Домашнее задание по стохастическому анализу к 26.10.2023

Выполнил: Милюшков Георгий

Отключаем предупреждения для более красивого вывода.

Подключаем нужные библиотеки.

```
library(readr)
library(ggmap)
library(ggplot2)
library(tidyr)
library(dplyr)
library(kableExtra)
library(data.table)
library(corrplot)
library(nortest)
```

Читаем данные.

```
df <- read.csv(
  file = "/Users/georgymilyuskhov/Documents/Учеба/4 курс /R/data.csv",stringsAsFactors=FALSE,fileEncoding="latin1
"
)
```

Функция для более красивого вывода.

```
print_df <- function(df)
{
  df |>
    kable(format = "html") |>
    kable_styling() |>
    kableExtra::scroll_box(width = "100%", height = "100%")
}
```

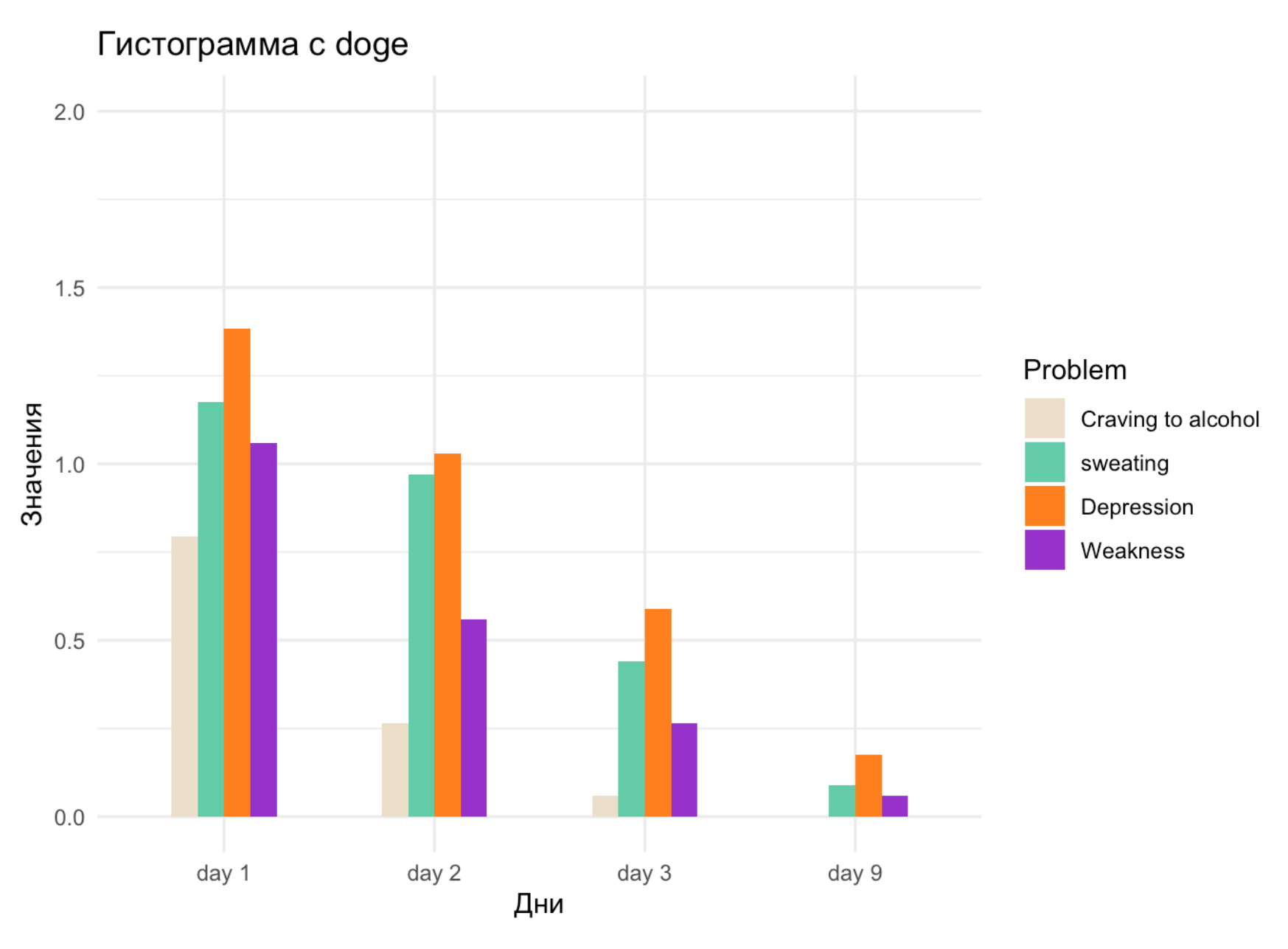
Печатаем первые 10 строк наших данных.

```
df |> head(10) |> print_df()
```

X	X.1	depressed.mood.1	anxiety.1	suspiciousness.1	irritability.1	craving.to.alcohol.1	weakness.1	insomia.1	headache.1
1	1	1	1	0	1	1	1	1	0
2	2	1	1	0	0	1	1	2	1
3	3	1	1	0	0	0	2	1	0
4	4	2	2	0	0	0	2	0	0
5	5	1	1	0	0	2	2	1	0
6	6	1	1	0	1	1	2	0	0
7	7	1	1	0	1	1	2	2	1
8	8	1	1	0	1	0	1	2	1
9	9	1	1	0	0	2	1	0	0
10	10	1	1	0	1	1	1	2	1

Ради интереса смотрим, что наши значения корректны и соответствуют логике, что самочувствие пациентов улучшается. Было выбрано “слабость”, “плотность”, “тяга к алкоголю” и “депрессивность”. Наблюдаем, что тяга к алкоголю полностью исчезает после третьего дня. Слабость быстро сходит на нет. Также видно, что депрессия не так быстро лечится за один день, но к девятому дню ее тоже почти нет.

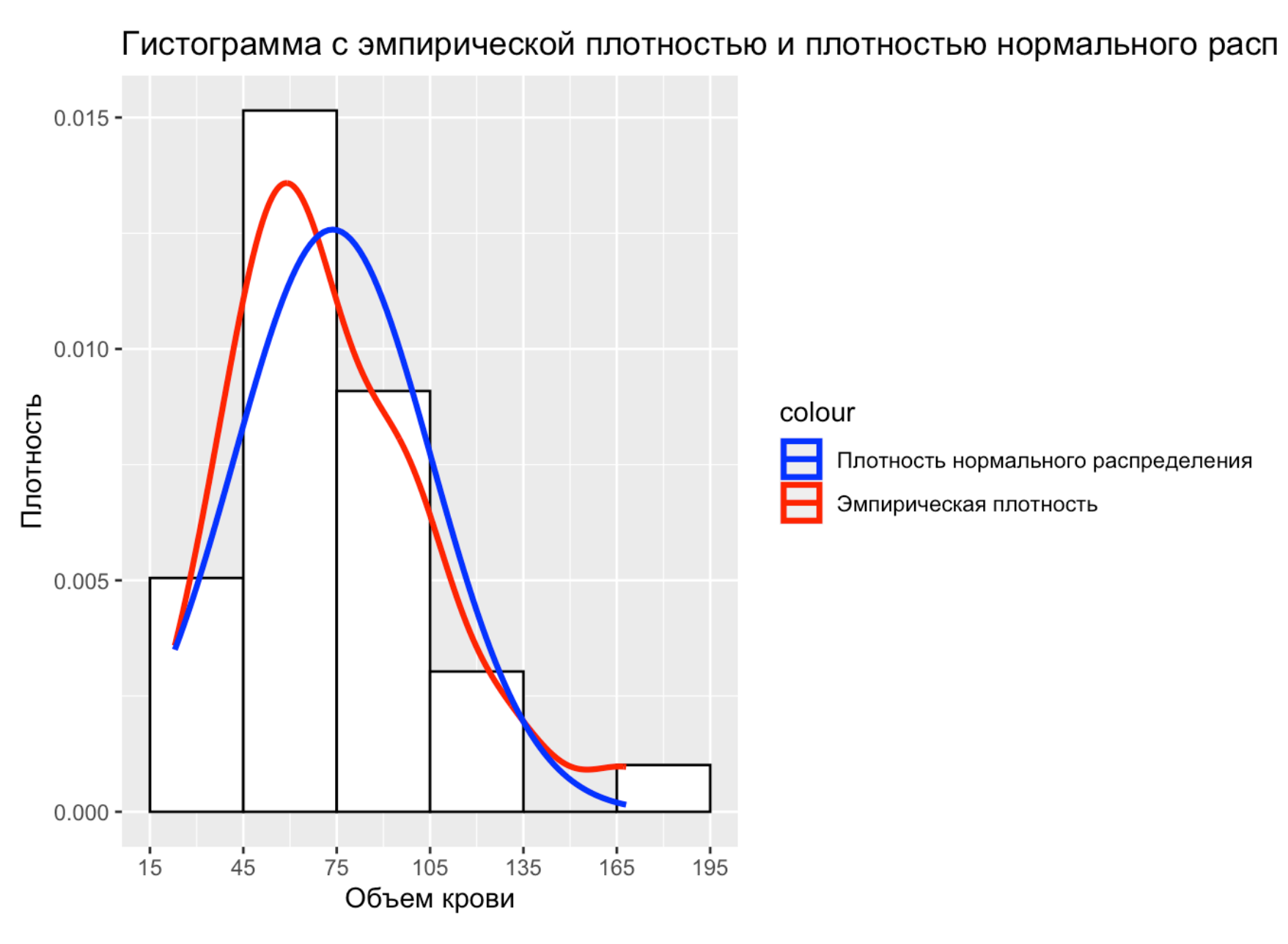
```
aclday1mean <- sum(df$craving.to.alcohol.1,na.rm = TRUE)/nrow(df)
aclday2mean <- sum(df$craving.to.alcohol.2,na.rm = TRUE)/nrow(df)
aclday3mean <- sum(df$craving.to.alcohol.3,na.rm = TRUE)/nrow(df)
aclday9mean <- sum(df$craving.to.alcohol.9,na.rm = TRUE)/nrow(df)

depreday1mean <- sum(df$depressed.mood.1,na.rm = TRUE)/nrow(df)
depreday2mean <- sum(df$depressed.mood.2,na.rm = TRUE)/nrow(df)
depreday3mean <- sum(df$depressed.mood.3,na.rm = TRUE)/nrow(df)
depreday9mean <- sum(df$depressed.mood.9,na.rm = TRUE)/nrow(df)

weaknday1mean <- sum(df$weakness.1,na.rm = TRUE)/nrow(df)
weaknday2mean <- sum(df$weakness.2,na.rm = TRUE)/nrow(df)
weaknday3mean <- sum(df$weakness.3,na.rm = TRUE)/nrow(df)
weaknday9mean <- sum(df$weakness.9,na.rm = TRUE)/nrow(df)

sweatday1mean <- sum(df$sweating.1,na.rm = TRUE)/nrow(df)
sweatday2mean <- sum(df$sweating.2,na.rm = TRUE)/nrow(df)
sweatday3mean <- sum(df$sweating.3,na.rm = TRUE)/nrow(df)
sweatday9mean <- sum(df$sweating.9,na.rm = TRUE)/nrow(df)


dat <- data.frame(
names = c("day 1","day 2","day 3", "day 9"),
values = c(aclday1mean,aclday2mean,aclday3mean,aclday9mean),
values4=c(sweatday1mean,sweatday2mean,sweatday3mean,sweatday9mean),
values2=c(depreday1mean,depreday2mean,depreday3mean,depreday9mean),
values3=c(weaknday1mean,weaknday2mean,weaknday3mean,weaknday9mean)
)

# Преобразование данных в длинный формат
dat_long <- dat %>%
  pivot_longer(cols = c(values, values4,values2,values3), names_to = "Problem", values_to = "Value")

# Создание графика
ggplot(dat_long, aes(x = names, y = Value, fill = Problem)) +
  geom_bar(stat = "identity", position = "dodge",width=0.5) +
  scale_fill_manual(labels = c("Craving to alcohol", "sweating","Depression","Weakness"), values = c("antiquewhit
e2", "aquamarine3","chocolate1","darkorchid")) +
  labs(title = "Гистограмма с doge", x = "Дни", y = "Значения") +
scale_y_continuous(limits = c(0, 2))+
  theme_minimal()
```



Рассматриваем теперь данные в столбце (SV.1), предполагая, что это “ударный объем”- объём крови, выбрасываемый левым желудочком в аорту, правым - в лёгочную артерию.

Выводим гистограмму ударного объема. Для нее выводим также эмпирическую плотность. Также выводим плотность нормального распределения с параметрами N(73.7272, 31.71), полученные по выборки.

```
mean_val <- mean(df$SV.1,na.rm = TRUE)
sd_val <- sd(df$SV.1,na.rm = TRUE)



ggplot(df, aes(x = SV.1))+
geom_histogram(binwidth=30, colour="black", fill="white", aes(y = ..density..))+
  scale_x_continuous(breaks = seq(0, 220, 30)-15)+
  geom_density(aes(color="Эмпирическая плотность"),size=1.1)+
  stat_function(fun = function(x) dnorm(x, mean = mean_val, sd = sd_val), aes(color = "Плотность нормального расп
ределения"),size=1.1)+
  scale_color_manual(values = c("Плотность нормального распределения" = "blue", "Эмпирическая плотность" = "red")
) +
  labs(title = "Гистограмма с эмпирической плотностью и плотностью нормального распределения", x = "Объем крови",
y = "Плотность")
```



Наблюдаем сходство с нормальным распределением. Объем крови разбит на интервалы с размахам равным 30. Это важно для возможности применения критерия хи-квадрат Пирсона. Потому что нужно, чтобы у нас были качественные переменные. Также важно, что предполагаем, что у нас случайная выборка.

```
normval <- rnorm(nrow(df)-1,mean_val,sd_val)
obsval <- df$SV.1|>na.omit()



br <- seq(0,200,30)
interv <- cut(obsval,br)
intervtheor <- cut(normval,br)
inter_count <- table(interv)
inter_count_theor <- table(intervtheor)|>as.data.frame()
tabl <- as.data.frame(inter_count)
tabl <- cbind(tabl,inter_count_theor$Freq)
```

Нулевая гипотеза(H0): Распределение выборки имеет нормальное распределение.

Альтернатива (H1): Распределение выборки не имеет нормальное распределение.

Для уровня значимости 0.05

Сначала проверяем, что ожидаемое значение всех значений больше 5. Согласно интернету, это важно для проверки критерий Пирсона-хи квадрат.

```
obsval <- tabl$Freq
total_observations <- sum(obsval)

expected_frequency <- total_observations / length(obsval)

chisq_test_result <- chisq.test(obsval, p = rep(1/length(obsval), length(obsval)))

expected_values <- chisq_test_result$expected
valid_test <- all(expected_values > 5)

if (valid_test) {
  cat("Матожидание для каждого значения больше 5. Можно пользоваться методом хи-квадрат.\n")
} else {
  cat("Матожидание для как минимум одного из значений меньше либо равен 5. Пользоваться методом хи-квадрат может
быть ошибочным.\n")
}
```

```
## Матожидание для каждого значения больше 5. Можно пользоваться методом хи-квадрат.
```

```
pearson_chi_squared_test <- pearson.test(tabl$Freq)
print(pearson_chi_squared_test)
```

```
##
##  Pearson chi-square normality test
##
## data:  tabl$Freq
## P = 0.66667, p-value = 0.7165
```

Получили выокий p-value=0.7165, значительно выше уровня значимости 0,05. Это означает, что у нас нет оснований отвергнуть нулевую гипотезу.

Посчитаем теперь вручную

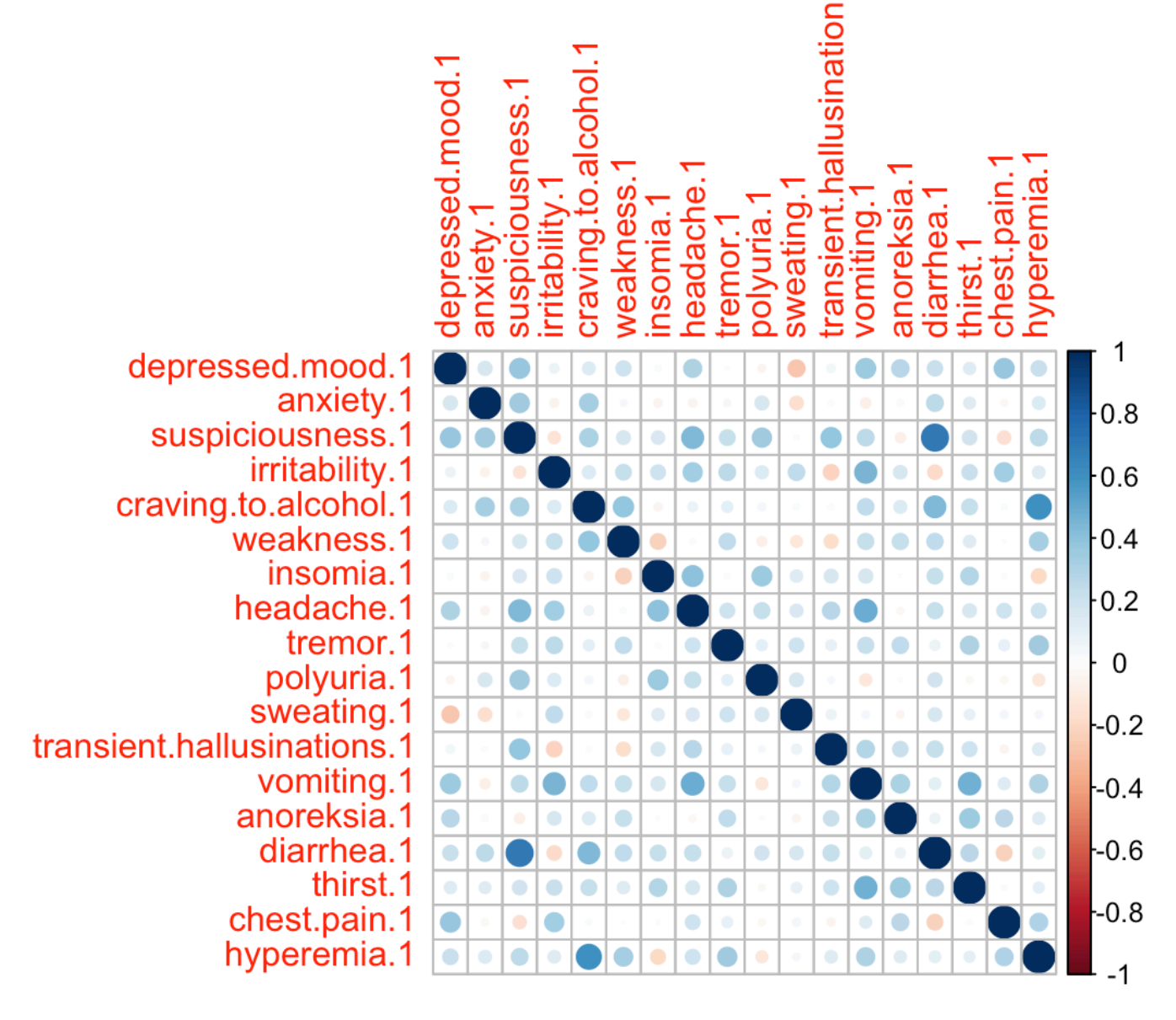
```
tabl0 <- head(tabl, n = nrow(tabl) - 1)
s <- sum(((tabl0$Freq-tabl0$`inter_count_theor$Freq`)**2)/tabl0$`inter_count_theor$Freq`)
s
```

```
## [1] 3.681818
```

Критическое значение для уровня значимости 0.05 равняется 9.4888. Мы туда не попадаем. Нет оснований отвергать нулевую гипотезу.

Посмотрим на корреляцию между категориями болезнями в первый день. Это ничего не говорит про независимость, но это указывает на линейную связь между категориями.

```
new_df <- df[,3:20]
corrplot(cor(new_df))
```



Проверим независимость головной боли и слабости **методом хи квадрат Пирсона.** Формулируем гипотезу:

- **Нулевая гипотеза (H0)**: Головная боль и боль в груди в первый день независимы.
- **Альтернативная гипотеза (H1)**: Головная боль и боль в груди в первый день зависимы.

Для данного теста желательно больше данных, но ради интереса проведем его.

```
tes2 <- table(df$chest.pain.1,df$weakness.1)
tes2
```

```
##
##     0 1 2
##   0 2 6 9
##   1 0 9 4
##   2 1 0 3
```

```
test <- chisq.test(tes2)
test
```

```
##
##  Pearson's Chi-squared test
##
## data:  tes2
## X-squared = 7.9447, df = 4, p-value = 0.09362
```

Получаем p-value=0.09362, это больше 0.05, следовательно делаем вывод, что нет оснований отвергать нулевую гипотезу. Боль в голове и в груди независимы.

Также интересно посмотреть, зависимы или нет головная боль в первый день и во второй.

```
tes2 <- table(df$headache.1,df$headache.2)
tes2
```

```
##
##      0  1  2
##   0 18  3  0
##   1  7  3  1
##   2  1  0  1
```

```
test <- chisq.test(tes2)
test
```

```
##
##  Pearson's Chi-squared test
##
## data:  tes2
## X-squared = 9.7652, df = 4, p-value = 0.04457
```

Значение не сильно, но меньше уровня значимости 0,05, следовательно принимает альтернативу, что данные признаки зависимы

Теперь будем использовать **критерий фишера**, для проверки независимости. Для этого мы данные где пациенты отвечали 1 или 2 объединяем, чтобы была таблица 2x2.

Задаем гипотезу:

- **Нулевая гипотеза (H0)**: Бессоница и жажда в первый день независимы.
- **Альтернативная гипотеза (H1)**: Они зависимы

  Выводим таблицу и p-value, при оценки критерия Фишера

```
df2 <- df
df2[df2==2] <- 1

count_0_A <- sum((df2$insomia.1 == 0))
count_1_A <- sum((df2$insomia.1== 1))

count_0_B <- sum(df2$thirst.1 == 0)
count_1_B <- sum(df2$thirst.1 == 1)


tes <- data.frame(Insomia = c(count_0_A, count_1_A), Thirst = c(count_0_B, count_1_B),
                  row.names = c("No(0)", "Yes(1,2)"),stringsAsFactors = FALSE)

print_df(tes)
```

	Insomia	Thirst
No(0)	7	3
Yes(1,2)	27	31

```
test <- fisher.test(table(df2$insomia.1, df2$thirst.1))
test
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(df2$insomia.1, df2$thirst.1)
## p-value = 0.5112
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   0.03031251 45.51424798
## sample estimates:
## odds ratio
##   2.030953
```

Получили p-value, значительно больше уровня значимости 0.05, следовательно у нас нет оснований отвергнуть основную гипотезу о том что бессоница и жажда независимы.

В заключении воспользуемся **критерием Мак-Немара.** Чтобы условие независимости не выполнялось будем проверять гипотезу на одной болезни в разные дни. То есть одно и то же наблюдение сравнивается до и после воздействия лечением на пациентов. В данном тесте также используется таблица 2x2.

- **Нулевая гипотеза (H0)**: Нет различий в частоте болей в груди между первым и вторым днем в больнице. Это означает, что вероятность испытания боли в груди не зависит от дня в больнице.
- **Альтернативная гипотеза (H1)**: Существуют различия в частоте болей в груди между первым и вторым днем в больнице. Это означает, что вероятность испытания боли в груди различается в зависимости от дня в больнице.

Выводим таблицу и сам тест. Проводим тест на уровне значимости 0,05.

```
count_0_A <- sum((df2$chest.pain.1 == 0))
count_1_A <- sum((df2$chest.pain.1== 1))

count_0_B <- sum(df2$chest.pain.2 == 0)
count_1_B <- sum(df2$chest.pain.2 == 1)

tes3 <- data.frame(Chest_pain_Day1 = c(count_0_A, count_1_A), Chest_pain_Day2 = c(count_0_B, count_1_B),row.names
= c("No(0)", "Yes(1,2)"),stringsAsFactors = FALSE)
print_df(tes3)
```

	Chest_pain_Day1	Chest_pain_Day2
No(0)	17	29
Yes(1,2)	17	5

```
tab <- table(df2$chest.pain.1,df2$chest.pain.2)
mcnemar.test(tab, y = NULL, correct = FALSE)
```

```
##
##  McNemar's Chi-squared test
##
## data:  tab
## McNemar's chi-squared = 12, df = 1, p-value = 0.000532
```

P-value равное 0.000532 является гораздо меньше чем заданный уровень значимости. Следовательно мы отвергаем нулевую гипотезу и принимаем альтернативу, которая утверждает, что существуют различия в частоте болей в груди между первым и вторым днем в больнице. Это означает, что вероятность испытания боли в груди различается в зависимости от дня в больнице. По данным видно, что оно уменьшается. Следовательно лечение в больнице эффективно.