СОФИЙСКИ УНИВЕРСИТЕТ
„СВ. КЛИМЕНТ ОХРИДСКИ“

SOFIA UNIVERSITY
St. KLIMENT OHRIDSKI

Курсов проект по Статистика и Емпирични Методи

Изготвен от:

Манол Джермански, 72003

1. Въвеждане на данните

Данните са взети под формата на csv файл, генериран от анкета. Файлът съдържа информация относно пола, трудов статус, приходи, разходите за хранителни продукти, разходите за хранене навън на над 50 души, попълнили анкетата.

Въвеждаме данните със следната команда:

```
survey_data <- read.csv("D:\\R project\\Food costs.csv")
```

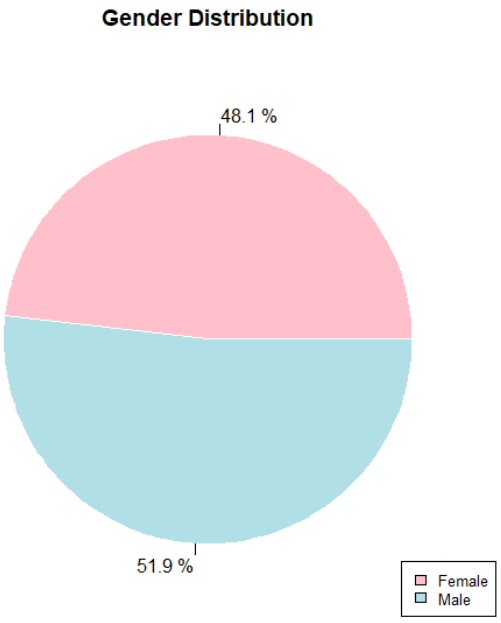
2. Анализ на данните

2.1. Анализ на една променлива

2.1.1. Анализ на категорийни променливи

Ще започнем с анализ на променливата Gender като представим съотношението на участвалите в анкетата чрез Pie chart :

```
# Изчисляване проценти на пола на участниците с точност до 1 знака след десетичната запетая

gender_distribution <- table(survey_data$Gender.)

gender_distribution_percentage <- prop.table(gender_distribution)

percent <- round(gender_distribution_percentage*100,1)

percent <-paste(percent,"%",sep = " ")

# Използваме изчислените проценти за визуализация:

pie(gender_distribution,col = c("pink", "powderblue"),main ="Gender Distribution",labels = percent, border = 0)
```



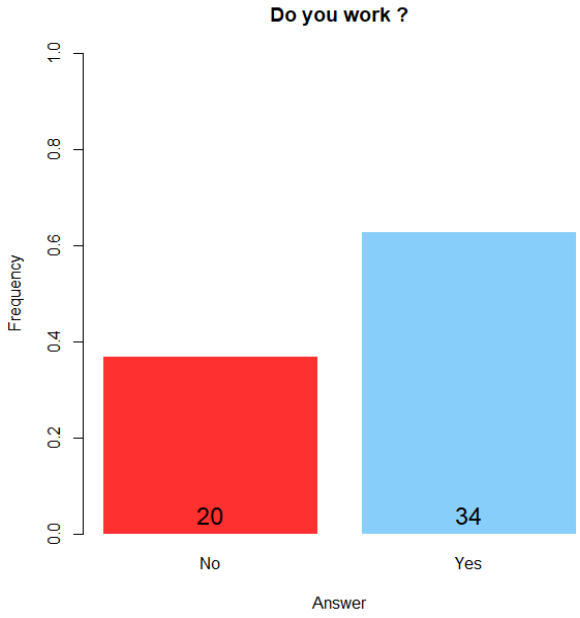
legend("bottomright",c("Female","Male"), fill =c("pink", "powderblue"),cex = 0.8)

Ще направим анализ на променливата трудов статус, като за целта използваме Bar plot:

working_status_percentage <-prop.table(table(survey_data$Do.you.work..))

working_status <-barplot(working_status_percentage, main = "Do you work ?", xlab = "Answer", ylab = "Frequency", col = c("firebrick1","lightskyblue"),border = 0, ylim =c(0,1))

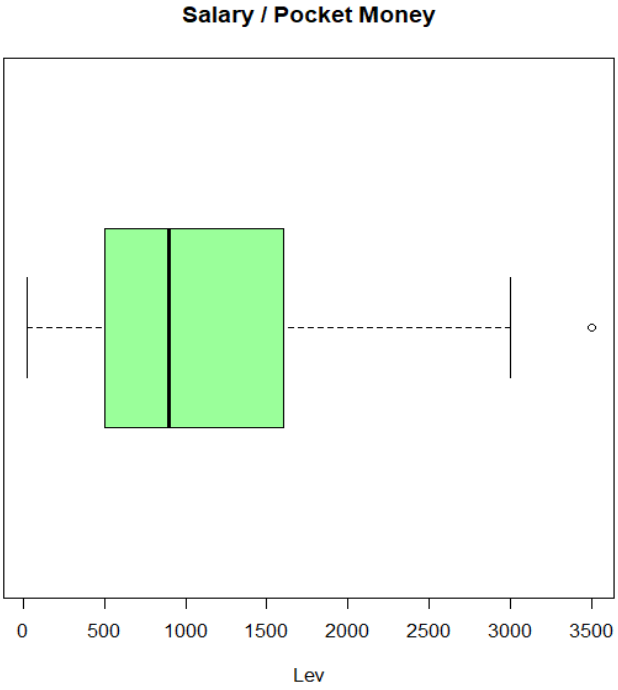
text(working_status, 0, table(survey_data$Do.you.work..), cex = 1.5, pos=3)



2.1.2. Анализ на числови променливи

Ще направим анализ на променливата Salary/Pocket Money с помощта на Box plot:

boxplot(survey_data$Salary.Pocket.Money, las = 1, col = I("palegreen1"), main = "Salary / Pocket Money", horizontal = TRUE, xlab = "Lev")



Както се вижда – минималната стойност е малко над 0 лв., а максималната е 3500 лв. Също така се забелязва, че 1-ви квартил е около 500 лв., а 3-ти е малко над 1500. Медианата е малко под 1000 лв. За по-голяма точност относно разпределението може да използваме:

```
summary(survey_data$Salary.Pocket.Money)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
20	500	900	1107	1600	3500

Ще анализираме стандартното отклонение, median absolute deviation, обхвата и интерквартилния обхват на променливата:

round(sd(survey_data$Salary.Pocket.Money),2) - 800.62

mad(survey_data$Salary.Pocket.Money) - 741.3

range(survey_data$Salary.Pocket.Money) - 20 500

IQR(survey_data$Salary.Pocket.Money) - 1100

Ще проверим дали променливата Salary/Pocket Money е в нормално разпределение.

За целта, ще използваме qqnorm:

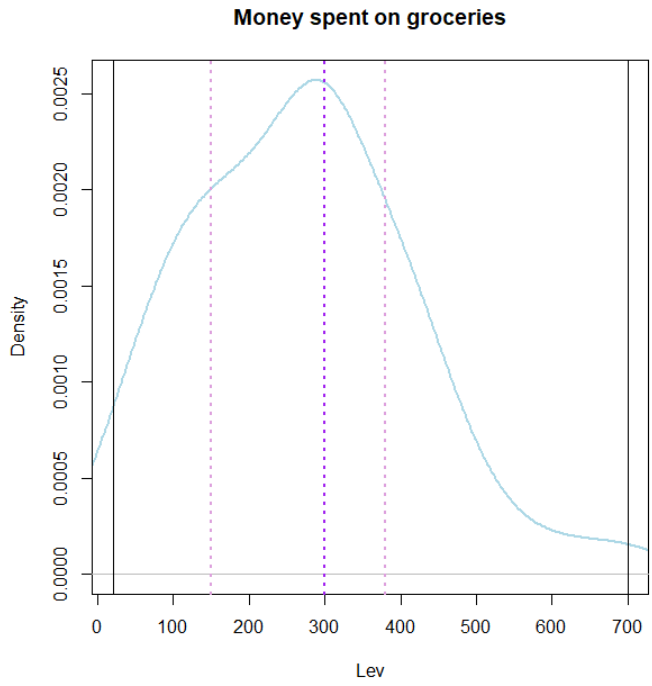
```
qqnorm(survey_data$Salary.Pocket.Money, pch = 1, main = "Check for normal distribution", xlab = "Salary / Pocket Money")
```

```
qqline(survey_data$Salary.Pocket.Money, col = "steelblue", lwd = 2)
```

Забелязва се, че разпределението наподобява експоненциалното разпределение.

Аналогично, ще анализираме променливата Groceries:

```
plot(density(survey_data$Groceries..), lwd = 2, main = "Money spent on groceries", xlab =
"Lev", ylab = "Density",
    col = "lightblue", xlim = range(survey_data$Groceries..))
options(scipen=999)
abline(v = fivenum(survey_data$Groceries..), lwd = c(1.5, rep(2, 3), 1.5),
      col = c("black","plum", "purple", "plum", "black"),lty = c(1, rep(3, 3), 1))
```



Анализирайки графиката, забелязваме, че минималната стойност е малко над 0 лв., максималната е 700 лв., първи квартил е приблизително 150 лв., 3-ти квартил е малко под 400 лв., а медианата е 300 лв.

За по-голяма точност, отново използваме:

```
summary(survey_data$Groceries..)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
20.0	150.0	300.0	268.4	372.5	700.0

Ще анализираме стандартното отклонение, median absolute deviation и интерквартилния обхват на променливата:

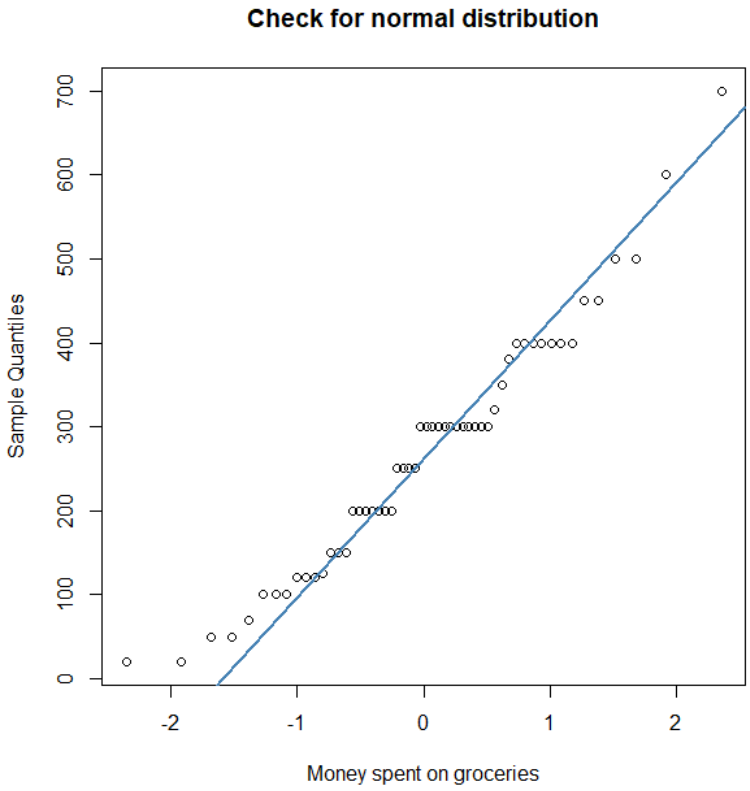
round(sd(survey_data$Groceries..),2) – 146.05

mad(survey_data$Groceries..) - 148.26

IQR(survey_data$Groceries..) – 222.5

Ще проверим дали данните са в нормално разпределение:

```
qqnorm(survey_data$Groceries.., pch = 1, main = "Check for normal distribution", xlab =
"Money spent on groceries")

qqline(survey_data$Groceries.., col = "steelblue", lwd = 2)
```



Тъй като графиката поражда съмнение дали данните са в нормално разпределение, ще използваме Shapiro-Wilk normality test:

```
shapiro.test(survey_data$Groceries..)
```

```
        Shapiro-Wilk normality test

data:  survey_data$Groceries..

W = 0.96506, p-value = 0.1161
```

Тъй като p-value > 0.05, приемаме че данните са в нормално разпределение

Остава да анализираме последната ни числова променлива – Eating out.

```
summary(survey_data$Eating.out..)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   20.0    50.0   100.0   146.2   200.0  1000.0
```

Ще анализираме стандартното отклонение, median absolute deviation и интерквартилния обхват на променливата

round(sd(survey_data$Eating.out..),2) – 154.18

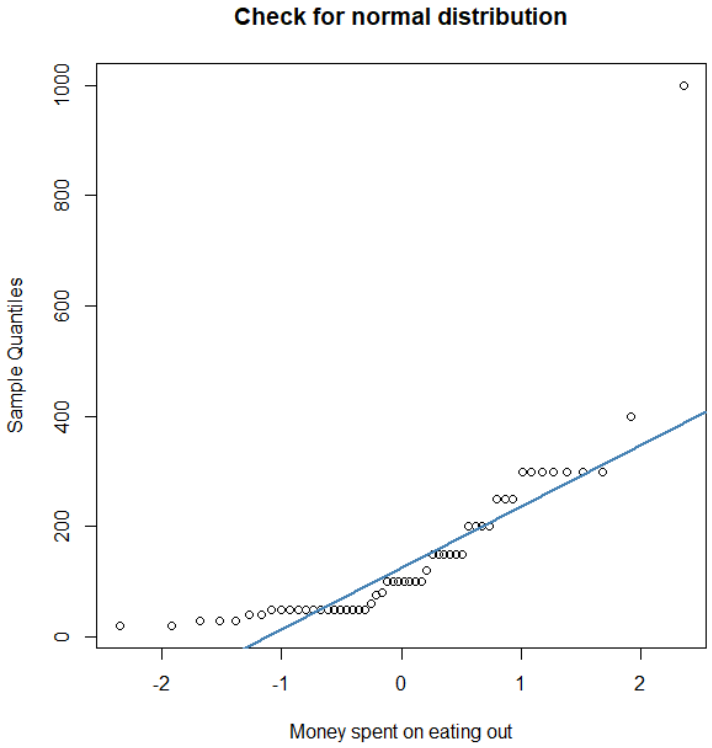
mad(survey_data$Eating.out..) – 74.13

IQR(survey_data$Eating.out..) – 150

Ще проверим дали променливата е в нормално разпределение:

qqnorm(survey_data$Eating.out.., pch = 1, main = "Check for normal distribution", xlab = "Money spent on eating out")

qqline(survey_data$Eating.out.., col = "steelblue", lwd = 2)



По графиката се забелязва, че променливата не е в нормално разпределение. За потвърждение ще направим Shapiro-Wilk normality test.

shapiro.test(survey_data$Eating.out..)

Shapiro-Wilk normality test

data: survey_data$Eating.out..

W = 0.66263, p-value = 0.0000000007152

Тъй като p-value < 0.05 , данните действително не са в нормално разпределение.

2.2. Многомерен анализ

2.2.1. Анализ на категорийна спрямо числова променлива

Ще направим анализ на променливата Salary/Pocket Money спрямо променливата Gender.

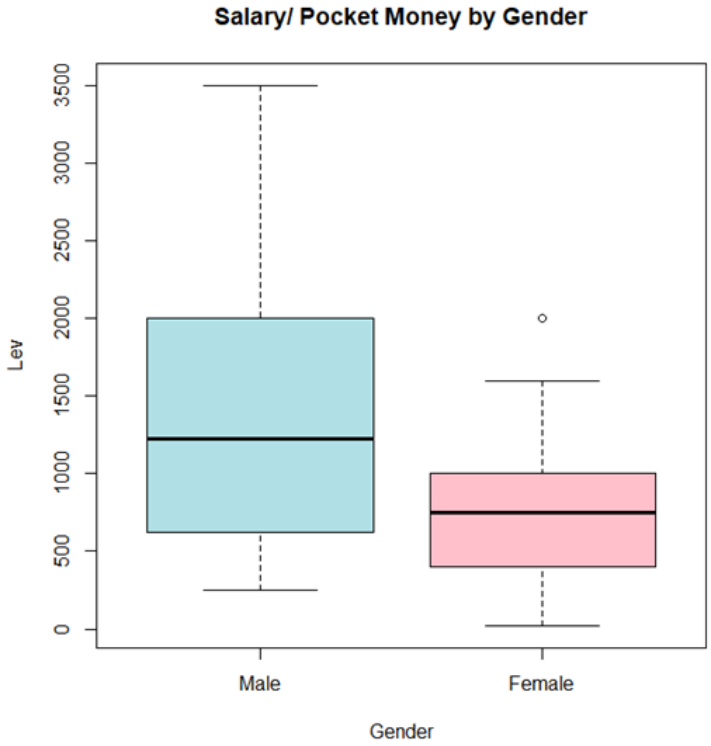
Нулевата хипотеза е, че няма разлика в приходите между мъжете и жените, а алтернативната е, че има разлика в приходите.

Тъй като данните не са в нормално разпределение, ще ползваме Wilcox тест:

```
wilcox.test( survey_data_frame$Salary.Pocket.Money[survey_data_frame$Gender. == "Male"],
 survey_data_frame$Salary.Pocket.Money[survey_data_frame$Gender. == "Female"], paired = F,exact = T,conf.int = T)
```

```
        Wilcoxon rank sum test with continuity correction

data:  survey_data_frame$Salary.Pocket.Money[survey_data_frame$Gender. == "Male"] and survey_data_frame$Salary.Pocket.Money[survey_data_frame$Gender. == "Female"]
W = 502.5, p-value = 0.01667
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 99.99999  999.99998
sample estimates:
difference in location
              488.7192
```

От резултата на теста можем да отречем нулевата хипотеза, тоест има разлика в средните приходите на двата пола.



Ще направим анализ на променливата Groceries спрямо променливата Gender. Тъй като Groceries е с нормално разпределение, ще ползваме t-test".

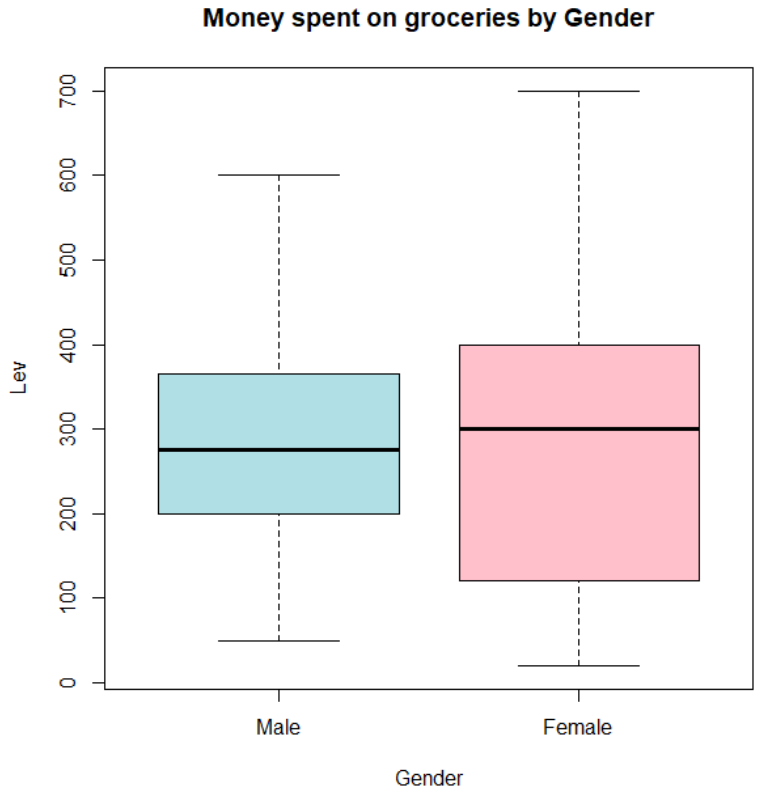
Нулевата хипотеза е, че двата пола харчат средно еднакво за хранителни стоки, а алтернативната е, че харчат различно.

```
t.test(survey_data_frame$Groceries..[survey_data_frame$Gender. == "Male"],
survey_data_frame$Groceries..[survey_data_frame$Gender. == "Female"], paired = F,exact =
T,conf.int = T)
```

Welch Two Sample t-test

data: survey_data_frame$Groceries..[survey_data_frame$Gender. == "Male"] and survey_data_frame$Groceries..[survey_data_frame$Gender. == "Female"]
t = 0.38489, df = 49.953, p-value = 0.702
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-65.42480 96.44128
sample estimates:
mean of x mean of y
275.8929 260.3846

Тъй като p-value > 0.05, можем да приемам нулевата хипотеза, тоест двата пола средно харчат еднакво за хранителни стоки.

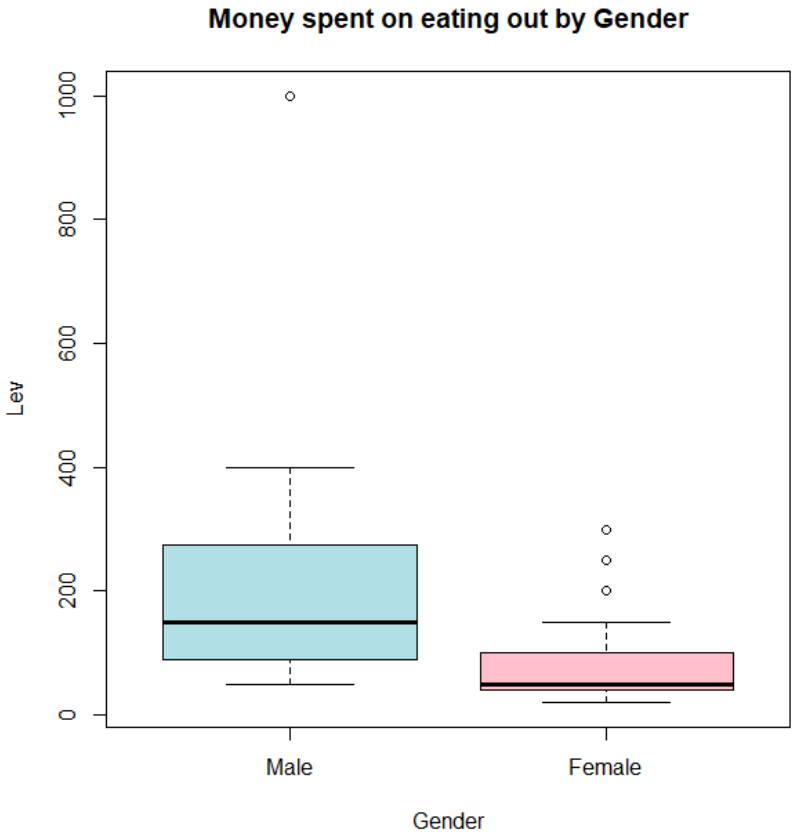


Ще анализираме променливата Eating out спрямо променливата Gender. Отново, ще се позовем на Wilcox test, тъй като Eating out не е нормално разпределена:

```
wilcox.test( survey_data_frame$Eating.out..[survey_data_frame$Gender. == "Male"],
survey_data_frame$Eating.out..[survey_data_frame$Gender. == "Female"],
 paired = F,exact = T,conf.int = T)
```

```
        Wilcoxon rank sum test with continuity correction

data:  survey_data_frame$Eating.out..[survey_data_frame$Gender. == "Male"] and
survey_data_frame$Eating.out..[survey_data_frame$Gender. == "Female"]
W = 561, p-value = 0.0005756
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 29.99996 139.99999
sample estimates:
difference in location
              70.00008
```

Тъй като p-value е значително по-малко от 0.05, можем да отречем нулевата хипотеза и да кажем, че има значителна разлика в средната сума, която двата пола отделят за хранене навън.



Ще направим анализ на променливата Salary/Pocket Money спрямо променливата Working Status чрез Wilcox test.

Нулевата хипотеза е, че няма разлика в средните приходи спрямо работния статус, а алтернативната е, че има разлика в средните приходи спрямо работния статус.

```
wilcox.test(
survey_data_frame$Salary.Pocket.Money[survey_data_frame$Do.you.work.. == "Yes"],
survey_data_frame$Salary.Pocket.Money[survey_data_frame$Do.you.work.. == "No"],
paired = F,exact = T,conf.int = T)
```

Wilcoxon rank sum test with continuity correction

data: survey_data_frame$Salary.Pocket.Money[survey_data_frame$Do.you.work.. == "Yes"] and survey_data_frame$Salary.Pocket.Money[survey_data_frame$Do.you.work.. == "No"]

W = 644, p-value = 5.113e-08

alternative hypothesis: true location shift is not equal to 0

95 percent confidence interval:

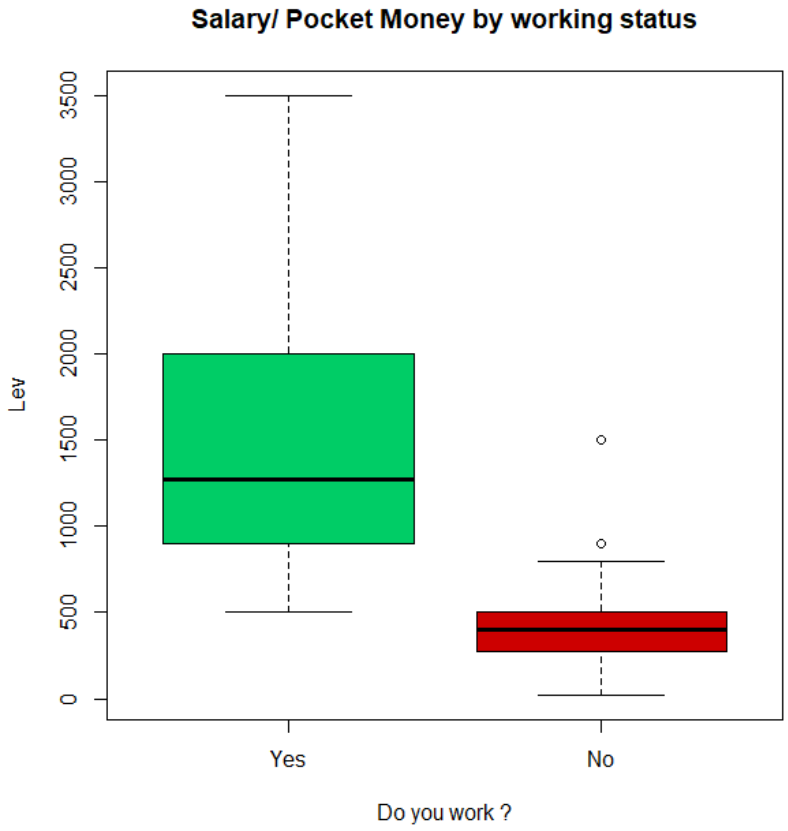
600 1350

sample estimates:

difference in location

900

От p-value може отречем нулевата хипотеза, т.е. има разлика в приходите спрямо работния статус.



Ще анализираме променливата Groceries спрямо работния статус чрез t-test

```
t.test(survey_data_frame$Groceries..[survey_data_frame$Do.you.work.. == "Yes"], survey_data_frame$Groceries..[survey_data_frame$Do.you.work.. == "No"], paired = F,exact = T,conf.int = T)
```

Welch Two Sample t-test

data: survey_data_frame$Groceries..[survey_data_frame$Do.you.work.. == "Yes"] and survey_data_frame$Groceries..[survey_data_frame$Do.you.work.. == "No"]

t = 5.0623, df = 42.056, p-value = 8.669e-06

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

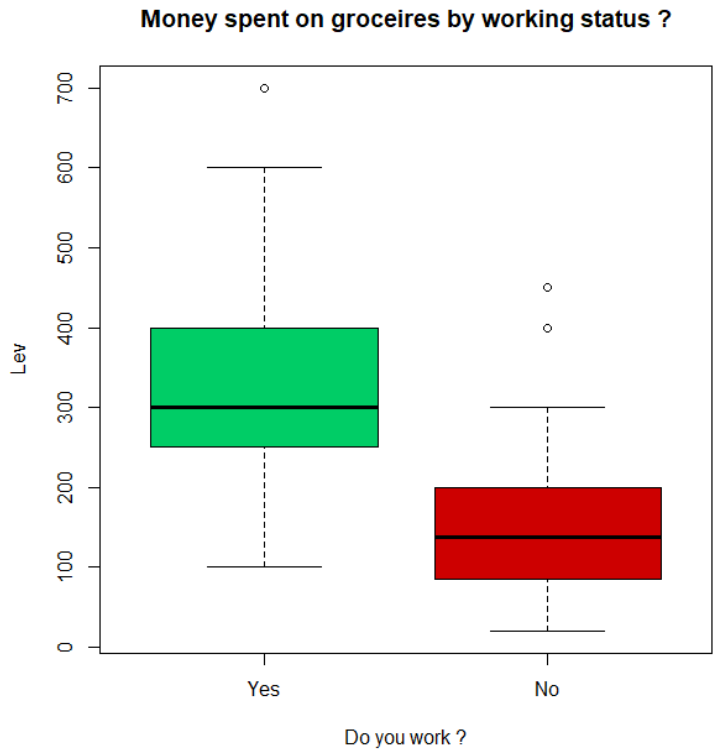
102.3645 238.0766

sample estimates:

mean of x mean of y

331.4706 161.2500

Тъй като p-value е доста по-малка оп 0.05, можем да отречем нулевата хипотеза – има разлика в средната стойност, отделена за хранителни продукти спрямо работния статус.



Остава да анализираме променливата Eating out спрямо работния статус. Това ще стане чрез Wilcox test.

Нулевата хипотеза е, че няма разлика в средната сума отделена за хранене навън спрямо работния статус, а алтернативната е, че има.

```
wilcox.test(survey_data_frame$Eating.out..[survey_data_frame$Do.you.work.. == "Yes"], survey_data_frame$Eating.out..[survey_data_frame$Do.you.work.. == "No"], paired = F,exact = T,conf.int = T)
```

Wilcoxon rank sum test with continuity correction

data: survey_data_frame$Eating.out..[survey_data_frame$Do.you.work.. == "Yes"] and survey_data_frame$Eating.out..[survey_data_frame$Do.you.work.. == "No"]

W = 520.5, p-value = 0.001102

alternative hypothesis: true location shift is not equal to 0

95 percent confidence interval:

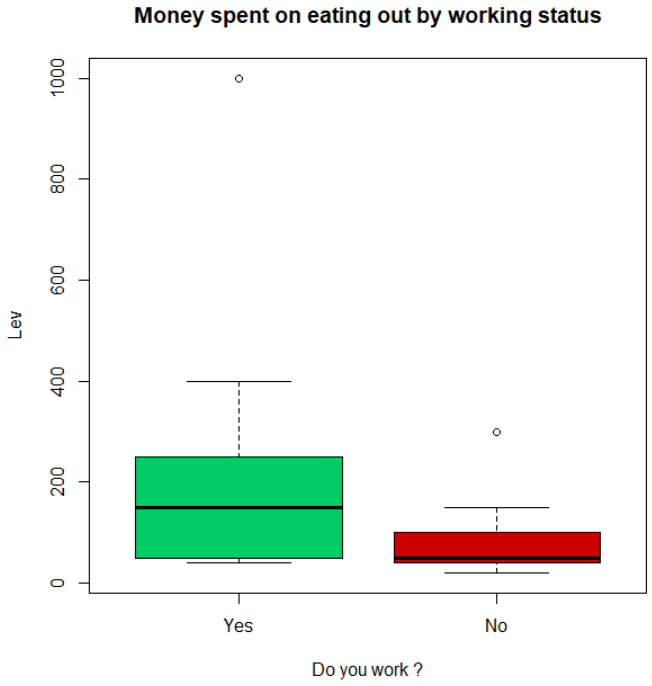
24.99999 149.99993

sample estimates:

difference in location

79.99995

От p-value можем да заключим, че има разлика в средната сума отделена за хранене навън спрямо работния статус.



2.2.2. Анализ на числова променлива спрямо друга

Ще направим анализ на променливата Salary/Pocket Money спрямо променливата Eating out. Тъй като променливата Eating out не е нормално разпределена , ще се насочим към корелационен анализ на Spearman:

Нулевата хипотеза е, че няма корелация между променливите, а алтернативната е, че има.

```
cor.test(survey_data_frame$Salary.Pocket.Money,survey_data_frame$Eating.out..,method = "spearman")
```

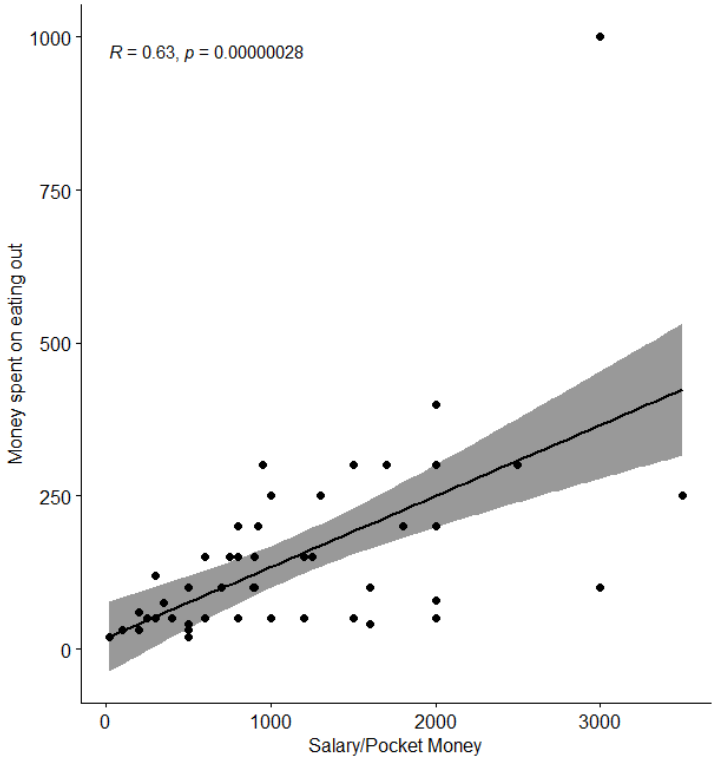
Spearman's rank correlation rho

data: survey_data_frame$Salary.Pocket.Money and survey_data_frame$Eating.out..

S = 9631.6, p-value = 0.0000002827

alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.6328708

Тъй като p-value е по-малка от 0.05, приемаме, че между двете променливи има корелация. Още повече, коефициентът на корелация rho е 0.6328708, което говори средна корелация.



Ще направим анализ на променливата Groceries спрямо променливата Eating out. Тъй като Eating out не е нормално зависима, ще направим корелационен анализ на Spearman.

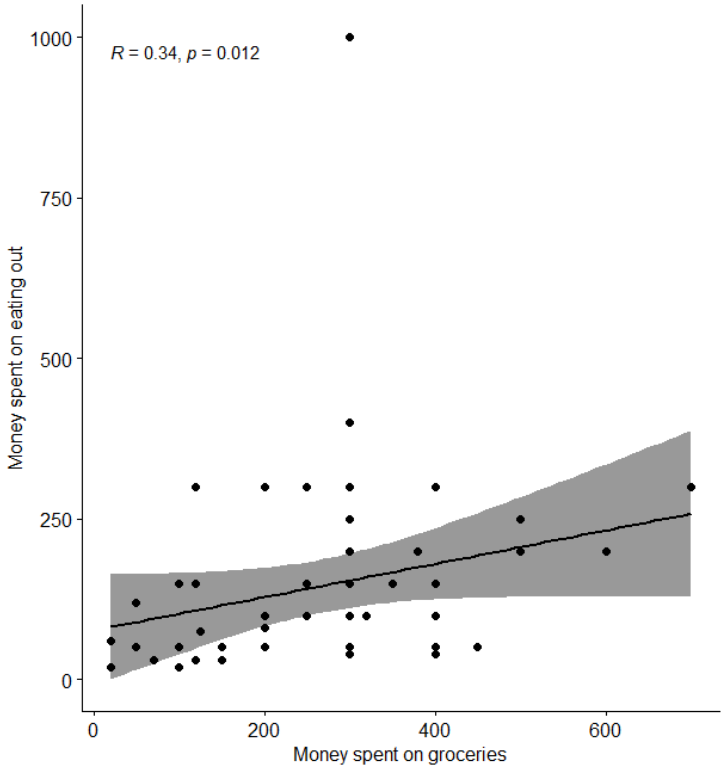
Нулевата хипотеза е, че няма корелация между променливите, а алтернативната гласи обратното.

```
cor.test (survey_data_frame$Groceries.., survey_data_frame$Eating.out.., method =
"spearman")
```

Spearman's rank correlation rho

data: survey_data_frame$Groceries.. and survey_data_frame$Eating.out..
S = 17281, p-value = 0.01155
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.3413023

Тъй като p-value е под 0.05, можем да кажем, че данните са в корелация, но коефициентът на корелация rho е 0.3413023, което говори за много слаба корелация.

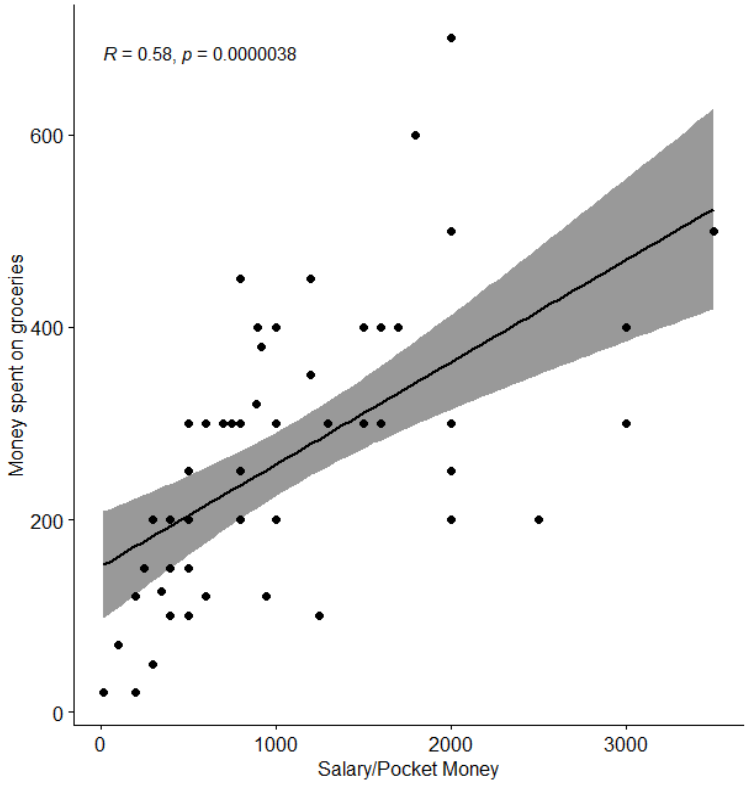


За финал ще направим корелационен анализ на променливите Salary/Pocket Money и Groceries:

```
cor.test(survey_data_frame$Salary.Pocket.Money,survey_data_frame$Groceries..,method = "pearson")
```

```
        Pearson's product-moment correlation

data:  survey_data_frame$Salary.Pocket.Money and survey_data_frame$Groceries..
t = 5.1691, df = 52, p-value = 0.000003801
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3730381 0.7356113
sample estimates:
      cor
0.5826052
```

Забелязваме, че има средна корелация между променливите.



3. Изводи

Проектът представя анализ на приходите, разходите за хранителни стоки и разходите за хранене навън на 54 анкетирани души. От тях 51.9 % са мъже и 49.1% жени. На въпроса дали работят 20 човека са отговорили с “Не”, а 34 с “Да”.

При изследване на приходите, забелязваме че най-ниската сума е 20 лв. , а максималната е 3500 лв. на месец. Медианата е 900 лв. , а средната заплата е 1107 лв.

Данните за разходите за хранителни стоки сочат следното:

Най-ниската отделена сума е 20 лв., а най-голямата – 700 лв. Медианата е 300 лв., а средната стойност – 268.4 лв.

Разходите за хранене навън са следните: Минималната сума е 20 лв., максималната 1000 лв. Медианата е 100 лв., а средната стойност 146.2 лв.

При анализ на приходите на мъжете и жените се установява, че за мъжете: минималната сума е 250 лв., медианата – 1225 лв., средната стойност -1520 лв. и максималната е 3500 лв. От друга страна, за жените – минималната сума е 20 лв, медианата е 750 лв. , средната стойност 754 лв., а максималната – 1600 лв.

Относно разходите за хранителни стоки за двата пола – минималната стойност за мъжете е 50 лв., медианата 275 лв., следната стойност – 298 лв., а максималната е 600 лв. За жените минималната е 20 лв., медианата – 300 лв., средната стойност е 308 лв., а максималната е 700 лв.

За хранене навън, минималната сума за мъжете е 40 лв., медианата – 150 лв., средната стойност – 178 лв., а максималната 400 лв. За жените минималната е 20 лв., медианата – 50 лв., средата стойност 78 лв., а максималната 150 лв.

Спрямо трудовия статус, за работещите приходите са следните: минимална стойност – 500 лв., медиана – 1275 лв., средна стойност – 1635 лв. и максимална стойност 3500 лв. За неработещите, минималната стойност е 20 лв., медианата е 400 лв., средната стойност - 399 лв. и максималната стойност е 800 лв.

При анализ на разходите за хранителни стоки: Минималната стойност за работещите е 100 лв., медианата – 300 лв., средната стойност 330 лв., а максималната е 600 лв. От друга страна за неработещите – минималната стойност е 20 лв., медианата – 137.5 лв., средната стойност е 148.5 лв., а максималната стойност е 300 лв.

За хранене навън, минималната сума за работещите е 40 лв., медианата – 150 лв., средната стойност – 178 лв., а максималната е 400 лв. За неработещите – минималната стойност е 20 лв., медианата – 50 лв., средната стойност – 72 лв., а максималната стойност е 150 лв.

При изследване на връзките между приходите и разходите за хранителни стоки и хранене навън се забелязва средна положителна корелация.