RUDN university

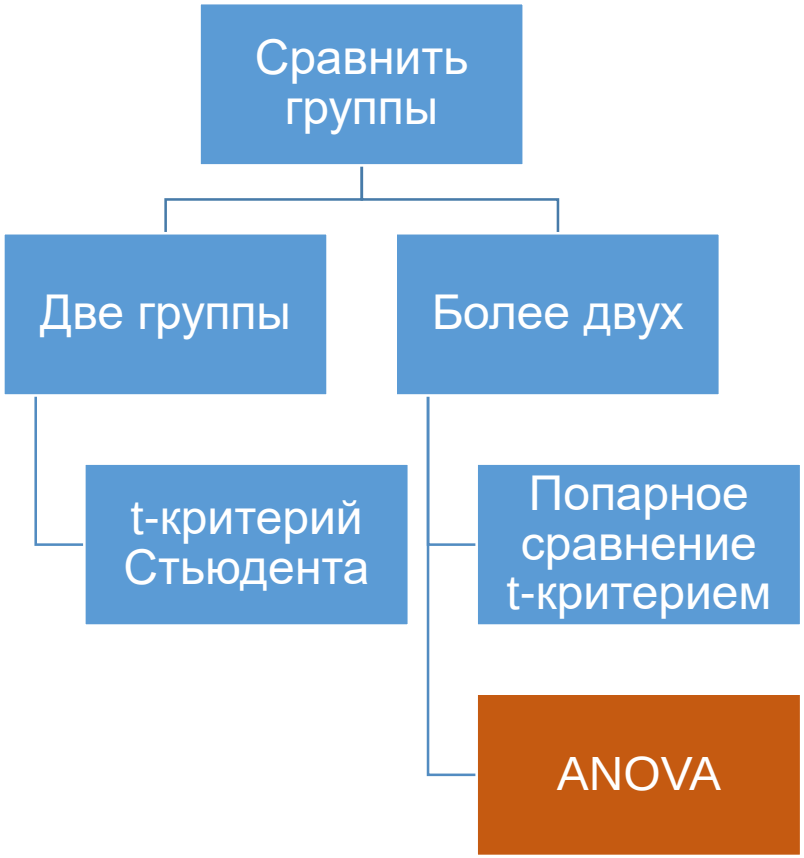
5100

Программирование в среде R

Шевцов Василий Викторович,
директор ДИТ РУДН, shevtsov_vv@rudn.university

Дисперсионный анализ для анализа нескольких групп и попарного сравнения

Дисперсионный анализ ANalysis Of VAriance (ANOVA)



Дисперсионный анализ — метод в математической статистике, направленный на поиск зависимостей в экспериментальных данных путём исследования значимости различий в средних значениях. В отличие от t-критерия, позволяет сравнивать средние значения трёх и более групп.

Суть дисперсионного анализа сводится к изучению влияния одной или нескольких независимых переменных (факторов), на зависимую переменную.

Требования:

- Количественный непрерывный тип данных, дискретные данные менее желательны.
- Независимые между собой выборки.
- Нормальное распределение признака в статистических совокупностях, из которых извлечены выборки.
- Равенство (гомогенность) дисперсий изучаемого признака в статистических совокупностях из которых извлечены выборки.
- Независимые наблюдения в каждой из выборок.

Определение статистических моделей, формулы

DV - dependent variable – зависимая переменная
IV - independent variable – независимая переменная

DV ~ IV1	На зависимую переменную оказывает влияние одна независимая переменная
DV ~ IV1 + IV2	На зависимую переменную оказывают влияние две независимых переменные
DV ~ IV1 : IV2	Влияние IV1 на DV зависит от уровня переменной IV2
DV ~ IV1 + IV2 + IV1 : IV2 DV ~ IV1 * IV2	На зависимую переменную оказывают влияние две независимых переменные, между независимыми переменными существует взаимосвязь (главный эффект + взаимодействие)
DV ~ (IV1 + IV2 + IV3)^2 DV ~ IV1 + IV2 + IV3 + IV1 : IV2 + IV2 : IV3 + IV1 : IV3	Основные эффекты + взаимодействие до 2-го уровня
DV ~ IV1 + Error(Subject/IV1)	Повторное измерение На зависимую переменную оказывает влияние межгруппова независимая переменная и внутригрупповая переменная (например, множество измерения одного и того же объекта)

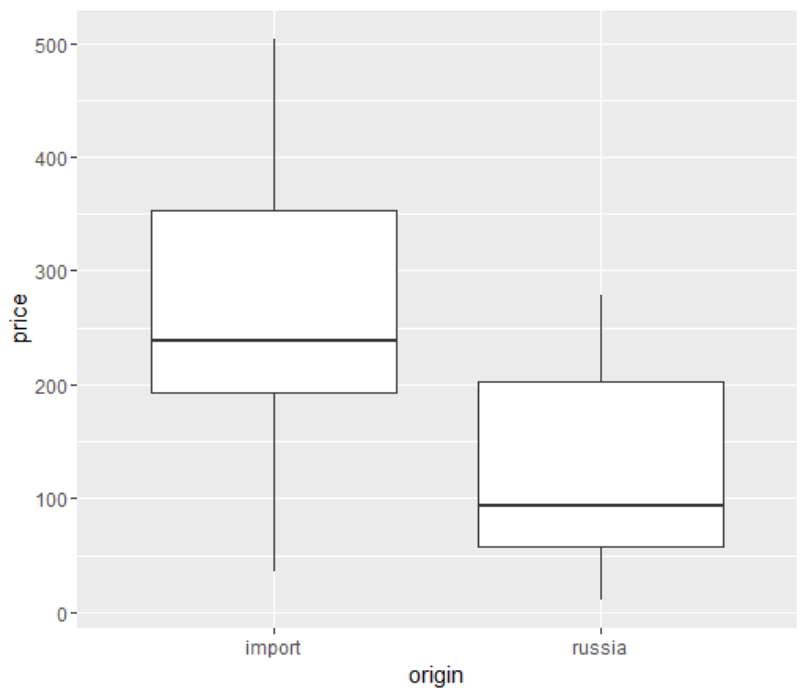
Пример

```
> df1 <- read.csv("Z:\\ФА\\2018-2019 уч.г\\R\\Занятие 14\\shops.csv")
> str(df1)
'data.frame':   20 obs. of  4 variables:
 $ food  : Factor w/ 5 levels "bread","cheese",..: 3 3 3 3 1 1 1 5 1 5 ...
 $ price : num  100.3 55.6 268.6 196.8 10.9 ...
 $ store : Factor w/ 2 levels "minimarket","supermarket": 2 1 1 2 1 2 2 1
 $ origin: Factor w/ 2 levels "import","russia": 2 2 1 1 2 2 1 2 1 2 ...
```

food	price	store	origin
chocolate	100.30	supermarket	russia
chocolate	55.57	minimarket	russia
chocolate	268.62	minimarket	import
chocolate	196.81	supermarket	import
bread	10.91	minimarket	russia
bread	25.84	supermarket	russia
bread	35.44	supermarket	import
vegetables	64.93	minimarket	russia
bread	116.23	minimarket	import
vegetables	226.39	supermarket	russia

```
> library(ggplot2)
> ggplot(df1,aes(x=origin,y=price)) +
+   geom_boxplot()
```

boxplot(price ~ origin, df1)



Однофакторный анализ

```
f1 <- aov(price ~ origin, data = df1)
```

```
> f1
Call:
   aov(formula = price ~ origin, data = df1)

Terms:
                  origin Residuals
Sum of Squares   94106.85 254729.45
Deg. of Freedom         1        18

Residual standard error: 118.9606
Estimated effects may be unbalanced
```

формула
сумма квадратов
степени свободы

Residuals характеризует внутригрупповую дисперсию (ее еще называют шумовой или остаточной дисперсией - в том смысле, что она не может быть объяснена влиянием экспериментального фактора)

Интерпретация результатов

```
> f1 <- aov(price ~ origin,data=df1)
> summary(f1)
             Df Sum Sq Mean Sq F value Pr(>F)
origin        1  94107   94107    6.65 0.0189 *
Residuals    18 254729   14152
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Вывод:
есть основания отвергнуть нулевую гипотезу об отсутствии различий между группами origin, т.е. стоимость групп товаров зависит от страны-производителя

Df	Степень свободы
Sum Sq	Сумма квадратов разброс наблюдений внутри групп
Mean Sq	Среднее квадратов разброс между группами (разброс групповых средних)
F value	Чем ближе F к 1, тем меньше оснований утверждать, что внутри- и межгрупповая дисперсии различаются. Иными словами, нет оснований отклонить сформулированную выше нулевую гипотезу. Если же F значительно выше 1, нулевую гипотезу можно отклонить. Если F-значение, рассчитанное по экспериментальным данным, превышает критическое значение, мы можем отклонить нулевую гипотезу об отсутствии эффекта изучаемого фактора.
Pr(>F)	Вероятность получить F-значение, равное или превышающее то значение, которое мы в действительности рассчитали по имеющимся выборочным данным (при условии, что нулевая гипотеза верна). Пороговое значение – 5%

Двухфакторный анализ

```
f1 <- aov(price ~ origin + store, data=df1)
```

```
> f1 <- aov(price ~ origin + store, data=df1)
> f1
Call:
   aov(formula = price ~ origin + store, data = df1)

Terms:
                  origin     store Residuals
Sum of Squares   94106.85   2980.95 251748.50
Deg. of Freedom         1         1        17

Residual standard error: 121.6911
Estimated effects may be unbalanced
```

```
> summary(f1)
            Df Sum Sq Mean Sq F value Pr(>F)
origin       1  94107   94107   6.355  0.022 *
store        1   2981    2981   0.201  0.659
Residuals   17 251749   14809
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Страна – производитель влияет на цену

Тип магазина не влияет на цену

model.tables()

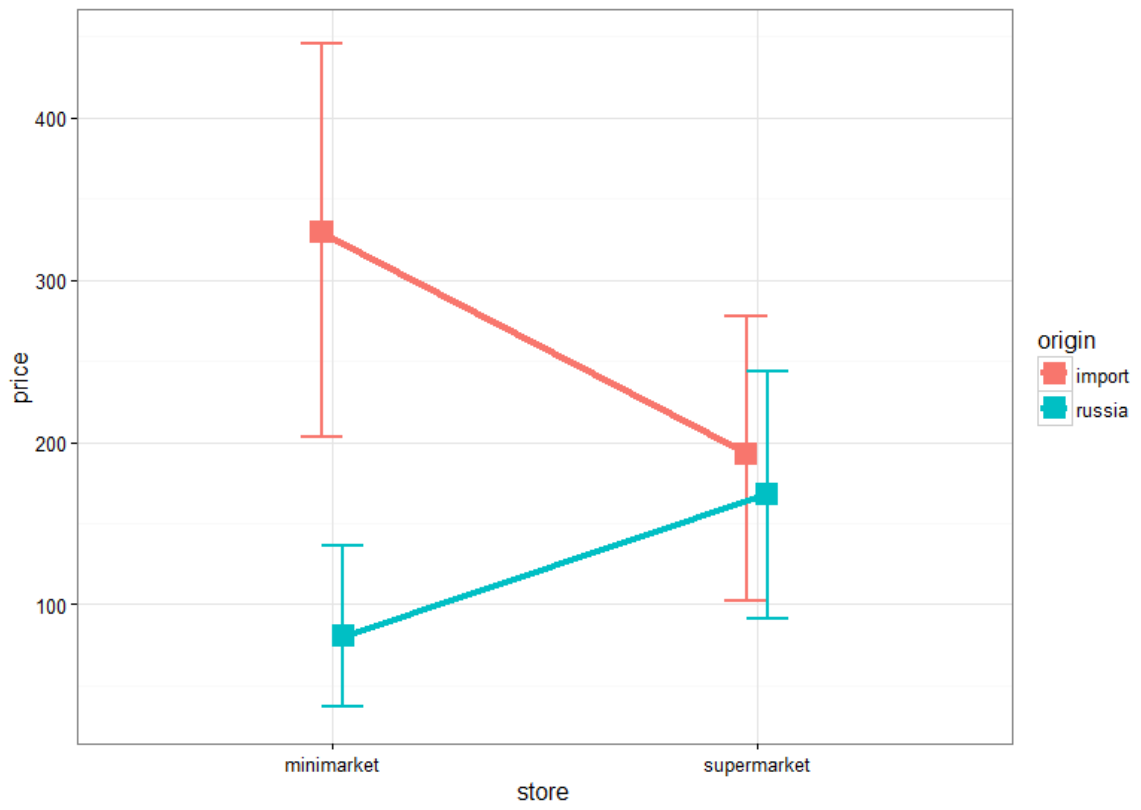
model.tables()
Compute Tables Of Results From An Aov Model Fit
x - a model object, usually produced by aov
type - type of table: currently only "effects" and "means" are implemented. Can be abbreviated.

```
> model.tables(f1,"mean")
Tables of means
Grand mean

192.7745

 origin
origin
import russia
261.37 124.18

 store
store
 minimarket supermarket
     204.98      180.57
```

Разница между средними ценами у типов товаров значительна,

между средними ценами у типов магазинов - нет

```
pd = position_dodge(0.1)
ggplot(df1, aes(x = store, y = price, color = origin, group = origin)) +
  stat_summary(fun.data = mean_cl_boot, geom = 'errorbar', width = 0.2, lwd = 0.8, position = pd)+
  stat_summary(fun.data = mean_cl_boot, geom = 'line', size = 1.5, position = pd) +
  stat_summary(fun.data = mean_cl_boot, geom = 'point', size = 5, position = pd, pch=15) +
  theme_bw()
```



В минимаркетах разброс цен по стране-производителю больше, чем в супермаркетах.

Отчетливо видно взаимодействие типа магазина и производителя

```
f1 <- aov(price ~ origin + store + origin : store, data=df1)
```

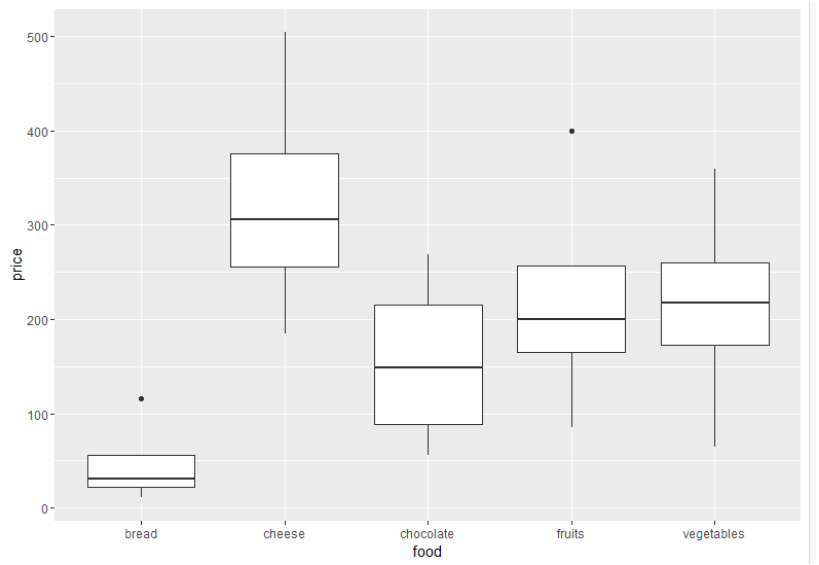
```
> f1 <- aov(price ~ origin + store + origin : store, data=df1)
> summary(f1)
             Df Sum Sq Mean Sq F value Pr(>F)
origin        1  94107   94107   7.968 0.0123 *
store         1   2981    2981   0.252 0.6222
origin:store  1  62777   62777   5.315 0.0349 *
Residuals    16 188971   11811
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Страна – производитель влияет на цену
Тип магазина не влияет на цену
Взаимодействие страны-производителя и типа магазина значимо

```
> f1 <- aov(price ~ origin * store, data=df1)
> summary(f1)
             Df Sum Sq Mean Sq F value Pr(>F)
origin        1  94107   94107   7.968 0.0123 *
store         1   2981    2981   0.252 0.6222
origin:store  1  62777   62777   5.315 0.0349 *
Residuals    16 188971   11811
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Многофакторный анализ

```
ggplot(df1, aes(x = food, y = price)) +
  geom_boxplot()
```




```
> f1 <- aov(price ~ food, data=df1)
> summary(f1)
            Df Sum Sq Mean Sq F value Pr(>F)
food         4 165823   41456   3.398 0.0362 *
Residuals   15 183013   12201
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Зависимость стоимости от типа продуктов является статистически значимой

TukeyHSD()

Реализует множественные сравнения групповых средних при помощи теста Тьюки.

Для выполнения большого числа попарных сравнений групповых средних без потери статистической мощности было разработано несколько методов, один из наиболее популярным является критерий Тьюки, или критерий достоверно значимой разности Тьюки.

```
> TukeyHSD(f1)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = price ~ food, data = df1)

$food
                          diff        lwr       upr     p adj
cheese-bread          278.0525   36.86938 519.23562 0.0204058
chocolate-bread       108.2200 -132.96312 349.40312 0.6453667
fruits-bread          174.3125  -66.87062 415.49562 0.2209202
vegetables-bread      167.7625  -73.42062 408.94562 0.2512881
chocolate-cheese     -169.8325 -411.01562  71.35062 0.2413687
fruits-cheese        -103.7400 -344.92312 137.44312 0.6789317
vegetables-cheese    -110.2900 -351.47312 130.89312 0.6297401
fruits-chocolate       66.0925 -175.09062 307.27562 0.9117335
vegetables-chocolate   59.5425 -181.64062 300.72562 0.9375222
vegetables-fruits      -6.5500 -247.73312 234.63312 0.9999874
```

Из всех пар сравнений можно выделить как статистически значимую cheese-bread

$p<0,05$

Дисперсионный анализ с повторным измерением

Цель его заключается в анализе различий между ответами одних и тех же респондентов (subject) на одни и те же вопросы в несколько приемов, то есть в течение ряда дискретных временных промежутков.

В качестве примера можно привести панельные исследования, когда одни и те же респонденты (потребители какого-либо продукта) отвечают на одни и те же вопросы через определенные интервалы времени (скажем, каждый квартал). Одной из основных целей дисперсионного анализа в рассматриваемом случае будет оценка влияния на ответы респондентов временного фактора. Таким образом, в частности, можно установить уровень лояльности к продуктам различных марок: если с течением времени средние оценки продукта марки X существенно не меняются/возрастают/убывают, следовательно, и отношение респондентов к данной марке сохраняется на прежнем уровне/улучшается/ухудшается. Иными словами, дисперсионный анализ с повторными измерениями может применяться для оценки значимости тенденций.

```
> str(df1)
'data.frame':   30 obs. of  5 variables:
 $ subject   : int  1 1 1 2 2 2 3 3 3 4 ...
 $ sex       : Factor w/ 2 levels "female","male": 1 1 1 1 1 1 2 2 2 2 ...
 $ therapy   : Factor w/ 3 levels "placebo","therapy1",..: 2 3 1 2 3 1 2 3 1 2 ...
 $ price     : Factor w/ 2 levels "high","low": 2 2 2 2 2 2 2 2 2 2 ...
 $ well_being: num  54.3 62.6 28.1 35.9 28.1 ...
```

well_being – зависимая переменная, показывает состояние здоровья

df1$subject <- as.factor(df1$subject)

subject	sex	therapy	price	well_being
1	female	therapy1	low	54.29100
1	female	therapy2	low	62.55823
1	female	placebo	low	28.13934
2	female	therapy1	low	35.90391
2	female	therapy2	low	28.12927
2	female	placebo	low	40.74942
3	male	therapy1	low	67.90081
3	male	therapy2	low	35.97446
3	male	placebo	low	42.59402
4	male	therapy1	low	28.46630
4	male	therapy2	low	42.94644
4	male	placebo	low	50.33725

Учет дисперсии испытуемого

```
f1 <- aov(well_being ~ therapy, data = df1)
```

```
> summary(f1)
            Df Sum Sq Mean Sq F value Pr(>F)
therapy      2    427   213.4   0.668  0.521
Residuals   27   8625   319.4
```

терапия, в т.ч. плацебо, не оказывает влияния на здоровье

```
f2 <- aov(well_being ~ therapy + Error(subject/therapy),
data = df1)
```

```
> summary(f2)

Error: subject
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  4    860     215

Error: subject:therapy
          Df Sum Sq Mean Sq F value Pr(>F)
therapy    2  426.8   213.4   0.619  0.563
Residuals  8 2760.3   345.0

Error: Within
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 15   5005   333.6
```

расчет с учетом множественности измерений по каждому объекту наблюдений с расчетом внутригрупповой ошибки

терапия, в т.ч. плацебо, не оказывает влияния на здоровье

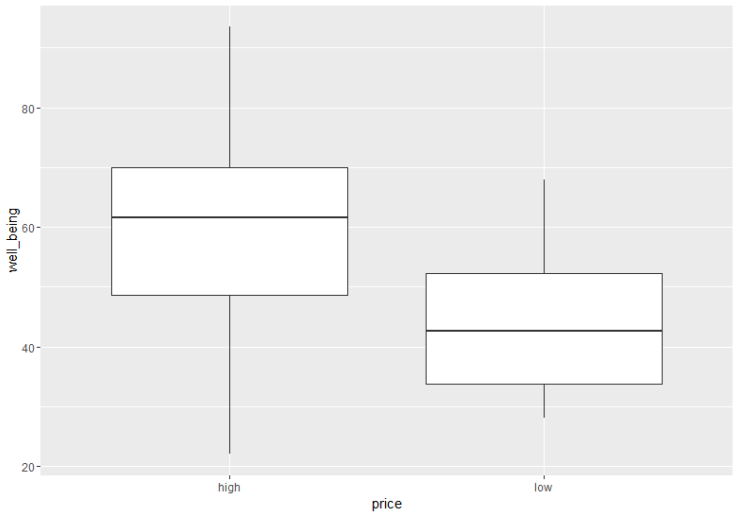
Второй фактор

```
f2 <- aov(well_being ~ therapy*price, data = df1)
```

```
> f2 <- aov(well_being ~ therapy*price, data = df1)
> summary(f2)
             Df Sum Sq Mean Sq F value Pr(>F)
therapy       2    427   213.4   0.743 0.4863
price         1   1675  1674.8   5.831 0.0237 *
therapy:price 2     57    28.6   0.100 0.9057
Residuals    24   6893   287.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
ggplot(df1, aes(x = price, y = well_being)) +
  geom_boxplot()
```

При более высокой цене состояние здоровья выше



С учетом дисперсии испытуемых

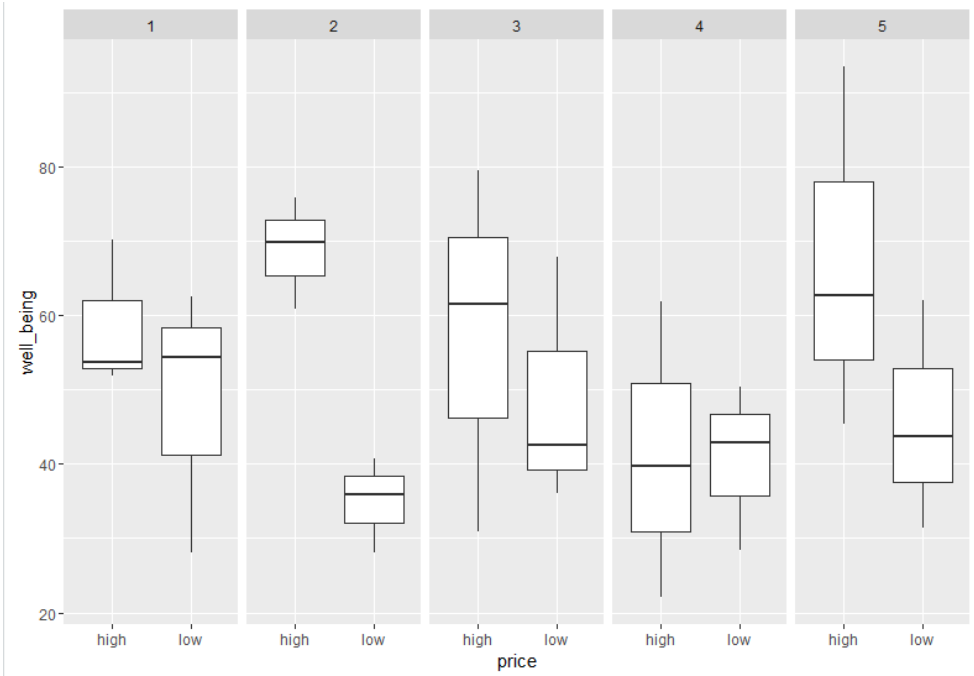
```
f3 <- aov(well_being ~ therapy*price +
       Error(subject/(therapy*price)),
       data = df1)
```

```
> summary(f3)

Error: subject
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  4    860     215

Error: subject:therapy
          Df Sum Sq Mean Sq F value Pr(>F)
therapy    2  426.8   213.4   0.619  0.563
Residuals  8 2760.3   345.0

Error: subject:price
          Df Sum Sq Mean Sq F value Pr(>F)
price      1   1675  1674.8   6.667 0.0612 .
Residuals  4   1005   251.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Error: subject:therapy:price
              Df Sum Sq Mean Sq F value Pr(>F)
therapy:price  2   57.2   28.58   0.101  0.905
Residuals      8 2267.8  283.48
```

При учете дисперсии по испытуемым зависимость состояние здоровья от цены уже не является столь значимым

Разбивка по испытуемым

```
ggplot(df1, aes(x = price, y = well_being)) +
  geom_boxplot() +
  facet_grid(~subject)
```



Не у всех испытуемые есть зависимость состояния здоровья от цены

Добавление независимого фактора

```
f4 <- aov(well_being ~ therapy*price*sex, data = df1)
```

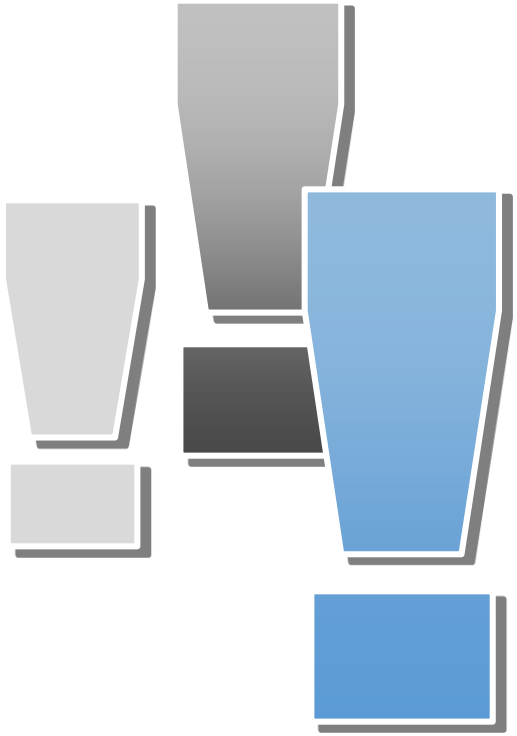
```
> summary(f4)
                Df Sum Sq Mean Sq F value Pr(>F)
therapy          2    427   213.4   0.609 0.5548
price            1   1675  1674.8   4.778 0.0423 *
sex              1     46    45.8   0.131 0.7219
therapy:price    2     57    28.6   0.082 0.9220
therapy:sex      2     70    35.1   0.100 0.9053
price:sex        1    255   255.3   0.728 0.4046
therapy:price:sex  2    212   106.1   0.303 0.7426
Residuals       18   6310   350.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
f4 <- aov(well_being ~
        therapy*price*sex +
        Error(subject/(therapy*price)),
        data = df1)
```

В итоге нельзя выделить какие-либо значимые факторы, влияющие на состояние здоровья испытуемого

```
> summary(f4)

Error: subject
          Df Sum Sq Mean Sq F value Pr(>F)
sex        1   45.8   45.82   0.169  0.709
Residuals  3  814.2  271.40

Error: subject:therapy
           Df Sum Sq Mean Sq F value Pr(>F)
therapy     2  426.8   213.4   0.476  0.643
therapy:sex 2   70.1    35.1   0.078  0.926
Residuals   6 2690.2   448.4

Error: subject:price
          Df Sum Sq Mean Sq F value Pr(>F)
price      1 1674.8  1674.8   6.704 0.0811 .
price:sex  1  255.3   255.3   1.022 0.3865
Residuals  3  749.5   249.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Error: subject:therapy:price
                 Df Sum Sq Mean Sq F value Pr(>F)
therapy:price     2   57.2    28.6   0.083  0.921
therapy:price:sex 2  212.2   106.1   0.310  0.745
Residuals         6 2055.7   342.6
```

Спасибо за внимание!



Шевцов Василий Викторович

shevtsov_vv@rudn.university
+7(903)144-53-57