Харинаев Артём 316 группа 25.10.21

1. Задания из файла с семинара

```
drug <- array(c(11, 10, 25, 27,
                16, 22, 4, 10,
                14, 7, 5, 12,
                2, 1, 14, 16,
                6, 0, 11, 12,
                1, 0, 10, 10,
                1, 1, 4, 8,
                4, 6, 2, 1),
              dim = c(2, 2, 8),
              dimnames = list(
              Group = c("Препарат", "Контроль"),
              Response = c("Успешно", "Неудачно"),
              Center = c("1", "2", "3", "4", "5", "6", "7", "8")))

library(reshape)

#чуть подправим функцию, чтобы не появлялось предупреждение
meltnew <- reshape::melt.matrix
body(meltnew)[8][[1]] <- 'dn[char] <- lapply(dn[char], type.convert, as.is = TRUE)'

drug.df <- meltnew(drug, varnames = names(dimnames(drug)))

#Пересчитаем процентные соотношения по группам в клиниках
library(dplyr)
```

```
##
## Присоединяю пакет: 'dplyr'
```

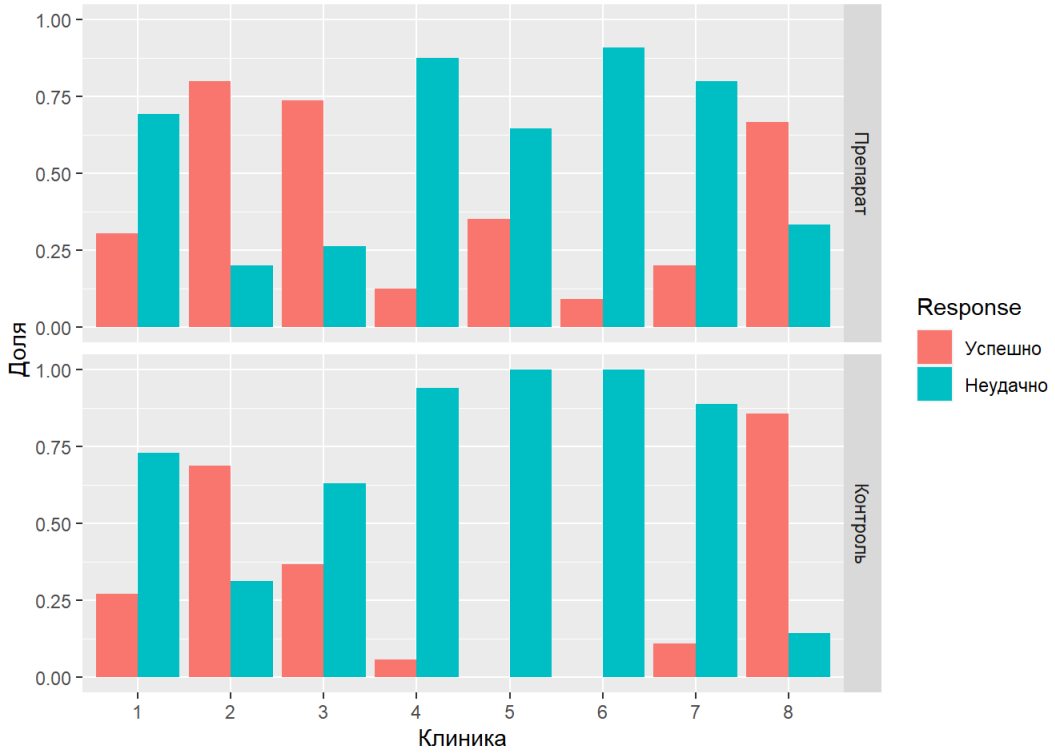
```
## Следующий объект скрыт от 'package:reshape':
##
##     rename
```

```
## Следующие объекты скрыты от 'package:stats':
##
##     filter, lag
```

```
## Следующие объекты скрыты от 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
drug_per <- group_by(drug.df, Group,Center) %>% transmute(Response, percent = value/sum(value))

library(ggplot2)
p <- ggplot(data = drug_per, aes(x = Center, y = percent,
                                 fill = Response))+xlab("Клиника")+ylab("Доля")
p + geom_bar(stat = "identity", position = "dodge") + facet_grid(Group~.) + scale_x_discrete(limits=facto
r(1:8))
```

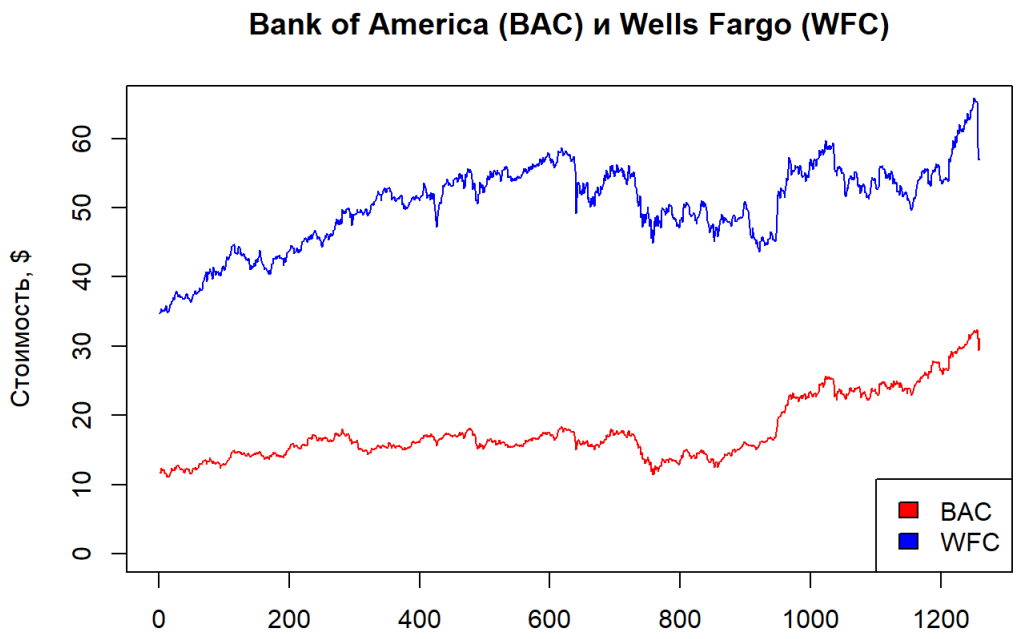


2. Корреляция в собственных данных

2.1 Коэффициент корреляции Пирсона

```
data <- read.csv(file='..\\dataset.csv')
data$date <- as.Date(data$date)
data$year <- as.numeric(format(data$date, format='%Y'))
bac <- subset(data, data$Name=='BAC')
wfc <- subset(data, data$Name=='WFC')
```

```
plot(1:1259, bac$open, type='l', col='red', ylim=c(0,65),
     main = 'Bank of America (BAC) и Wells Fargo (WFC)',
     xlab='', ylab='Стоимость, $')
lines(1:1259, wfc$open, col='blue')
legend('bottomright', legend=c('BAC', 'WFC'), fill=c('red', 'blue'))
```



Условия применимости критерия:

1. данные близки к нормальному распределению
2. длины выборок равны

Нормальность данных

```
library(nortest)
```

```
normalize <- function (set, col, year, top, bottom){
  set_year <-  set[col][set['year']==year]
  set_norm <- sapply(set_year, FUN=function (x) {(x-mean(set_year))/var(set_year)})
  set_norm_q <- set_norm[(set_norm <= quantile(set_norm, top)) &
                         (set_norm >= quantile(set_norm, bottom))]
  set_norm_q
}
```

```
bac_open_n <- normalize(bac, 'open', 2015, 0.6, 0.3)
wfc_open_n <- normalize(wfc, 'open', 2015, 0.7, 0.4)
lillie.test(bac_open_n)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  bac_open_n
## D = 0.07809, p-value = 0.311
```

```
lillie.test(wfc_open_n)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  wfc_open_n
## D = 0.075419, p-value = 0.3637
```

p-value достаточно велико, данные скорее всего близки к нормальному распределению

Рассмотрим корреляцию цен акций компаний Bank of America (BAC) и Wells Fargo (WFC)

```
if (length(bac_open_n) != length(wfc_open_n)){
  len <- min(length(bac_open_n), length(wfc_open_n))
  bac_open_n <- bac_open_n[1:len]
  wfc_open_n <- wfc_open_n[1:len]
}
cor.test(bac_open_n, wfc_open_n)
```

```
##
##  Pearson's product-moment correlation
##
## data:  bac_open_n and wfc_open_n
## t = 1.2815, df = 73, p-value = 0.2041
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.08136954  0.36307051
## sample estimates:
##       cor
## 0.1483317
```

p-value достаточно велико (20%), значит нулевую гипотезу о некоррелированности величин нельзя отвергнуть с уверенностью, возможно, величины некоррелированы, однако вычисленный коэффициент корреляции не равен 0

Теперь рассмотрим корреляцию цен Bank of America (BAC) и Kellogg (K) (компания специализируется на производстве сухих завтраков и продуктов питания быстрого приготовления)

```
kellogg <- subset(data, data$Name=='K')
kellogg_open_n <- normalize(kellogg, 'open', 2015, 0.7, 0.4)
lillie.test(kellogg_open_n)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  kellogg_open_n
## D = 0.075638, p-value = 0.3592
```

p-value достаточно велико, данные скорее всего близки к нормальному распределению

```
if (length(bac_open_n) != length(kellogg_open_n)){
  len <- min(length(bac_open_n), length(kellogg_open_n))
  bac_open_n <- bac_open_n[1:len]
  kellogg_open_n <- kellogg_open_n[1:len]
}
cor.test(bac_open_n, kellogg_open_n)
```

```
##
##  Pearson's product-moment correlation
##
## data:  bac_open_n and kellogg_open_n
## t = 1.6551, df = 73, p-value = 0.1022
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.03844738  0.39987624
## sample estimates:
##       cor
## 0.1901739
```

Коэффициент корреляции мал, а p-value недостаточно мало, это значит, что нельзя отвергнуть гипотезу о некоррелированности этих величин полагаясь на этот тест, и скорее всего цены не коррелированы. Что довольно логично, т.к. компании взяты из разных секторов

Рассмотрим корреляцию цены и объема продаж

```
bac_vol_n <- normalize(bac, 'volume', 2015, 0.6, 0.3)
lillie.test(bac_vol_n)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  bac_vol_n
## D = 0.082781, p-value = 0.2308
```

p-value достаточно велико, данные скорее всего близки к нормальному распределению

```
if (length(bac_open_n) != length(bac_vol_n)){
  len <- min(length(bac_open_n), length(bac_vol_n))
  bac_open_n <- bac_open_n[1:len]
  bac_vol_n <- bac_vol_n[1:len]
}
cor.test(bac_open_n, bac_vol_n)
```

```
##
##  Pearson's product-moment correlation
##
## data:  bac_open_n and bac_vol_n
## t = -0.54681, df = 73, p-value = 0.5862
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2866747  0.1654924
## sample estimates:
##         cor
## -0.06386836
```

p-value достаточно велико, нельзя отвергнуть гипотезу о некоррелированности данных. Однако, можно заметить, что коэффициент корреляции отрицательный, что логично, т.к. при низкой цене ликвидность наоборот повышается

2.2 Коэффициент корреляции Спирмена

Т.к. критерий ранговый, то для корректного вычисления p-value, необходимы данные без повторений

```
bac_uniq <- bac_open_n[!duplicated(bac_open_n) & !duplicated(wfc_open_n)]
wfc_uniq <- wfc_open_n[!duplicated(bac_open_n) & !duplicated(wfc_open_n)] #длины выборок равны
lillie.test(bac_uniq)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  bac_uniq
## D = 0.084429, p-value = 0.6354
```

```
lillie.test(wfc_uniq)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  wfc_uniq
## D = 0.074774, p-value = 0.8033
```

p-value достаточно велико, данные скорее всего близки к нормальному распределению

```
cor.test(bac_uniq, wfc_uniq, method='spearman')
```

```
##
##  Spearman's rank correlation rho
##
## data:  bac_uniq and wfc_uniq
## S = 10706, p-value = 0.4016
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.1324852
```

p-value велико, нулевую гипотезу нельзя отвергнуть, полагаясь на этот тест

2.3 Коэффициент корреляции Кендалла

Т.к. критерий ранговый, то для корректного вычисления p-value, необходимы данные без повторений

```
cor.test(bac_uniq, wfc_uniq, method='kendall')
```

```
##
##  Kendall's rank correlation tau
##
## data:  bac_uniq and wfc_uniq
## T = 473, p-value = 0.3649
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.09872242
```

p-value велико, нулевую гипотезу нельзя отвергнуть, полагаясь на этот тест

3.

3.1 Метод Хи-квадрат

Создадим таблицу сопряженности. По строкам расположим банковские компании. По столбцам - годы. В ячейках - кол-во дней в году, когда цена акции увеличивается

```
library(reshape2)
```

```
##
## Присоединяю пакет: 'reshape2'
```

```
## Следующие объекты скрыты от 'package:reshape':
##
##     colsplit, melt, recast
```

```
data$year <- as.numeric(format(data$date, format='%Y'))
data$day_profit <- data$close > data$open
banks <- subset(data, (data$Name=='JPM' | data$Name == 'BAC' | data$Name == 'WFC') &
                  data$year<=2017 & data$year>=2013)
banks_table <- dcast(banks, Name ~ year, value.var = 'day_profit', fun.aggregate = sum)
banks_ct <- data.matrix(banks_table)
row.names(banks_ct) <- c('BAC', 'JPM', 'WFC')
banks_ct <- banks_ct[,-1]
banks_ct
```

```
##     2013 2014 2015 2016 2017
## BAC  103  125  122  131  112
## JPM  118  143  130  140  128
## WFC  114  133  119  125  127
```

```
chisq.test(banks_ct)
```

```
##
##  Pearson's Chi-squared test
##
## data:  banks_ct
## X-squared = 1.4728, df = 8, p-value = 0.9932
```

Заметно, что цены акций компаний из схожего сектора растут и падают коррелировано друг с другом

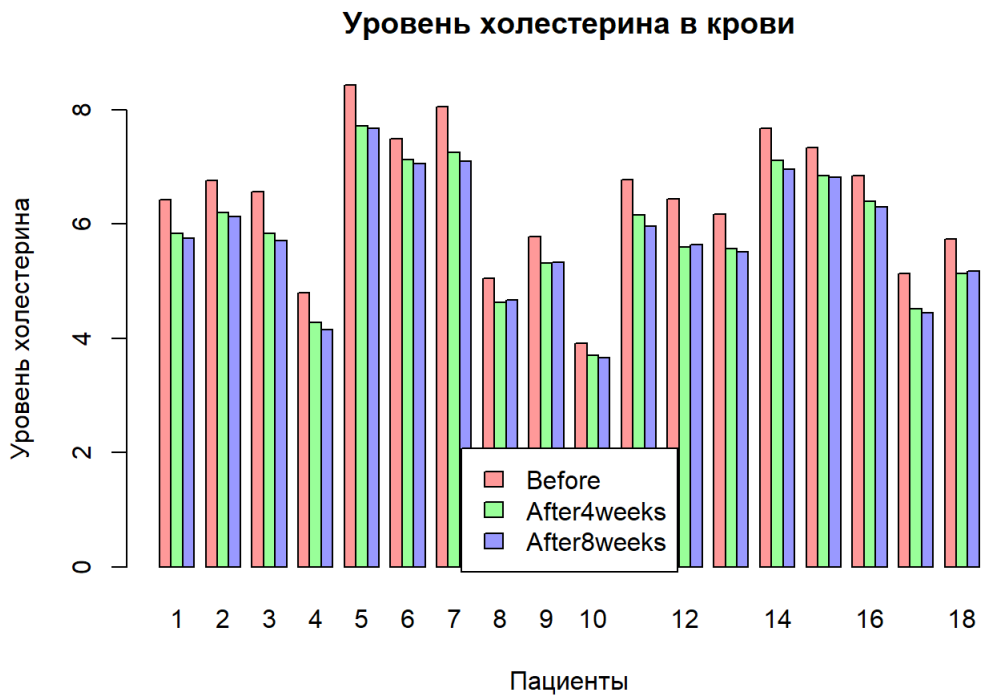
3.2 Тест МакНемара

Тест проводится для таблицы 2*2, в которой содержатся данные, разделенные по бинарным категориям, причем данные

измеряются дважды (например, до какого-то преобразования и после, либо просто двумя разными способами) и заносятся в столбцы и строки

Рассмотрим датасет, содержащий наблюдения над уровнем холестерина в крови у 18-ти людей, употреблявших особый вид маргарина без транс-жиров

```
chol <- read.csv(file='..\\Cholesterol_R.csv')
data <- chol[,2:4]
barplot(t(data.matrix(data)), beside=TRUE, col = rainbow(3, alpha=0.4),
        names.arg=chol$п.їID,
        main='Уровень холестерина в крови',
        xlab='Пациенты', ylab='Уровень холестерина')
legend('bottom', colnames(data), fill=rainbow(3, alpha=0.4))
```



Составим таблицу, где укажем в строках сождеражание холестерина до диеты выше и ниже порогового значения, а в столбцах сождеражание холестерина после 4 недель диеты

```
lim <- 6.4
less_less <- sum((chol$Before < lim) & (chol$After4weeks < lim))
less_great <- sum((chol$Before < lim) & (chol$After4weeks >= lim))
great_less <- sum((chol$Before >= lim) & (chol$After4weeks < lim))
great_great <- sum((chol$Before >= lim) & (chol$After4weeks >= lim))
mat_chol <- matrix(c(less_less, great_less, less_great, great_great), nrow = 2,
                   dimnames = list("Before" = c("Less", "Greater"),
                    "After 4 weeks" = c("Less", "Greater")))
mat_chol
```

```
##          After 4 weeks
## Before    Less Greater
##   Less       7       0
##   Greater    5       6
```

```
mcnemar.test(mat_chol)
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  mat_chol
## McNemar's chi-squared = 3.2, df = 1, p-value = 0.07364
```

p-value удовлетворяет уровню значимости 0.1, отвергаем нулевую гипотезу. То есть, диета влияет на превышение порогового значения уровня холестерина в крови

3.3 Тест Кохрана-Мантеля-Хензеля

Необходима 3-х мерная таблица. Третьим измерением добавим вид маргарина (в датасете A или B)

```
lim <- 6.4

less_less_A <- sum((chol$Before < lim) & (chol$After4weeks < lim) & (chol$Margarine=='A'))
less_great_A <- sum((chol$Before < lim) & (chol$After4weeks >= lim) & (chol$Margarine=='A'))
great_less_A <- sum((chol$Before >= lim) & (chol$After4weeks < lim) & (chol$Margarine=='A'))
great_great_A <- sum((chol$Before >= lim) & (chol$After4weeks >= lim) & (chol$Margarine=='A'))

less_less_B <- sum((chol$Before < lim) & (chol$After4weeks < lim) & (chol$Margarine=='B'))
less_great_B <- sum((chol$Before < lim) & (chol$After4weeks >= lim) & (chol$Margarine=='B'))
great_less_B <- sum((chol$Before >= lim) & (chol$After4weeks < lim) & (chol$Margarine=='B'))
great_great_B <- sum((chol$Before >= lim) & (chol$After4weeks >= lim) & (chol$Margarine=='B'))

arr_chol <- array(c(less_less_A, great_less_A, less_great_A, great_great_A,
                    less_less_B, great_less_B, less_great_B, great_great_B),
                  dim = c(2,2,2),
                  dimnames = list("Before" = c("Less", "Greater"),
                    "After 4 weeks" = c("Less", "Greater"),
                    "Margarine type" = c('A', 'B')))
arr_chol
```

```
## , , Margarine type = A
##
##          After 4 weeks
## Before    Less Greater
##   Less       5       0
##   Greater    1       3
##
## , , Margarine type = B
##
##          After 4 weeks
## Before    Less Greater
##   Less       2       0
##   Greater    4       3
```

```
mantelhaen.test(arr_chol)
```

```
##
##  Mantel-Haenszel chi-squared test with continuity correction
##
## data:  arr_chol
## Mantel-Haenszel X-squared = 3.5588, df = 1, p-value = 0.05923
## alternative hypothesis: true common odds ratio is not equal to 1
## 95 percent confidence interval:
##  NaN NaN
## sample estimates:
## common odds ratio
##               Inf
```

p-value мало, значит, что тип маргарина влияет на результат диеты (B оказался лучше, чем A)