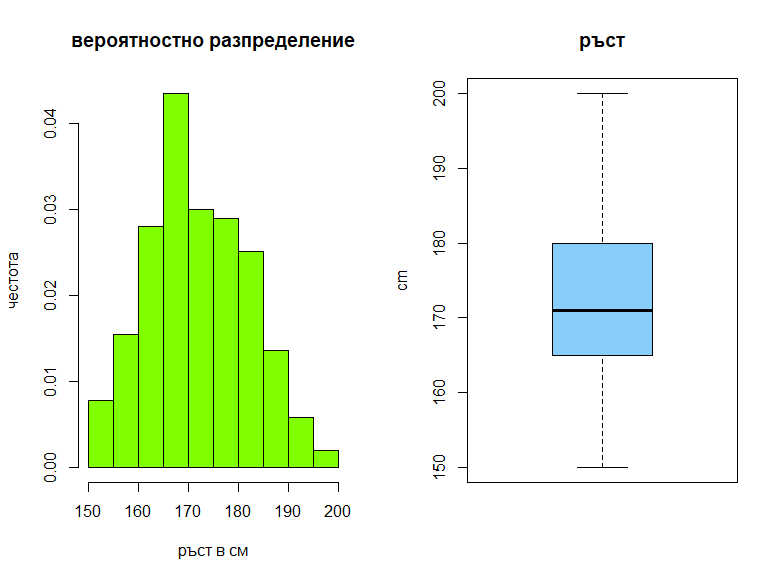
1. **Изследване на променливите поотделно**
   1. **пол, категорийна номинална**

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. Също така ако разглеждаме стойности само върху единия пол и бройката значително различава, това ще доведе до неправилни изводи.

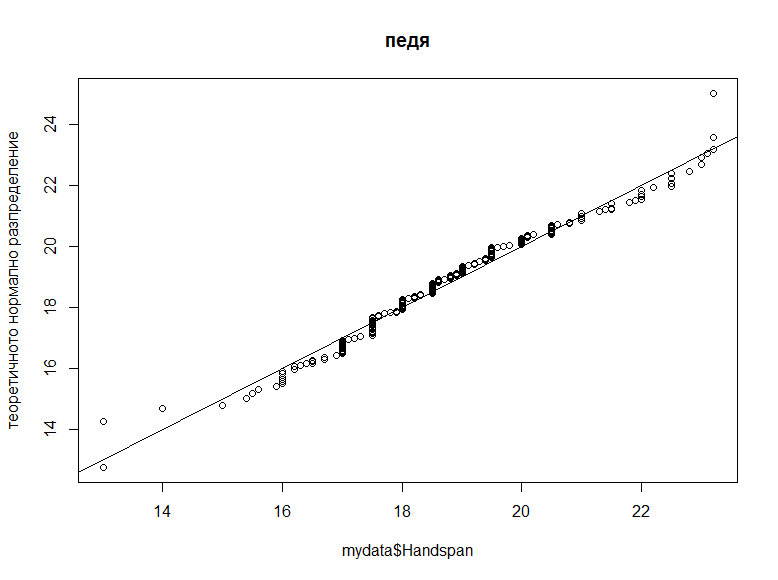
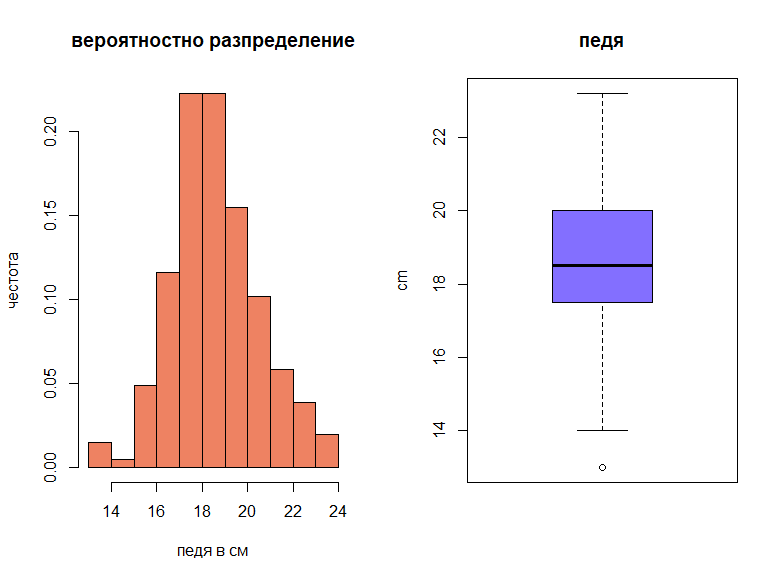
* 1. **ръст, числова непрекъсната**



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 параметрична 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
  1. **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 това разпределението може да не е като нормалното.

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

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

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

1. Изследване на взаимодействия между променливите
   1. категорийни обясняващи и числови зависими
   2. числови обясняващи и числови зависими
2. Заключение