

Практика №3 Проверка гипотез однородности

Долаева А.Р., г. 20.М04-мм

12/05/2021

Вариант №9.

Данные (addicts.xls). Варианты метрической переменной (variable), категориальной с двумя градациями (factor.2), категориальной с четырьмя градациями (factor.4) соответственно по 9 варианту:

variant	variable	factor.2	factor.4
9	rabdru	curwor	prcod

rabdru - оценка риска в употреблении наркотиков

curwor - занятость:

0 – не работает,

1 – работает

prcod - лечебный препарат:

1 – ноутрексон и плацебо,

2 – плацебо и плацебо,

3 – ноутрексон и антидепрессант,

4 – плацебо и антидепрессант.

```
var <- addicts$rabdru;
fact2 <- addicts$curwor;
fact4 <- addicts$prcod
```

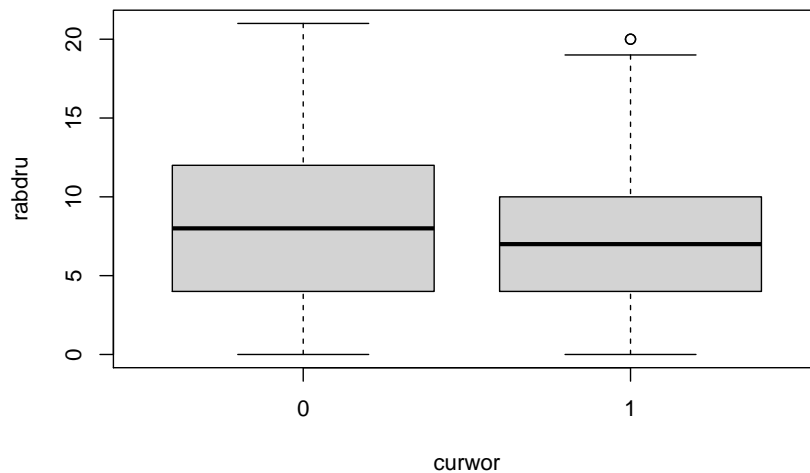
1. Проверить гипотезу о равенстве дисперсий двух выборок и в соответствии с выводом применить критерий Стьюдента для проверки равенства средних. Использовать вариант группирующей переменной factor.2 Распределение метрических данных по категориальным признакам

```
df2<-na.omit(data.frame(group=as.factor(fact2),X=as.numeric(var)));
table(fact2);
```

```
## fact2
##    0    1
## 204   76
```

График “ящика с усами”

```
boxplot(X~group, df2, xlab = "curwor", ylab = "rabdru")
```



У тех, кто не работает(0), риск употребления наркотиков в среднем выше, чем у тех, кто работает(1). Требуется проверить значимость различия дисперсий и средних между выборками категориальных признаков curwor (0 и 1).

Применение критерия Шапиро–Уилка для проверки на согласованность с нормальным распределением

```
p.Sh<-with(df2, tapply(X, group, function(x) shapiro.test(x)$p.value));p.Sh;
```

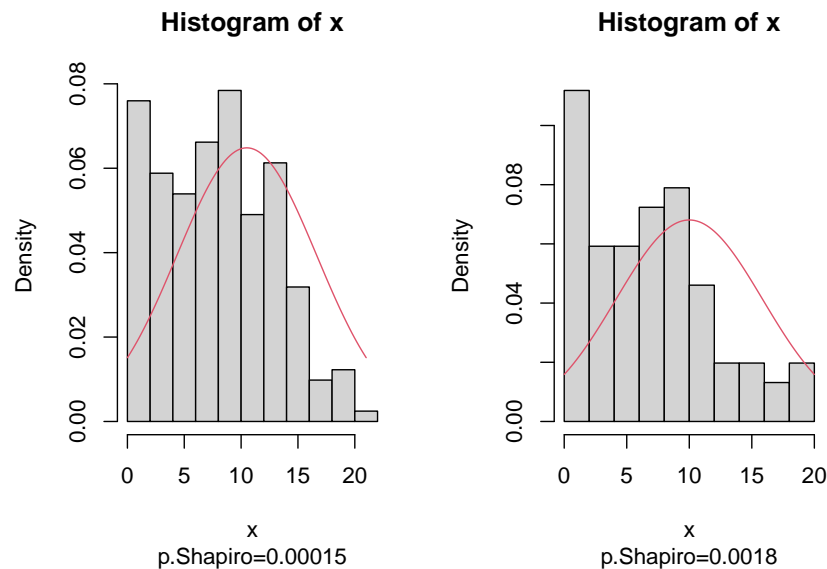
```
##           0           1
## 0.0001468372 0.0018119557
```

Доверительные уровни ниже 0.05, значит гипотеза согласия с нормальным распределением отвергается для обеих выборок.

(Следует применить непараметрические критерии. Вычислим позже.)

```
Fig<-function(x)
{
  hist(x,freq=FALSE);
  f1<-function(x)dnorm(x,mean(x,na.rm=TRUE),sd(x,na.rm=TRUE));
  curve(f1,min(x),max(x),col=2,add=TRUE);
  title(sub=paste("p.Shapiro",format(shapiro.test(x)$p.value,4,2),sep="="))
};

op<-par(mfrow=c(1,2));
Fig(df2$X[df2$group==0]);
Fig(df2$X[df2$group==1])
```



Как видно из гистограмм распределений выборок по категориальным признакам, графики имеют выраженную правостороннюю асимметрию: 0.21 и 0.61.

```
c(skewness(df2$X[df2$group==0]), skewness(df2$X[df2$group==1]))
```

```
## [1] 0.2098638 0.6141859
```

Хоть выборки не согласованы с нормальным распределением, применим критерии Фишера и Стьюдента для проверки равенства дисперсий. (скрепя сердцем)

параметрические критерии: статистика Фишера и Стьюдента

```
p.F<-var.test(X~group, df2)$p.value;p.F;
```

```
## [1] 0.8163512
```

Гипотеза равенства дисперсий не отвергается, так как из статистики Фишера получаем доверительный уровень $0.82 > 0.05$. Дисперсии оказались одинаковыми, значит мы можем применить статистику Стьюдента для вычисления значимости различия в риске употребления наркотиков у работающих и нет.

```
p.T<-t.test(X~group, df2, var.equal=TRUE)$p.value;p.T
```

```
## [1] 0.215805
```

$pvalue = 0.22 > 0.05$, значит различие может быть случайным.

```
mm<-tapply(df2$X,df2$group, function(x) mean(x, na.rm = TRUE));
Sd<-tapply(df2$X,df2$group, function(x) sd(x, na.rm = TRUE));
nn<-tapply(df2$X,df2$group, function(x) length(na.omit(x)));
err<-Sd/sqrt(nn);
paste(mm[1], '±', err[1], ' / ', nn[1]);
```

```
## [1] "8.35588235294118 ± 0.352034658769551 / 204"
```

```
paste(mm[2], '±', err[2], ' / ', nn[2]);
```

```
## [1] "7.51315789473684 ± 0.588036704025884 / 76"
```

Среднее в группе (0) равно 8.36 ± 0.35 по 204 объему выборки, в группе (1) равно $7.51 \pm 0.59/12$ по 76, различие незначимо, так как $p=0.22$

1.1.Повторить обработки с применением непараметрических аналогов для независимых выборок. Для проверки гипотезы однородности двух медиан используется критерий Вилкоксона (точный) или его приближенный вариант критерий Манна-Уитни.

Вычисляем критерии: точный Вилкоксона, Манна-Уитни и Манна-Уитни с поправкой. И выводим квантили.

```
wilcox.test(X~group,df2,exact=TRUE);

## Warning in wilcox.test.default(x = c(6, 19, 3, 9, 13, 9, 12, 1, 4, 12, 12, :
## cannot compute exact p-value with ties

##
## Wilcoxon rank sum test with continuity correction
##
## data: X by group
## W = 8595.5, p-value = 0.1609
## alternative hypothesis: true location shift is not equal to 0

wilcox.test(X~group,df2,exact=FALSE,correct=FALSE);

##
## Wilcoxon rank sum test
##
## data: X by group
## W = 8595.5, p-value = 0.1607
## alternative hypothesis: true location shift is not equal to 0

wilcox.test(X~group,df2,exact=FALSE);

##
## Wilcoxon rank sum test with continuity correction
##
## data: X by group
## W = 8595.5, p-value = 0.1609
## alternative hypothesis: true location shift is not equal to 0

quantile(df2$X[df2$group==0]);

## 0% 25% 50% 75% 100%
## 0 4 8 12 21

quantile(df2$X[df2$group==1]);

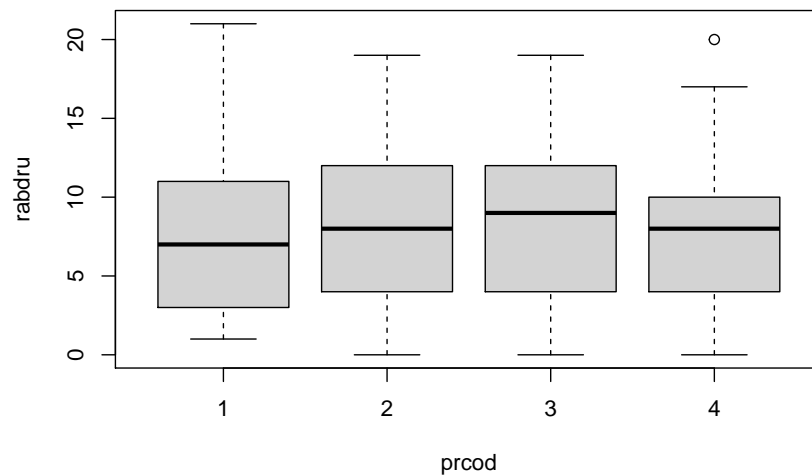
## 0% 25% 50% 75% 100%
## 0 4 7 10 20
```

Несколько видов непараметрического критерия Вилкоксона показывают доверительный уровень $0.16 > 0.05$. Значит различие в средних незначимо.

```
df4<-na.omit(data.frame(group=as.factor(fact4),X=as.numeric(var)));
table(df4$group)
```

2. Применить однофакторный дисперсионный анализ в случае фактора с четырьмя градациями и множественные сравнения с разными поправками. Проверить гипотезу о равенстве дисперсий

```
##
## 1 2 3 4
## 70 70 70 70
boxplot(X~group,xlab="prcod",ylab="rabdru",data=df4)
```



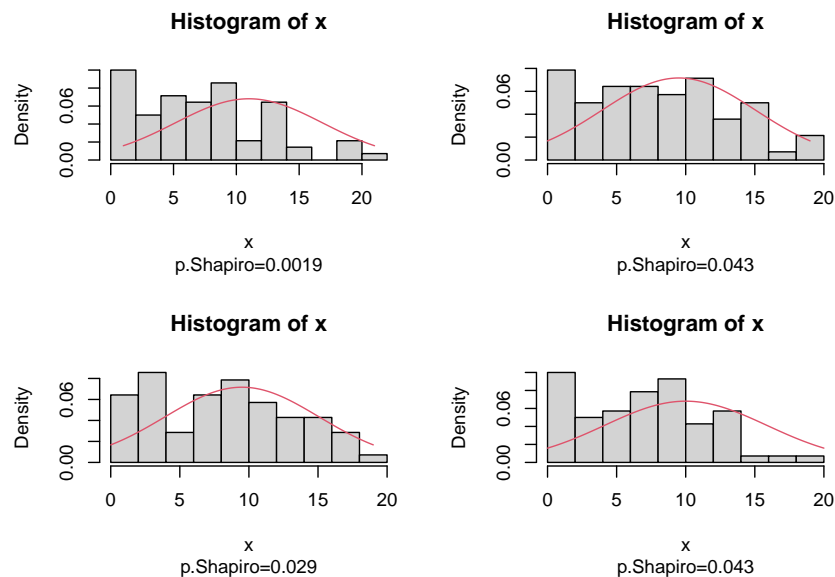
Из графика можно предположить, что риск употреблении наркотиков (rabdru) в среднем выше при приёме ноутрексона и антидепрессантов (prcod = 3) и ниже при приёме ноутрексона и плацебо (prcod = 1).

```
p.Sh<-with(df4, tapply(X, group, function(x) shapiro.test(x)$p.value));p.Sh
```

```
##          1          2          3          4
## 0.001872653 0.042797629 0.029384262 0.042809255
```

Выборки не согласуются с нормальным распределением (сердце всё ещё скрепим), но всё равно проведём тест.

```
op<-par(mfrow=c(2,2));
Fig(df4$X[df4$group==1]);
Fig(df4$X[df4$group==2]);
Fig(df4$X[df4$group==3]);
Fig(df4$X[df4$group==4])
```



Наблюдаем правостороннюю асимметрию в некоторых выборках.

```
c(skewness(df4$X[df4$group==1]), skewness(df4$X[df4$group==2]),
  skewness(df4$X[df4$group==3]), skewness(df4$X[df4$group==4]))
```

```
## [1] 0.5884398 0.1978376 0.1229115 0.2859349
```

Проверим гипотезу о равенстве дисперсий (хоть нет соответствия с нормальным распределением).

критерий Бартлетта для проверки равенства дисперсий

```
bartlett.test(X~group,df4)
```

```
##
## Bartlett test of homogeneity of variances
##
## data: X by group
## Bartlett's K-squared = 1.6776, df = 3, p-value = 0.6419
```

критерий Левена для проверки гипотезы о равенстве дисперсий

```
with(df4,levene.test(X,group))
```

```
##
## Modified robust Brown-Forsythe Levene-type test based on the absolute
## deviations from the median
##
## data: X
## Test Statistic = 0.71207, p-value = 0.5455
```

Оба критерия показывают высокий доверительный уровень $0.64 > 0.05$ (Бартлетт) и $0.55 > 0.05$ (Левен) Значит делаем вывод, что гипотеза о равенстве дисперсий не отвергается, они одинаковы.

Так как соблюдено требование о равенстве дисперсий, мы можем применить однофакторный дисперсионный анализ.

однофакторный дисперсионный анализ

```
ao<-aov(X~group,df4)
summary(ao)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## group          3      63    20.98   0.818   0.485
## Residuals     276    7079    25.65
```

pvalue = 0.485 > 0.05, что сообщает нам о возможной случайности различия дисперсий. Значит различие при применении разных лекарств и риска употреблении наркотиков незначимо.

Выведем статистические показатели

```
mm<-tapply(df4$X,df4$group, function(x) mean(x, na.rm = TRUE));
Sd<-tapply(df4$X,df4$group, function(x) sd(x, na.rm = TRUE));
nn<-tapply(df4$X,df4$group, function(x) length(na.omit(x)));
err<-Sd/sqrt(nn);
paste(mm[1], '±', err[1], ' / ', nn[1]);
```

```
## [1] "7.8 ± 0.626403186758299 / 70"
```

```
paste(mm[2], '±', err[2], ' / ', nn[2]);
```

```
## [1] "8.53714285714286 ± 0.625669567400976 / 70"
```

```
paste(mm[3], '±', err[3], ' / ', nn[3]);
```

```
## [1] "8.64285714285714 ± 0.618230867476973 / 70"
```

```
paste(mm[4], '±', err[4], ' / ', nn[4]);
```

```
## [1] "7.52857142857143 ± 0.547428178096912 / 70"
```

```
quantile(df4$X[df4$group==1]);
```

```
##      0%   25%   50%   75%  100%
```

```
##  1.00  3.00  7.00 10.75 21.00
```

```
quantile(df4$X[df4$group==2]);
```

```
##      0%   25%   50%   75%  100%
```

```
##  0.00  4.25  8.00 12.00 19.00
```

```
quantile(df4$X[df4$group==3]);
```

```
##      0%   25%   50%   75%  100%
```

```
##      0      4      9     12     19
```

```
quantile(df4$X[df4$group==4])
```

```
##      0%   25%   50%   75%  100%
```

```
##      0      4      8     10     20
```

2.1. Повторить обработки с применением непараметрических аналогов. Проверим гипотезу однородности двух медиан критерием Краскела-Уоллиса для более двух выборок.

```
kruskal.test(X~group,df4)
```

```
##
```

```
##  Kruskal-Wallis rank sum test
```

```
##
```

```
## data: X by group
## Kruskal-Wallis chi-squared = 2.4024, df = 3, p-value = 0.4932
```

Доверительный уровень $0.49 > 0.05$, различие в средних может быть случайным, незначимо.

Применим множественные сравнения: с поправкой Бонферони и Тьюки. (Для проверки различия среди групп).

```
out <- LSD.test(ao,"group", p.adj="none",group=FALSE)$comparison$pvalue
out1 <- LSD.test(ao,"group", p.adj="bonferroni",group=FALSE)$comparison$pvalue
out2 <- LSD.test(ao,"group", p.adj="holm",group=FALSE)$comparison$pvalue
out3 <- LSD.test(ao,"group", p.adj="BH",group=FALSE)$comparison$pvalue

res <- data.frame(pair = c('1 - 2', '1 - 3', '1 - 4', '2 - 3', '2 - 4', '3 - 4'),
                  none = out, bonferroni = out1, holm = out2, BH=out3)
res
```

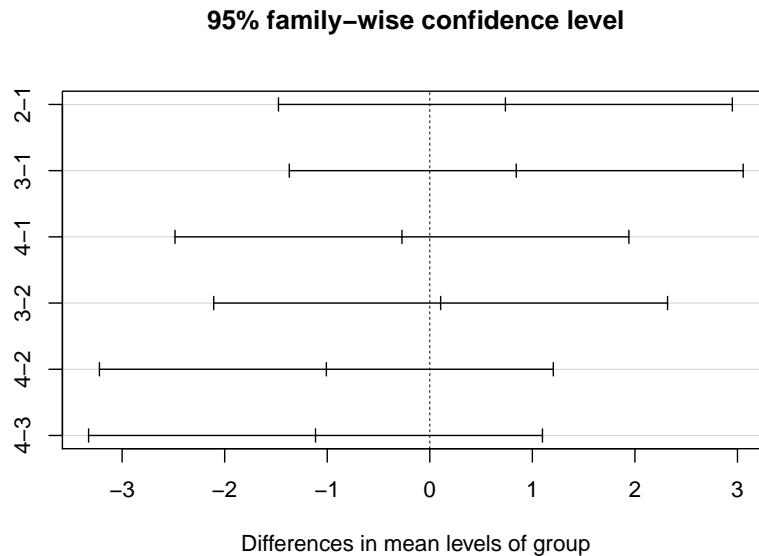
```
##      pair    none bonferroni holm    BH
## 1 1 - 2 0.3899          1    1 0.5849
## 2 1 - 3 0.3257          1    1 0.5849
## 3 1 - 4 0.7514          1    1 0.9017
## 4 2 - 3 0.9018          1    1 0.9018
## 5 2 - 4 0.2398          1    1 0.5849
## 6 3 - 4 0.1941          1    1 0.5849
```

Так как в выборках одинаковое число наблюдений, можем применить критерий Тьюки.

```
TukeyHSD(ao, "group", ordered = TRUE)
```

```
##      Tukey multiple comparisons of means
##      95% family-wise confidence level
##      factor levels have been ordered
##
## Fit: aov(formula = X ~ group, data = df4)
##
## $group
##      diff      lwr      upr      p adj
## 1-4 0.2714286 -1.941313 2.484170 0.9889482
## 2-4 1.0085714 -1.204170 3.221313 0.6411761
## 3-4 1.1142857 -1.098456 3.327028 0.5626670
## 2-1 0.7371429 -1.475599 2.949885 0.8248718
## 3-1 0.8428571 -1.369885 3.055599 0.7583756
## 3-2 0.1057143 -2.107028 2.318456 0.9993257
```

```
plot(TukeyHSD(ao, "group"))
```

Значимые отличия в группах, применявших разные лекарства, не наблюдаются ни по одному из методов множественных сравнений.

3. Повторить обработки с применением непараметрических аналогов. Вычисления приведены в разделах 1.1 и 2.1.

4. Для двух зависимых переменных BDI.1 BDI.2 проверить однородность изменений во времени по критерию Стьюдента для зависимых выборок и по ранговому критерию Вилкоксона. Шкала депрессии Бека (Beck Depression Inventory, BDI) в 1, 2, 5 и 7 дни

variant	X1	X2	X3	X4
9	BDI.1	BDI.2	BDI.5	BDI.7

```
colMeans(dataNF[,c("BDI.1", "BDI.2")], na.rm = TRUE)
```

```
##      BDI.1      BDI.2
## 20.35061 13.00398
```

Уменьшение среднего показателя депрессии на второй день.

критерий Стьюдента для проверки равенства средних

```
t.test(dataNF$BDI.1, dataNF$BDI.2, paired = TRUE)
```

```
##
## Paired t-test
##
## data: dataNF$BDI.1 and dataNF$BDI.2
## t = 13.81, df = 249, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  5.881656 7.838344
```

```
## sample estimates:
## mean of the differences
##                6.86
```

Доверительный уровень $2.2e-16 < 0.05$, значит уменьшение депрессии на второй день можно считать значимым.

Непараметрические критерии однородности для двух и нескольких зависимых выборок.

критерий знаков

```
tab<-table(sign(dataNF$BDI.2-dataNF$BDI.1));tab
```

```
##
##  -1   0   1
## 200   7  43

binom.test(min(tab), sum(tab), p = 0.5,
alternative = "less",
conf.level = 0.95)
```

```
##
## Exact binomial test
##
## data: min(tab) and sum(tab)
## number of successes = 7, number of trials = 250, p-value < 2.2e-16
## alternative hypothesis: true probability of success is less than 0.5
## 95 percent confidence interval:
##  0.000000 0.051947
## sample estimates:
## probability of success
##                0.028
```

ранговый Вилкоксона

```
wilcox.test(dataNF$BDI.2, dataNF$BDI.1,paired=TRUE,exact = FALSE)
```

```
##
## Wilcoxon signed rank test with continuity correction
##
## data: dataNF$BDI.2 and dataNF$BDI.1
## V = 2692.5, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Friedman

```
A<-apply(colnames(dataNF),function(x)substr(x,1,nchar("BDI")))
dat<-dataNF[,names(A[A=="BDI"])]
friedman.test(as.matrix(dat[,c(1,2,5,7)]))
```

```
##
## Friedman rank sum test
##
## data: as.matrix(dat[, c(1, 2, 5, 7)])
## Friedman chi-squared = 212.2, df = 3, p-value < 2.2e-16
```

Различие в средних значимо, $pvalue < 0.05$.

5. Для зависимых переменных с факторами “PRCOD.1” и “SEX.1” выполнить ANOVA Repeated Measures. Проверить значимость факторов “PRCOD.1” и “SEX.1” времени и эффектов взаимодействия. Исследование динамики показателей. Выбираем данные: группирующие переменные “PRCOD.1” и “SEX.1” и наблюдения показателей шкалы депрессии Бека в три момента времени BD1, BD2, BD5.

применим ANOVA Repeated Measures и выведем таблицу дисперсионного анализа

```
dat.AR<-na.omit(data.frame(PRCOD=dataNF$PRCOD.1,SEX =dataNF$SEX.1,
                          BD1 = dataNF$BDI.1, BD2 = dataNF$BDI.2, BD5 = dataNF$BDI.5))
m<-ncol(dat.AR)-2;
```

```
dat.AR.T<-data.frame(
  stack(dat.AR[, -seq(2)]),
  # cols of BDI
  sub=as.factor(rep(seq(nrow(dat.AR)),m)),
  # PRCOD factor
  gr1=as.factor(rep(dat.AR[,1],m)),
  # SEX factor
  gr2=as.factor(rep(dat.AR[,2],m)) );
dat.AR.T$values<-as.numeric(dat.AR.T$values);

formula <-values~(gr1 + gr2)*ind + Error(sub/ind);
ao <-aov(formula, dat.AR.T);
summary(ao)
```

```
##
## Error: sub
##           Df Sum Sq Mean Sq F value Pr(>F)
## gr1        3   1115    371.7   3.329 0.02148 *
## gr2        1   1065   1065.3   9.541 0.00242 **
## Residuals 140   15631    111.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Error: sub:ind
##           Df Sum Sq Mean Sq F value Pr(>F)
## ind         2  12136    6068 212.201 <2e-16 ***
## gr1:ind      6    325     54   1.892 0.0821 .
## gr2:ind      2    106     53   1.856 0.1582
## Residuals 280   8007     29
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$p = 0.02148 < 0.05$, следовательно, влияние фактора лекарств значимо для уровня депрессии,
 $p = 0.00242 < 0.05$, влияние фактора пола значимо для уровня депрессии.

средний уровень депрессии в группах соответственно равны:

```
FF<- function (name, krit) {
  temp <- dat.AR[krit ==name,];
  df<-rbind(data.frame(x = temp$BD1), data.frame(x = temp$BD2),
            data.frame(x = temp$BD5));
  mm<-mean(df$x);
  er<-sd(df$x)/sqrt(length(df$x));er;
  paste("for", name, mm, '±', er)
}
```

```

FF('male', dat.AR$SEX)

## [1] "for male 11.5922330097087 ± 0.487202479933983"
FF('female', dat.AR$SEX)

## [1] "for female 15.2539682539683 ± 0.96259810519184"
FF('NLTX+Framex', dat.AR$PRCOD)

## [1] "for NLTX+Framex 15.0900900900901 ± 0.906844440123583"
FF('NLTX+Placebo', dat.AR$PRCOD)

## [1] "for NLTX+Placebo 10.8405797101449 ± 0.691735864298655"
FF('Placebo+Framex', dat.AR$PRCOD)

## [1] "for Placebo+Framex 12.5061728395062 ± 1.06096907961457"
FF('Placebo+Placebo', dat.AR$PRCOD)

## [1] "for Placebo+Placebo 12.5714285714286 ± 0.998520622099373"

выводим средние
model.tables(ao, 'mean')

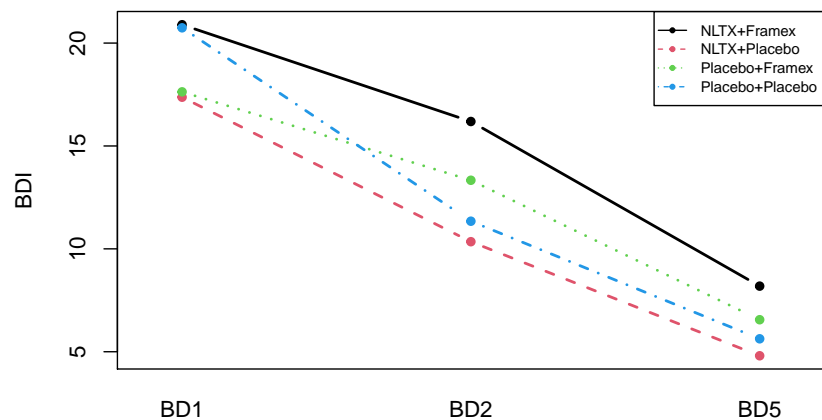
## Tables of means
## Grand mean
##
## 12.65287
##
## gr1
##      NLTX+Framex NLTX+Placebo Placebo+Framex Placebo+Placebo
##           15.09      10.84          12.51          12.57
## rep      111.00      138.00          81.00          105.00
##
## gr2
##      female  male
##           15.06  11.67
## rep  126.00 309.00
##
## ind
##      BD1  BD2  BD5
##      19.13 12.63 6.193
## rep 145.00 145.00 145.000
##
## gr1:ind
##
##      ind
## gr1      BD1  BD2  BD5
## NLTX+Framex 20.89 16.19 8.19
## rep      37.00 37.00 37.00
## NLTX+Placebo 17.37 10.35 4.80
## rep      46.00 46.00 46.00
## Placebo+Framex 17.63 13.33 6.56
## rep      27.00 27.00 27.00
## Placebo+Placebo 20.74 11.34 5.63
## rep      35.00 35.00 35.00

```

```
##
## gr2:ind
##      ind
## gr2    BD1    BD2    BD5
## female 22.50 14.98  7.70
## rep    42.00 42.00 42.00
## male   17.76 11.68  5.58
## rep    103.00 103.00 103.00

Names<-names(table(dat.AR[,1]));K<-length(Names)

interaction.plot(x.factor=dat.AR.T$ind,
trace.factor=dat.AR.T$gr1,
response=dat.AR.T$values,
fun = mean,
type = "b", legend = FALSE,
trace.label = "group",
xlab = "",
ylab = 'BDI',
lty = seq(K), col = seq(K), pch = 20, lwd=2
)
legend('topright',Names,lty = seq(K), col =seq(K), cex=0.7,pch=20)
```

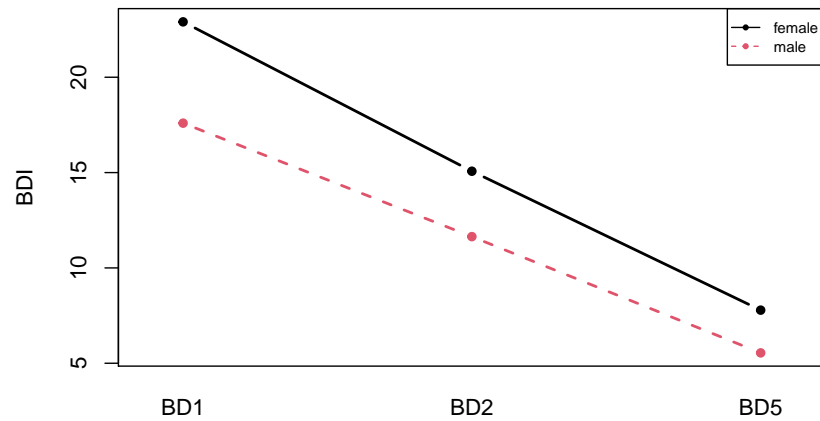


Присутствует некоторая взаимосвязь среди групп. Особенно среди групп, принимавших Placebo+FrameX и Placebo+Placebo.

```
Names<-names(table(dat.AR[,2]));K<-length(Names)

interaction.plot(x.factor=dat.AR.T$ind,
trace.factor=dat.AR.T$gr2,
response=dat.AR.T$values,
fun = mean,
type = "b", legend = FALSE,
trace.label = "group",
xlab = "",
```

```
ylab = 'BDI',
lty = seq(K), col = seq(K), pch = 20, lwd=2
)
legend('topright',Names,lty = seq(K), col =seq(K), cex=0.7,pch=20)
```



Две линии, соответствующие полам, проходят почти параллельно друг другу. Это означает, отсутствие эффекта взаимодействия факторов пола.