

Липецкий государственный технический университет

Кафедра прикладной математики

КОМПЬЮТЕРНЫЕ ТЕХНОЛОГИИ МАТЕМАТИЧЕСКИХ ИССЛЕДОВАНИЙ

Лекция 7

7. Дисперсионный анализ

Составитель - Сысоев А.С., к.т.н., доц.

Липецк – 2021

Outline

7.1. Основные понятия

7.2. Подгонка ANOVA-моделей

7.3. Однофакторный дисперсионный анализ

7.3.1. Множественные сравнения

7.3.2. Проверка справедливости допущений, лежащих в основе теста

7.4. Двухфакторный дисперсионный анализ

7.5. Многомерный дисперсионный анализ

7.1. Основные понятия

Задача: лечение состояния тревоги

Способ лечения	
КТ	ДПДГ
п1	п6
п2	п7
п3	п8
п4	п9
п5	п10

Межгрупповой фактор

Зависимая переменная

Независимая переменная

План эксперимента

Сбалансированный эксперимент

Несбалансированный эксперимент

При наличии одной классифицирующей переменной - **однофакторный дисперсионный анализ (one-way between-groups ANOVA)**.

Внутригрупповой фактор -> **однофакторный дисперсионный анализ для зависимых наблюдений (one-way within-groups ANOVA)**.

При измерении каждого объекта более одного раза -> **дисперсионный анализ для повторных измерений (repeated measures ANOVA)**.

7.1. Основные понятия

		Пациент	Время	
			5 недель	6 месяцев
Способ лечения	КТ	п1		
		п2		
		п3		
		п4		
		п5		
	ДПДГ	п6		
		п7		
		п8		
		п9		
		п10		

Исследование сочетания нескольких факторов называется **многофакторным дизайном (планом) дисперсионного анализа (factorial ANOVA design)**.

Когда многофакторный дизайн включает и межгрупповые, и внутригрупповые факторы, он называется **смешанной моделью дисперсионного анализа (mixed-model ANOVA)**.

Поскольку уровень депрессии также может объяснять межгрупповые различия зависимой переменной, это смешивающий фактор (confounding factor).

Мешающая переменная (nuisance variable).

7.2. Подгонка ANOVA-моделей

Функция `aov()`

`aov(формула, data=таблица_данных)`

Символ	Использование
~	Разделяет зависимые переменные в левой части уравнения и независимые – в правой. Например, предсказание значений y по значениям A , B и C будет закодировано в виде $y \sim A + B + C$
+	Разделяет независимые переменные
:	Обозначает взаимодействие между переменными. Предсказание значений y по значениям A , B и взаимодействия между A и B будет закодировано в виде $y \sim A + B + A:B$
*	Обозначает все возможные сочетания переменных. Выражение $y \sim A*B*C$ расшифровывается как $y \sim A + B + C + A:B + A:C + B:C + A:B:C$
^	Обозначает сочетание определенного числа переменных. Выражение $y \sim (A+B+C)^2$ расшифровывается как $y \sim A + B + C + A:B + A:C + A:B$
.	Символ-заполнитель для всех остальных переменных в таблице данных, за исключением зависимой переменной. Например, если таблица данных содержит переменные y , A , B и C , то выражение $y \sim .$ расшифровывается как $y \sim A + B + C$

7.2. Подгонка ANOVA-моделей

Функция `aov()`

Дизайн	Формула
Однофакторный дисперсионный анализ	$y \sim A$
Однофакторный ковариационный анализ с одной ковариатой	$y \sim x + A$
Двухфакторный дисперсионный анализ	$y \sim A * B$
Двухфакторный ковариационный анализ с двумя ковариатами	$y \sim x1 + x2 + A * B$
Случайный блочный	$y \sim B + A$ (где B – определяющий блоки фактор)
Однофакторный дисперсионный анализ для зависимых переменных	$y \sim A + \text{Error}(\text{Объект}/A)$
Дисперсионный анализ с повторными измерениями с одним внутригрупповым фактором (W) и одним межгрупповым фактором (B)	$y \sim B * W + \text{Error}(\text{Объект}/W)$

7.2. Подгонка ANOVA-моделей

Порядок членов в формуле важен, если есть больше одного фактора и эксперимент несбалансирован.

При двухфакторном дисперсионном анализе с неравным числом наблюдений для разных комбинаций факторов модель $y \sim A*B$ не даст того же результата, что и модель $y \sim B*A$.

$$Y \sim A + B + A:B$$

Существует три основных подхода для распределения дисперсии y между эффектами, перечисленными в правой части уравнения.

Тип I (последовательный). Эффекты масштабируются по эффектам, которые указаны в формуле раньше. A не масштабируется, B масштабируется по A . Взаимодействие $A:B$ масштабируется по A и B .

Тип II (иерархический). Эффекты масштабируются по тем эффектам, которые имеют такой же или более низкий уровень. A масштабируется по B , B масштабируется по A . Взаимодействие $A:B$ масштабируется и по A , и по B .

Тип III (краевой). Каждый эффект масштабируется по любому другому эффекту в модели. A масштабируется по B и по $A:B$. B масштабируется по A и $A:B$. Взаимодействие $A:B$ масштабируется по A и B . В R по умолчанию реализуется тип I.

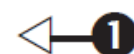
Более важные эффекты должны стоять в формуле раньше.

7.3. Однофакторный дисперсионный анализ

Задача: Пятьдесят пациентов проходили один из пяти снижающих уровень холестерина курсов лечения (переменная `trt`). Три из этих курсов заключались в использовании одного и того же препарата в количестве 20 мг один раз в день (`1time`), в количестве 10 мг дважды в день (`2times`) или 5 мг четыре раза в день (`4times`). Два других курса лечения заключались в приеме альтернативных препаратов (`drugD` и `drugE`). Какой из курсов лечения привел к наиболее заметному снижению уровня холестерина (переменная `response`)?

```
> library(multcomp)
> attach(cholesterol)
> table(trt)
trt
```

```
 1time 2times 4times  drugD  drugE
    10     10     10     10     10
```



1 Объем выборки
в каждой группе

```
> aggregate(response, by=list(trt), FUN=mean)
```

```
  Group.1      x
1  1time  5.78
2  2times  9.22
3  4times 12.37
4  drugD 15.36
5  drugE 20.95
```



2 Средние
значения в
каждой группе

7.3. Однофакторный дисперсионный анализ

```
> aggregate(response, by=list(trt), FUN=sd)
```

← 3 Стандартные отклонения в каждой группе

```
  Group.1      x
1  1time 2.88
2  2times 3.48
3  4times 2.92
4  drugD 3.45
5  drugE 3.35
```

```
> fit <- aov(response ~ trt)
```

← 4 Тест на межгрупповые различия (ANOVA)

Графическое изображение средних значений с доверительными интервалами для каждой группы

```
> summary(fit)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
trt	4	1351	338	32.4	9.8e-13 ***
Residuals	45	469	10		

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

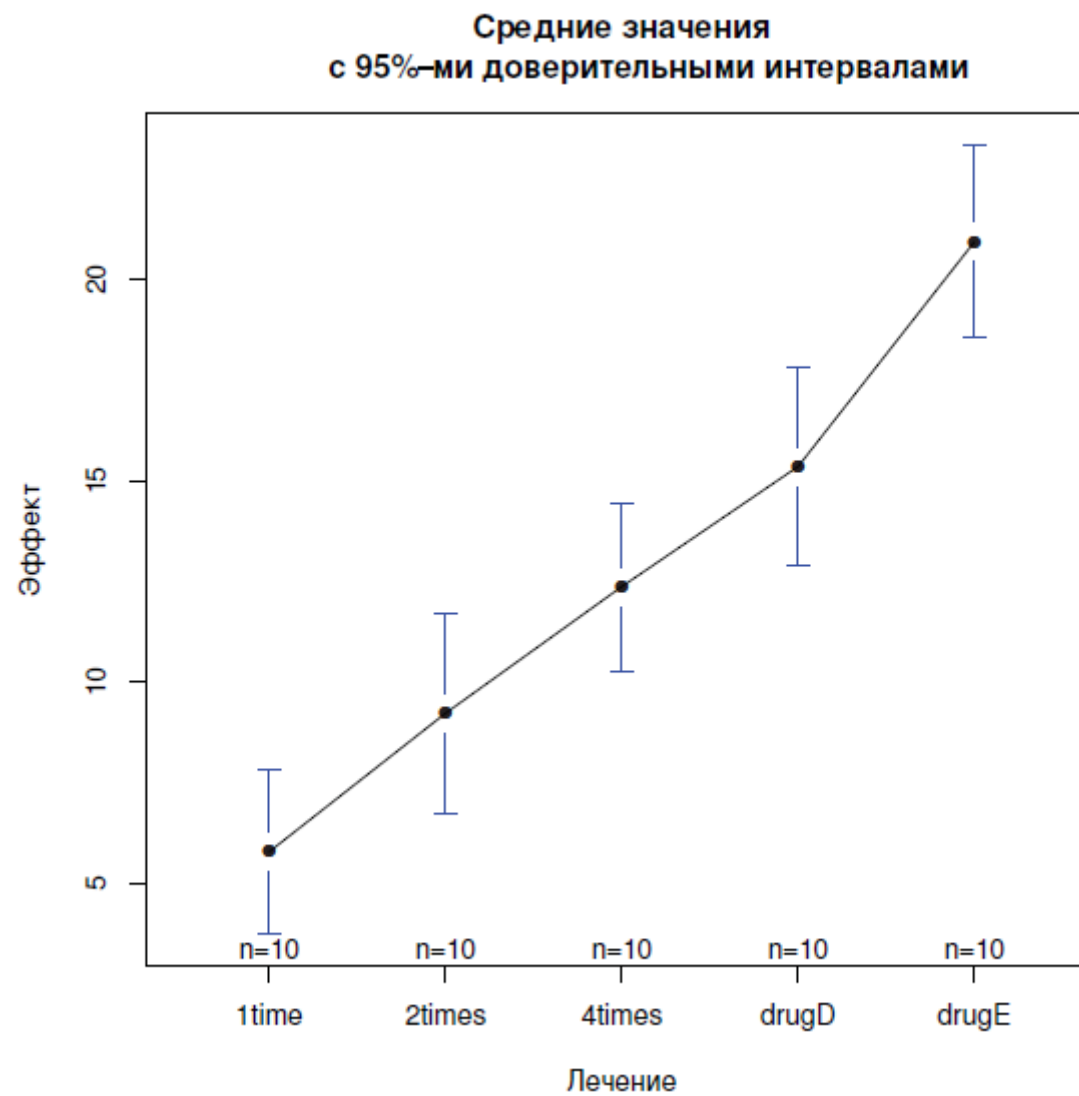
```
> library(gplots)
```

```
> plotmeans(response ~ trt, xlab="Лечение", ylab="Эффект",  
  main="Средние значения с 95%-ми доверительными интервалами")
```

← 5

```
> detach(cholesterol)
```

7.3. Однофакторный дисперсионный анализ



7.3. Однофакторный дисперсионный анализ

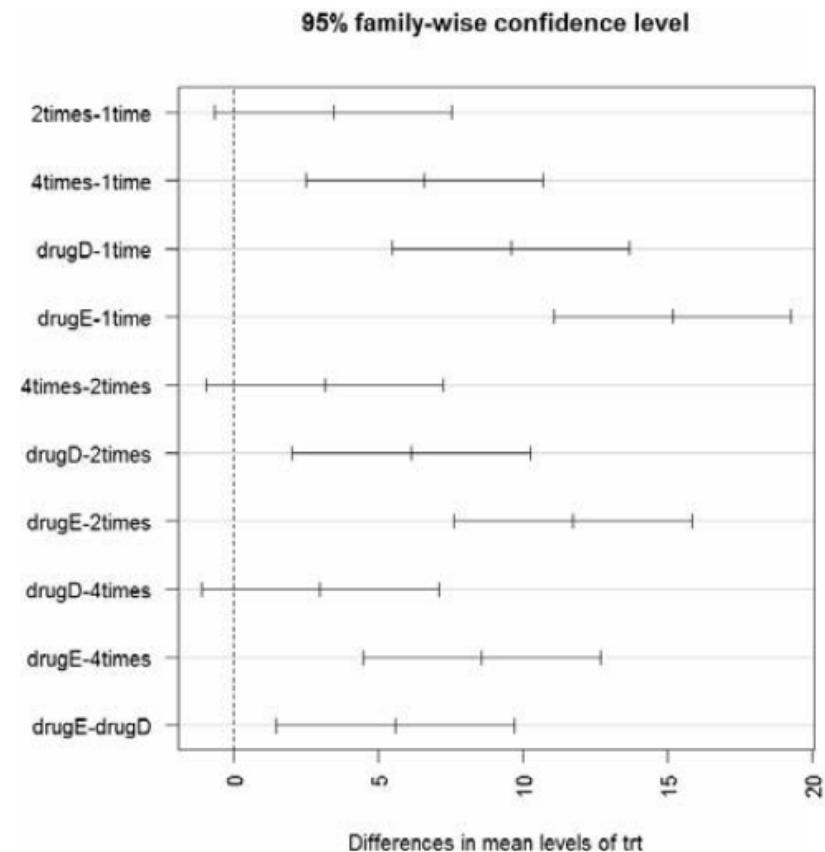
7.3.1. Множественные сравнения

Задача: Какие именно способы лечения различаются между собой?

```
> TukeyHSD(fit)
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = response ~ trt)
$trt

      diff      lwr      upr p adj
2times-1time  3.44 -0.658  7.54 0.138
4times-1time  6.59  2.492 10.69 0.000
drugD-1time   9.58  5.478 13.68 0.000
drugE-1time  15.17 11.064 19.27 0.000
4times-2times  3.15 -0.951  7.25 0.205
drugD-2times   6.14  2.035 10.24 0.001
drugE-2times  11.72  7.621 15.82 0.000
drugD-4times   2.99 -1.115  7.09 0.251
drugE-4times   8.57  4.471 12.67 0.000
drugE-drugD    5.59  1.485  9.69 0.003
```

```
> par(las=2)
> par(mar=c(5, 8, 4, 2))
> plot(TukeyHSD(fit))
```

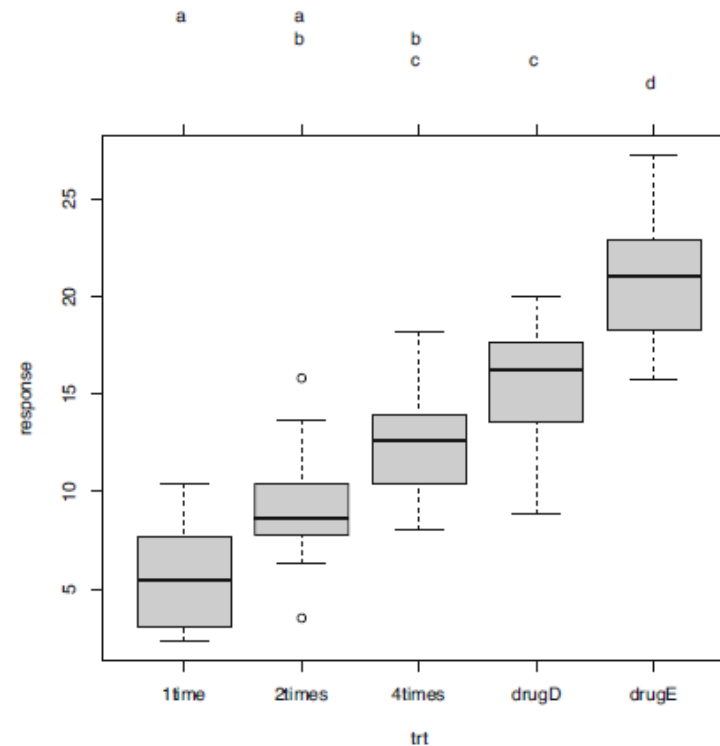


7.3. Однофакторный дисперсионный анализ

7.3.1. Множественные сравнения

Тест честных оценок достоверности различий Тьюки (Tukey honest significant differences test)

```
> library(multcomp)
> par(mar=c(5, 4, 6, 2))
> tuk <- glht(fit, linfct=mcp(trt="Tukey"))
> plot(cld(tuk, level=.05), col="lightgrey")
```



7.3. Однофакторный дисперсионный анализ

7.3.2. Проверка справедливости допущений, лежащих в основе теста

В случае однофакторного дисперсионного анализа предполагается, что значения зависимой переменной распределены нормально и имеют одинаковую дисперсию в каждой группе.

```
> library(car)
> qqPlot(lm(response ~ trt, data=cholesterol),
          simulate=TRUE, main="Q-Q Plot", labels=FALSE)
```

Тест Барлетта (Bartlett's test)

```
> bartlett.test(response ~ trt, data=cholesterol)
Bartlett test of homogeneity of variances
data: response by trt
Bartlett's K-squared = 0.5797, df = 4, p-value = 0.9653
```

Проверка данных на наличие выбросов

```
> library(car)
> outlierTest(fit)
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
      rstudent unadjusted p-value Bonferonni p
19 2.251149      0.029422      NA
```

7.4. Двухфакторный дисперсионный анализ

Задача: Шестьдесят морских свинок были случайно распределены по группам, получавшим три разных количества аскорбиновой кислоты (dose: 0.5, 1 или 2 мг) двумя способами (supp: апельсиновый сок – OJ или витамин C – VC) – по 10 свинок в каждой из шести групп. Зависимая переменная – это длина зубов (len).

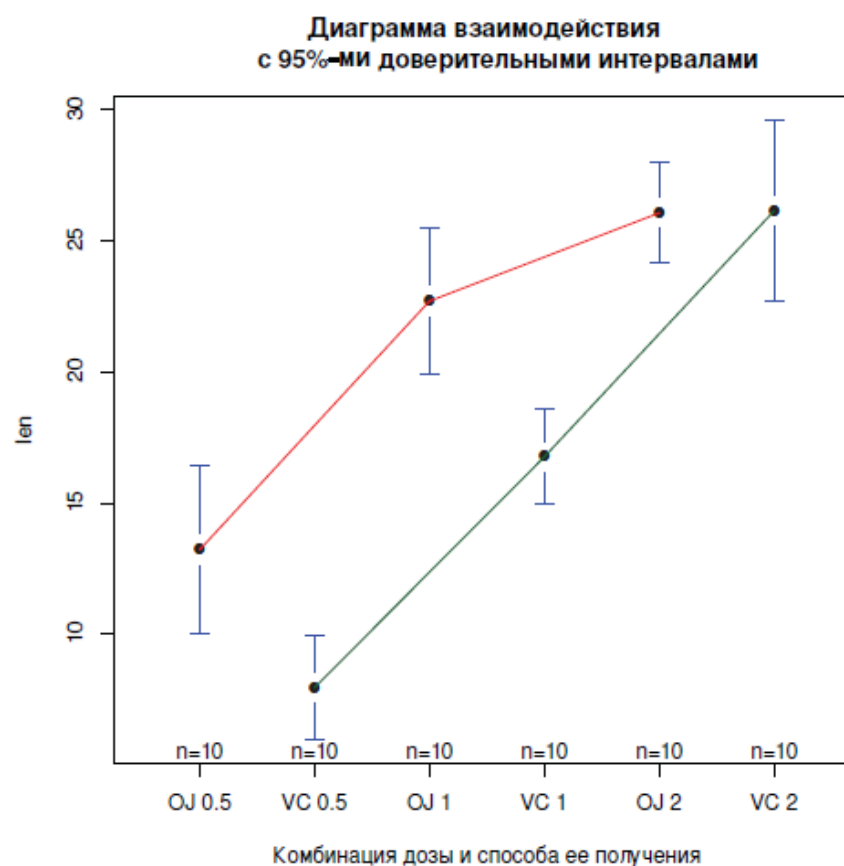
```
> attach(ToothGrowth)
> table(supp, dose)
      dose
supp 0.5  1  2
  OJ  10 10 10
  VC  10 10 10
> aggregate(len, by=list(supp, dose), FUN=mean)
> aggregate(len, by=list(supp, dose), FUN=sd)
> fit <- aov(len ~ supp*dose)
> summary(fit)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
supp	1	205	205	12.32	0.0009	***
dose	1	2224	2224	133.42	<2e-16	***
supp:dose	1	89	89	5.33	0.0246	*
Residuals	56	934	17			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

7.4. Двухфакторный дисперсионный анализ

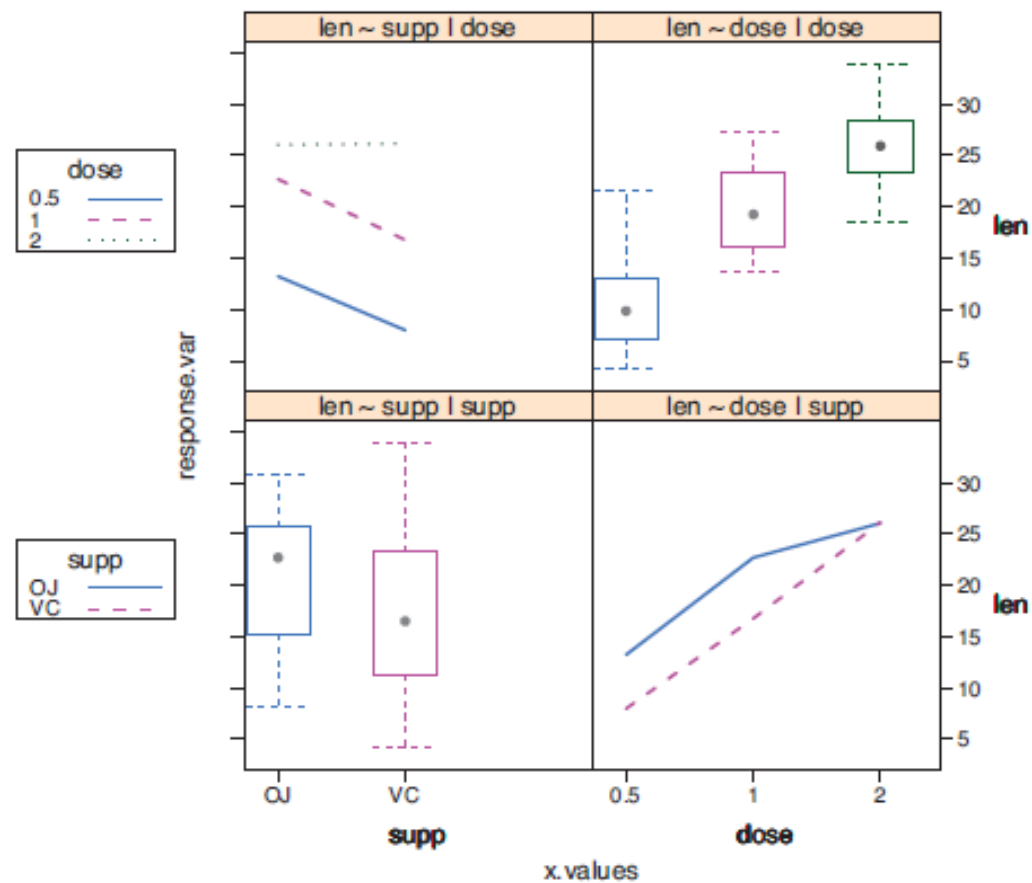
```
library(gplots)
plotmeans(len ~ interaction(supp, dose, sep=" "),
          connect=list(c(1,3,5),c(2,4,6)),
          col=c("red", "darkgreen"),
          main = "Диаграмма взаимодействия с 95%-ми доверительными  
интервалами",
          xlab= "Комбинация дозы и способа ее получения")
```



7.4. Двухфакторный дисперсионный анализ

```
library(HH)  
interaction2wt(len~supp*dose)
```

len: main effects and 2-way interactions



7.5. Многомерный дисперсионный анализ

Задача: Как зависит содержание калорий (calories), жиров (fat) и углеводов (sugars) в американских сухих завтраках от того, на какой полке в магазине они стоят (shelf: 1 – нижняя полка, 2 – средняя, 3 – верхняя). Количество калорий, жиров и углеводов – это зависимые переменные, а номер полки – независимая переменная с тремя уровнями (1, 2 и 3).

```
> library(MASS)
> attach(UScereal)
> y <- cbind(calories, fat, sugars)
> aggregate(y, by=list(shelf), FUN=mean)
  Group.1 calories    fat sugars
1      1      119 0.662    6.3
2      2      130 1.341   12.5
3      3      180 1.945   10.9
> cov(y)
      calories    fat sugars
calories 3895.2 60.67 180.38
fat       60.7  2.71   4.00
sugars    180.4  4.00  34.05
> fit <- manova(y ~ shelf)
> summary(fit)
      Df Pillai approx F num Df den Df  Pr(>F)
shelf   1 0.1959   4.9550     3    61 0.00383 **
Residuals 63
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

7.5. Многомерный дисперсионный анализ

```
> summary.aov(fit)                                     ← Вывод результатов для каждого признака
Response calories :
      Df Sum Sq Mean Sq F value    Pr(>F)
shelf   1  45313   45313  13.995 0.0003983 ***
Residuals 63 203982    3238
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response fat :
      Df Sum Sq Mean Sq F value    Pr(>F)
shelf   1  18.421   18.421   7.476 0.008108 **
Residuals 63 155.236    2.464
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response sugars :
      Df Sum Sq Mean Sq F value    Pr(>F)
shelf   1  183.34   183.34   5.787 0.01909 *
Residuals 63 1995.87    31.68
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Функция `manova()` осуществляет многомерный тест на межгрупповые различия. Статистически значимая величина F-статистики свидетельствует о том, что наши три группы злаков различаются по питательной ценности.

Библиографический список

Кабаков Р. К. (2014) R в действии. Анализ и визуализация данных на языке R Издательство: ДМК Пресс, 580 с.

Шипунов А. Б., Балдин Е. М., Волкова П. А., Коробейников А. И., Назарова С. А., Петров С. В., Суфиянов В. Г. (2012) Наглядная статистика. Используем R! - М.: ДМК Пресс, 298 с.