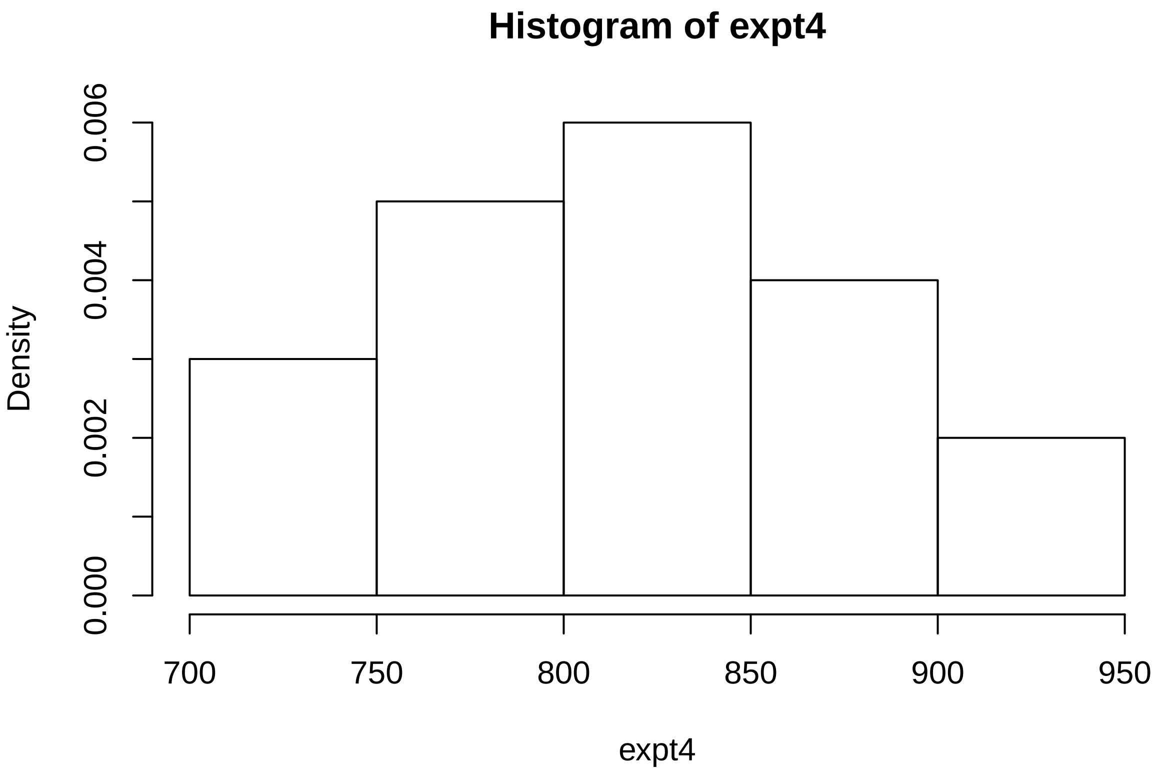
Домашно 2 по ВС (практикум) ФН: 45342

Ivo Stratev

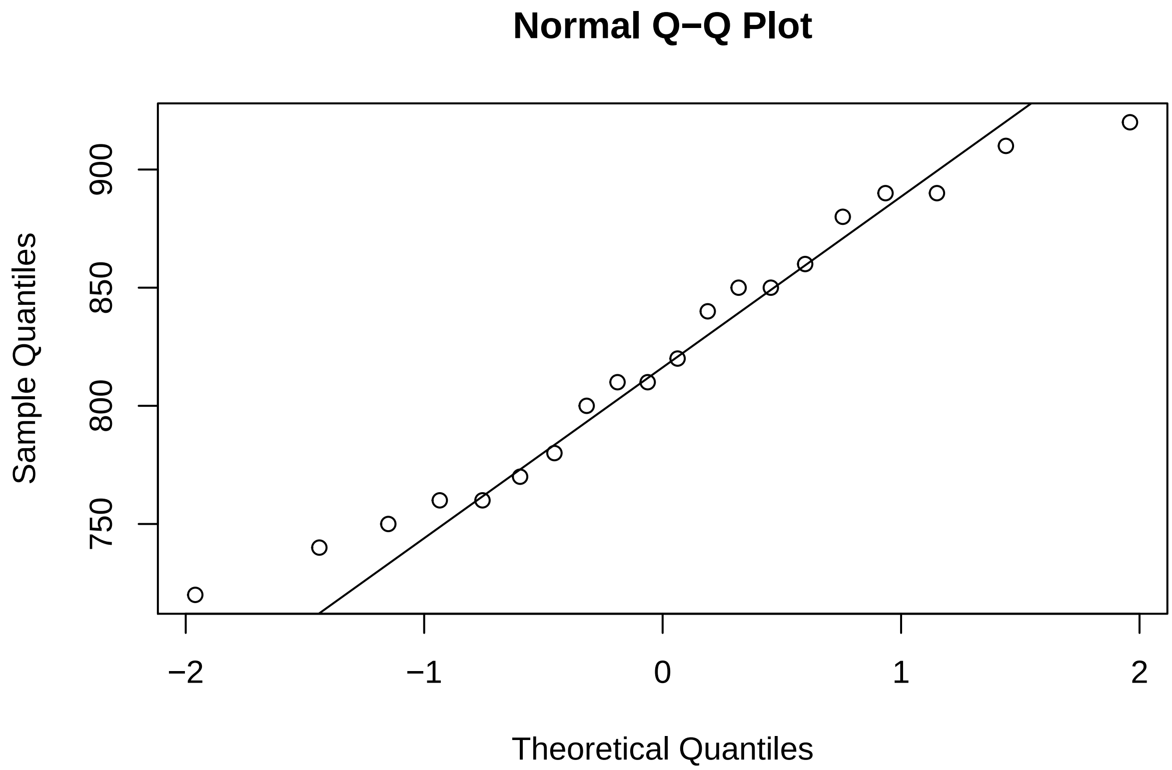
Задача 1.

```
library('MASS');
data("morley");
expt4 = morley[morley$Expt == 4, 3];
hist(expt4, prob=T)
```



Хистограмата изглежда като на нормално разпределение. Нека все пак сравним графично квантилите на разпределението на данните срещу това на нормално разпределение.

```
qqnorm(expt4)
qqline(expt4)
```



Сравнеието на квантилите потвърждава хипотезата, че данните са нормално разпределени. Следователно можем да намерим доверителен интервал с 97% точност за това какво е очакването на данните.

```
t.test(expt4, conf.level = 0.97)
```

```
##
##  One Sample t-test
##
## data:  expt4
## t = 61.114, df = 19, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 97 percent confidence interval:
##  789.008 851.992
## sample estimates:
## mean of x
##     820.5
```

Както можем да видим с 97% точност очакването на данните се намира в интервала [789, 852] като средната стойност е 820.5.

Задача 2.

Производителя твърди, че в плик с бонбони всички цветове с изключение на синия се срещат с еднаква вероятност, а той с два пъти по-голяма от останалите. Даден ни е плик с 83 сини, 32 червени, 42 оранжеви и 48 жълти искаме да проверим дали производителя казва истината. Тоест да проверим дали е възможно пакета да бъде извадка с твърдените пропорции. За тази цел ще направим Хи-квадрат тест. Първо пресмятаме вероятността за всеки цвят: 2p + p + p + p = 1. Следователно вероятността за син бонбон е 2/5, а за всеки друг цвят 1/5.

```
chisq.test(c(83, 35, 42, 48), p = c(2, 1, 1, 1) / 5)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(83, 35, 42, 48)
## X-squared = 2.0361, df = 3, p-value = 0.565
```

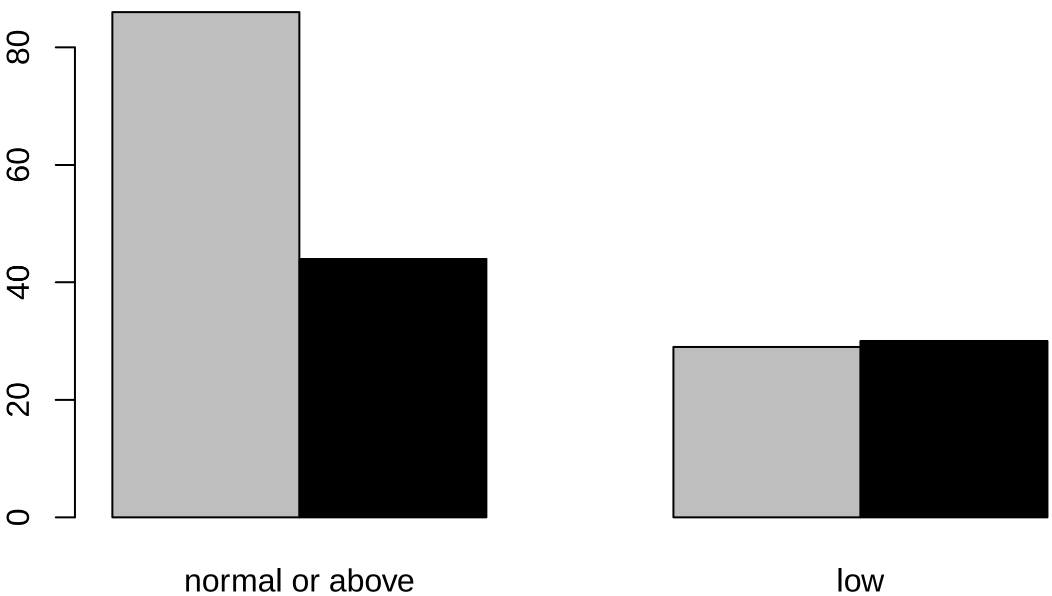
Следователно вероятността производителят да казва истината е 56.5%. Тоест можем да приемем, че той не лъже.

Задача 3.

```
library("knitr");
data("birthwt");
t = table(birthwt$smoke, birthwt$low);
colnames(t) = c("normal or above", "low");
row.names(t) = c("don't smoke", "smoke");
kable(t)
```

	normal or above	low
don't smoke	86	29
smoke	44	30

```
barplot(t, beside = T, col= c("grey", "black"))
```



Графично забелязваме, че има връзка между теглото и тютюнопушенето. Все пак ще проведем Хи-квадрат тест. Нулевата хипотеза е, че липсва зависимост между теглото и тютюнопушенето. Алтернативната, че има зависимост.

```
chisq.test(as.data.frame.matrix(t))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  as.data.frame.matrix(t)
```

X-squared = 4.2359, df = 1, p-value = 0.03958

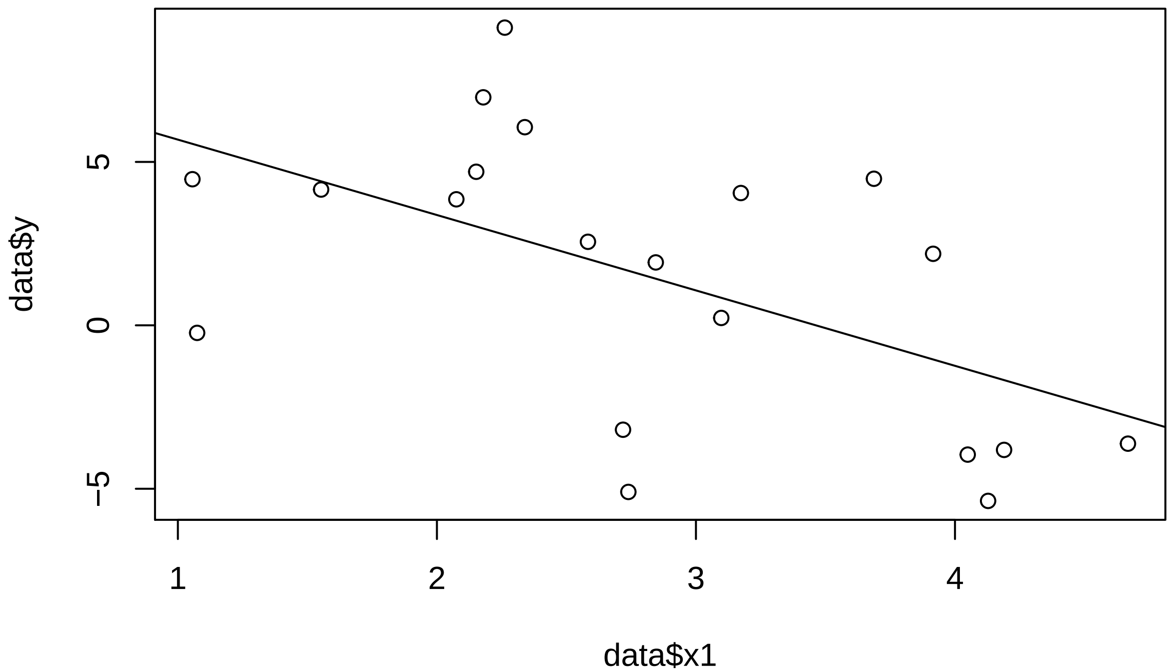
Нулевата хипотеза е подкрепена с вероятност 4%, за това я отхвърляме. Следователно има зависимост между теглото и тютюнопушенето.

Задача 4.

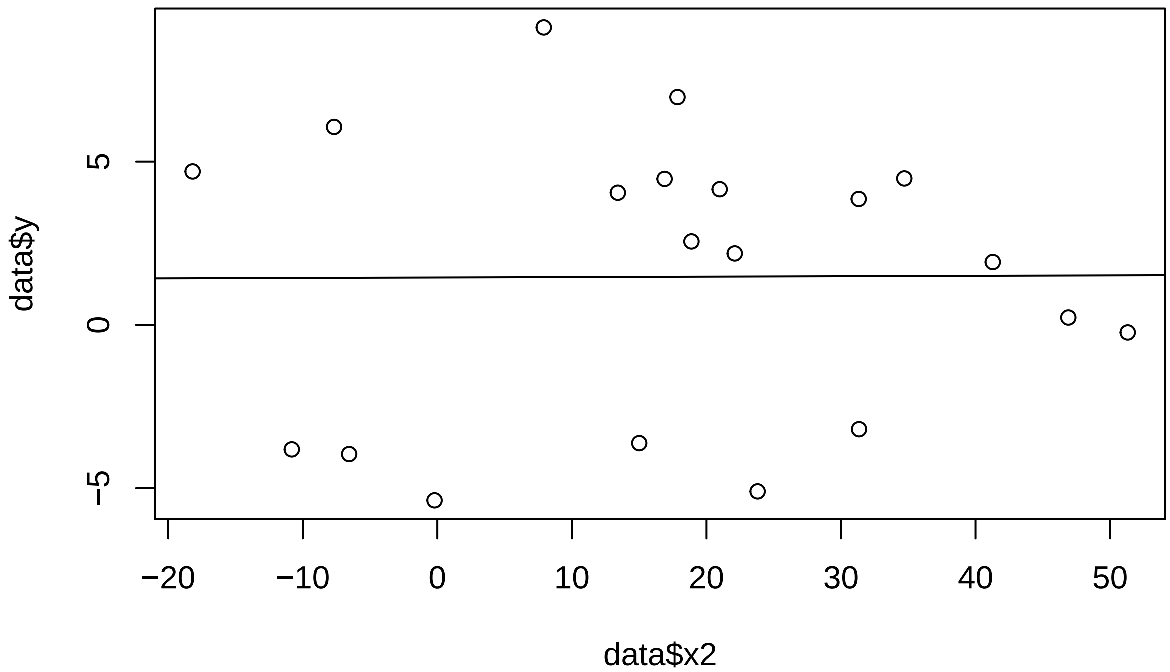
```
data = read.csv("DomR2.csv", header = T);
```

Изследваме как всяка от колоните: x1, x2, x3 влиае на колоната y.

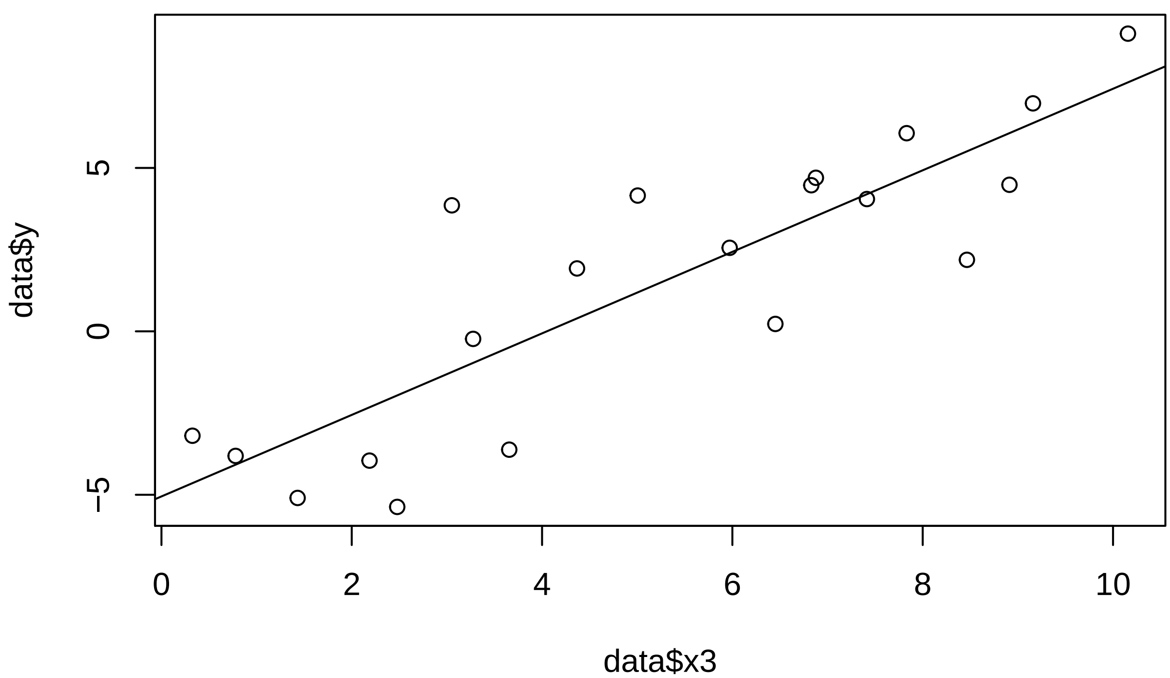
```
plot(data$x1, data$y);
abline(lm(data$y ~ data$x1));
```




```
plot(data$x2, data$y);
abline(lm(data$y ~ data$x2));
```




```
plot(data$x3, data$y);
abline(lm(data$y ~ data$x3));
```



Забелязваме, че между x3 и y твърдо има линейна зависимост. Между x1 и y има доста слаба линейна зависимост и между x2 и y твърдо няма линейна зависимост.

Все пак пресмятаме коефициентите на корелация.

```
cor(data$x1, data$y)

## [1] -0.550594

cor(data$x2, data$y)
```

```
## [1] 0.005742398
```

```
cor(data$x3, data$y)
```

```
## [1] 0.8571049
```

Стройм линеен модел ползвайки само x1 и x3.

```
s = summary(lm(data$y ~ data$x1 + data$x3));
s$r.squared
```

```
## [1] 0.871811
```

```
s$adj.r.squared
```

```
## [1] 0.85673
```

Проверяваме дали ако премахнем свободния член от модела ще получим по-добър R^2 резултат.

```
s = summary(lm(data$y ~ data$x1 + data$x3 - 1));
s$r.squared
```

```
## [1] 0.8856211
```

```
s$adj.r.squared
```

```
## [1] 0.8729124
```

И наистина има леко подобрение в R^2 резултата, така че ще използваме този модел. Тоест в модела участват променливите x1 и x3 и няма свободен член (получената права минава през центъра на координатната система).

```
s
```

```
##
## Call:
## lm(formula = data$y ~ data$x1 + data$x3 - 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4058 -1.3086  0.1345  0.9056  3.6719
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## data$x1  -1.5731     0.1930  -8.152 1.87e-07 ***
## data$x3   1.1300     0.0967  11.685 7.73e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.599 on 18 degrees of freedom
## Multiple R-squared:  0.8856, Adjusted R-squared:  0.8729
## F-statistic: 69.69 on 2 and 18 DF,  p-value: 3.351e-09
```

Очевидно е, че коефицията пред x1 не е -1, но все пак ще направим статистически тест за да се уверим в това :) .

```
beta1 = s$coefficients[1, 1];
se = s$coefficients[1, 2];
a = -1;
df = s$df[2];
t = (beta1 - a) / se;
2 * pt(t, df, lower.tail = t < 0)
```

[1] 0.008207128

Понеже вероятността е под 1% то директно отхвърляме хипотезата, че коефициента пред x1 е -1.