Санкт-Петербургский государственный университет

Кафедра математической теории игр и статистических решений

Семыкина Александра Андреевна

Выпускная квалификационная работа бакалавра

Применение многофакторного дисперсионного анализа в маркетинге

Направление 010400

Прикладная математика и информатика

Научный руководитель,
кандидат ф.-м. наук
доцент
Громова Е.В.

Санкт-Петербург
2016

Содержание

Введение

В условиях современной экономики для поддержания конкурентоспособности на высоком уровне, а также для грамотного ведения рыночной деятельности фирме необходимо проводить маркетинговые исследования для мониторинга её текущей деятельности.

Недостаток и недостоверность информации зачастую являются причиной неверных прогнозов, а значит, и принятия неэффективных управленческих решений, в то время как грамотный анализ данных в большинстве случаев позволяет избежать этой проблемы.

Одной из важнейших задач маркетинга и маркетингового исследования является установление причинно-следственных связей, выявление закономерностей бизнес-процесса. Многофакторный дисперсионный анализ служит инструментом исследования влияния набора факторов, являющихся качественными переменными, на зависимую количественную переменную (объем и частота покупок, размер дохода, потребительская оценка, рейтинг фирмы и др.). При этом, в роли качественных переменных могут выступать характеристики как потребителей (например пол, возраст, уровень дохода), так и самой фирмы (интенсивность и концепция рекламной кампании, варианты упаковки, географическое расположение). В качестве факторов могут быть рассмотрены и внешние влияния, такие как экономическая ситуация в стране, её климатические и культурные особенности, время года.

В данной работе рассматривается применение различных моделей многофакторного дисперсионного анализа. Объектом применения была выбрана конкретная фирма - аптечная сеть (по просьбе владельца фирмы анализируемые данные приводятся в обезличенном формате).

Необходимо также отметить, что область применения дисперсионного

анализа не ограничивается маркетинговыми исследованиями. Рассматриваемый метод широко используется в самых разных отраслях науки, в том числе в психологии, социологии, медицине, биологии и агрономии.

Постановка задачи

Целью работы являлось исследование особенностей применения различных моделей многофакторного дисперсионного анализа в маркетинге.

Для достижения поставленной цели были сформулированы следующие задачи:

1. Выбрать оптимальные модели для проведения исследования
2. Провести дисперсионный анализ по выбранным моделям
3. Провести интерпретацию полученных результатов
4. Информацию, полученную в ходе анализа, обобщить в выводе

По итогам исследования планировалось:

1. Выявить достоинства и недостатки исследуемого метода
2. Оценить потенциальную ценность информации, полученной в результате применения метода

Поставленные задачи предполагалось решить на примере данных из сети аптек. Прикладная задача может быть сформулирована так: «Провести исследование доходности сети аптек для случаев двух и трех факторов. Проверить гипотезу о наличии влияния времени года и категории товаров на размер дохода. Выяснить, различается ли размер дохода, приносимый разными филиалами».

Обзор литературы

Понятие дисперсионного анализа было впервые использовано в 1925 году британским статистиком Фишером Р. А. в его книге «Статистические методы для исследователей»[1]. С тех пор метод широко начал широко применяться в различных областях науки.

В отечественной литературе дисперсионный анализ наиболее полно описан в книге «Прикладная математическая статистика. Для инженеров и научных работников», автор Кобзарь А. И.[2]

В зарубежной литературе подробное описание метода представлено в следующих работах: Г. Шеффе «Дисперсионный анализ»[3], Г. Крамер «Математические методы статистики»[4].

В последние годы для проведения статистического анализа, в том числе и дисперсионного, все чаще используются различные приложения и статистические пакеты, такие как Excel, R, SPSS, Statistica и другие. В данном исследовании использовались такие пособия, как «Discovering Statistics Using R», A. Field[5], «Статистический анализ и визуализация данных с помощью R», С. Э. Мастицкий, В. К. Шитиков[6].

Количество публикаций на тему многофакторного дисперсионного анализа в области маркетинга на сегодняшний день остается высоким (20 900 - примерное количество публикаций на тему «ANOVA in marketing» с 2012 года по результатам поиска с помощью сервиса Google Scholar); это говорит о том, что метод остается востребованным и актуальным несмотря на то, что был введен достаточно давно. Из последних публикаций можно упомянуть статью «ANOVA in marketing research of consumer behavior of different categories in Georgian Market», Nugzar Todua, Teona Dotchviri[7]. Статья близка по тематике к данной работе, но исследование в ней проводится

по другой модели дисперсионного анализа.

Таким образом, подводя итог всего сказанного выше, можно заключить, что дисперсионный анализ имеет широкое применение в самых разных областях науки. А значит, использование этого метода является целесообразным. В данной работе будет рассмотрено приложение многофакторного дисперсионного анализа в маркетинге. Такой выбор области приложения обусловлен высоким уровнем востребованности маркетологами инструментов анализа и оценки текущей ситуации фирмы.

1 Многофакторный дисперсионный анализ

1.1. Принцип дисперсионного анализа

Дисперсионный анализ является статистическим методом анализа результатов наблюдений, зависящих от различных одновременно действующих факторов, с целью выбора наиболее значимых факторов и оценки их влияния на исследуемый процесс. С помощью дисперсионного анализа устанавливаются изменения дисперсии результатов эксперимента при изменении уровней изучаемого фактора. Если дисперсии будут отличаться значимо, то следует вывод о значимом влиянии фактора на среднее значение наблюдаемой случайной величины.

Нулевой гипотезой в дисперсионном анализе является утверждение о равенстве средних значений:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_j$$

Альтернативной гипотезой будет являться предположение о нарушении хотя бы одного из этих равенств. Пусть на случайную величину X воздействует некоторый качественный фактор F, имеющий p уровней, а количество наблюдений на каждом уровне фактора одинаково и равно q.

$\overline{x}$ - генеральное среднее значение всех наблюдений.

Введем обозначения:

$S_{total} = \sum_{j=1}^{p} \sum_{i=1}^{q} (x_{ij} - \overline{x})^2$ - общая сумма квадратов отклонений наблюдаемых значений от общего среднего;

$S_{BG} = q \sum_{i=1}^{q} (\overline{x}_i - \overline{x})^2$ - факторная (или межгрупповая, between-group) сумма квадратов отклонений групповых средних от общего среднего, характеризующая рассеяние между группами;

$S_{WG} = \sum_{j=1}^{p} \sum_{i=1}^{q} (x_{ij} - \overline{x}_j)^2$ - остаточная (или внутригрупповая, within-group)

сумма квадратов отклонений наблюдаемых значений группы от своего группового среднего, характеризующая рассеяние внутри групп, причем,

$$S_{total} = S_{BG} + S_{WG}.$$

Разделив суммы квадратов на соответствующее им число степеней свободы, получим общую, факторную и остаточную дисперсии:

$$MS_{total} = \frac{S_{total}}{N-1}, \qquad MS_{BG} = \frac{S_{BG}}{p-1}, \qquad MS_{WG} = \frac{S_{WG}}{N-p}.$$

Если справедлива гипотеза H_0, то экспериментальные группы являются случайными выборками из одной и той же генеральной совокупности, тогда факторная и остаточная дисперсии являются несмещенными оценками дисперсии этой совокупности, и, следовательно, различаются незначимо. Формально остаточная и факторная дисперсии сравниваются при помощи F-критерия, или критерия Фишера. Для этого по формуле $\frac{MS_{BG}}{MS_{WG}}$ вычисляется значение статистики. Критическое же значение F-критерия определяется желаемым уровнем значимости и свойствами F-распределения, форма которого полностью задается степенями свободы, соответствующими остаточной и факторной дисперсиям.[2]

При наличии нескольких факторов, аналогичные вычисления и проверка по критерию Фишера проводятся для каждого из них. Формулы для проведения двух- и трехфакторного дисперсионного анализа можно найти в приложении. Также, в этом случае, помимо основной нулевой гипотезы проводится проверка ещё одной гипотезы, согласно которой комбинация факторов не оказывает эффекта взаимодействия на значения зависимой переменной. Формальная запись этой гипотезы будет рассмотрена ниже.

1.2. Модель многофакторного дисперсионного анализа

Построим модель трехфакторного дисперсионного анализа. Пусть на случайную величину X воздействует факторы A, B, C, имеющие a, b и c уровней соответственно. Обозначим через y_{ijkt} результат t-го измерения, проведенного на уровне i фактора A, уровне j фактора B и уровне k фактора C. Модель будет иметь следующий вид [4]:

$$y_{ijkt} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkt}$$

где:

μ - глобальное среднее значение признака y;

μ_{ijk} - среднее значение наблюдений на пересечении уровней i, j и k факторов A, B и C соответственно;

$\alpha_i = \mu_{i..} - \mu$ - эффект уровня i фактора A, где $\mu_{i..}$ - среднее значение признака y на i-м уровне фактора A;

β_j и γ_k - эффекты уровня j фактора B и уровня k фактора C соответственно;

$(\alpha\beta)_{ij} = \mu_{ij.} - (\mu + \alpha_i + \beta_j)$ - эффект взаимодействия для комбинации уровня i фактора A и уровня j фактора B, где $\mu_{ij.}$ - среднее значение признака y на пересечении i-го уровня фактора A и j-го уровня фактора B;

Аналогично определяются $(\beta\gamma)_{jk}$ и $(\alpha\gamma)_{ik}$;

$(\alpha\beta\gamma)_{ijk} = \mu_{ijk} - (\mu + (\alpha\beta)_{ij} + (\beta\gamma)_{jk} + (\alpha\gamma)_{ik} + \alpha_i + \beta_j + \gamma_k)$ - эффект взаимодействия для комбинации уровня i фактора A, уровня j фактора B и уровня k фактор C;

ε_{ijkt} - случайная ошибка t-го измерения на пересечении уровней i, j и k факторов A, B и C соответственно;

Нулевые гипотезы можно записать следующим образом:

$H_{0_A} : \alpha_i = 0, \forall i$

$H_{0_B} : \beta_j = 0, \forall j$

$H_{0_C} : \gamma_k = 0, \forall k$

Каждая из сформулированных гипотез эквивалента гипотезе о равенстве средних уровней фактора, сформулированной в предыдущем параграфе.

Так как на зависимую переменную действует больше, чем один фактор, добавляются нулевые гипотезы о наличие эффекта взаимодействия факторов:

$H_{0_{AB}} : (\alpha\beta)_{ij} = 0, \forall i, j$

$H_{0_{BC}} : (\beta\gamma)_{jk} = 0, \forall j, k$

$H_{0_{AC}} : (\alpha\gamma)_{ik} = 0, \forall i, k$

$H_{0_{ABC}} : (\alpha\beta\gamma)_{ijk} = 0, \forall i, j, k$[3]

1.3. Основные предположения дисперсионного анализа

Классические методы дисперсионного анализа основываются на следующих предпосылках[2]:

- Все выборки носят случайный и независимый характер
- Распределение исходных случайных величин нормально
- Дисперсии экспериментальных данных одинаковы на различных уровнях изучаемого фактора (условие гомоскедастичности)

1.3.1. Проверка условия нормальности распределения генеральной совокупности

В данной работе было принято решение об использовании критерия Шапиро-Уилка для проверки гипотезы о нормальности распределения генеральной совокупности. Этот критерий был выбран, так как изучение его мощности показало, что он является одним из наиболее эффективных критериев проверки нормальности распределения случайных величин[2]. В ка-

честве недостатков этого метода можно упомянуть его смещенность при малых объемах выборок по отношению к альтернативам, более плосковершинным по сравнению с нормальным законом[8]. Но так как в этом исследовании объем выборок достаточно велик ($n = 100$), критерий Шапиро-Уилка можно считать оптимальным и принять в качестве основного инструмента проверки гипотезы нормальности распределения генеральной совокупности.

Тест основан на отношении оптимальной линейной несмещенной оценки дисперсии к её обычной оценке методом максимального правдоподобия. Статистика критерия имеет вид

$$W = \frac{1}{s^2}[\sum_{i=1}^{k} a_{n-i+1}(x_{n-i+1} - x_i)]^2,$$

где $s^2 = \sum_{i=1}^{n}(x_i - \bar{x})$; $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Коэффициенты a_{n-i+1} можно найти в таблице коэффициентов критерия Шапиро-Уилка. Если $W < W(\alpha)$, то нулевая гипотеза нормальности распределения отклоняется на уровне значимости α, критические значения $W(\alpha)$ можно найти в таблице процентных точек критерия $W(\alpha)$.[1]

1.3.2. Проверка условия гомоскедастичности

Условие гомоскедастичности может быть проверено несколькими способами, включающими критерии Хартли, Кохрана, Левене, Флигнера-Киллина и Бартлетта. Некоторые из этих критериев являются слишком чувствительными к нарушению условия нормальности (критерий Бартлетта), критерий Флигнера-Киллина хоть и является непараметрическим, но предполагает равенство медиан тестируемых выборок. Кроме того, непараметрические критерии значительно уступают в мощности параметрическим. Исследования[9] показали, что критерий Кохрана является самым мощным из перечисленных

критериев, сохраняя это свойство и при нарушении условия нормальности. Поэтому выбор был сделан в пользу критерия Кохрана.

Нулевая гипотеза для m выборок может быть записана так:

$H_0 : \sigma_1 = \sigma_2 = \ldots = \sigma_m$.

Альтернативной является гипотеза о нарушении хотя бы одного из этих равенств. Статистика этого критерия выражается формулой

$$Q = \frac{S_{max}}{S_1^2 + S_2^2 + \ldots + S_m^2},$$

где $S_{max} = max(S_1^2, S_2^2, \ldots, S_m^2)$, m - число выборок, S_i^2 - оценки выборочных дисперсий. Критическое значение критерия может быть вычислено следующим образом:

$$C_{UL}(\alpha, n, m) = [1 + \frac{m-1}{F_c(\alpha/m, (n-1), (m-1)(n-1))}]^{-1},$$

где n - количество наблюдений в каждой выборке, F_c критическое значение распределения Фишера. Если $C_j > C_{UL}$ хотя бы для одного j, то нулевая гипотеза отклоняется.[10]

1.4. Метод контрастов

Если нулевая гипотеза дисперсионного анализа отклоняется, то требуется определить, какие именно группы имеют значимое различие средних. Метод контрастов позволяет провести необходимые сравнения.

1.4.1. Анализ различия средних значений между уровнями фактора

Формулы приведены для случая двухфакторного дисперсионного анализа, но могут быть распространены и на случай большего числа факторов. Пусть нулевая гипотеза отклоняется для фактора A. Контраст Lk определя-

ется как линейная комбинация

$$Lk = \sum_{j=1}^{k_1} c_j a_j,$$

где c_j, $j = 1, ..., k_1$, - задаваемые контрасты, k_1 - число уровней фактора A, причем $\sum\limits_{j=1}^{k_1} c_j = 0$. Оценка контраста имеет следующий вид:

$$\hat{Lk} = \sum_{j=1}^{k_1} c_j \bar{X}_j,$$

где $\bar{X}_j$ - среднее значение уровня j фактора A. Рассмотрим нулевые гипотезы метода $H_0^{rs} : a_r = a_s, s \neq r$ против двусторонних альтернативных гипотез $H_1^{rs} : a_r \neq a_s, s \neq r$. Гипотеза $H_0^{rs} : a_r = a_s$ равносильна гипотезе $H_0^{rs} : Lk_{rs} = 0$, где

$$Lk_{rs} = a_r - a_s, c_r = 1, c_s = -1, c_j = 0, j \neq r, j \neq s.$$

Проверка нулевой гипотезы проводится по критерию Фишера, статистика критерия вычисляется по формуле

$$F_{Lk_{rs}} = \frac{MS_{Lk_{rs}}}{MS_{Err}},$$

где MS_{Err} - дисперсия, возникающая вследствие случайной ошибки (для случая классической формы дисперсионного анализа совпадает с внутригрупповой дисперсией). $MS_{Lk_{rs}}$ вычисляется на основании следующей величины:

$$SS_{Lk_{rs}} = \frac{k_2 n \hat{Lk}_{rs}^2}{\sum\limits_{j=1}^{k1} c_j^2},$$

где k_2 - число уровней фактора B, n - общее количество наблюдений[10][11].

1.4.2. Анализ взаимодействий факторов

Метод контрастов применим не только для сравнения уровней одного фактора, но также и для анализа взаимодействий факторов. Пересечение контрастов факторов может помочь установить источник эффекта взаимодействия. Контраст Lk_{AB} определяется как линейная комбинация:

$$Lk_{AB} = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} c_{ij} \bar{X}_{ij},$$

где k_1, k_2 - количество уровней факторов A и B соответственно,
$\sum_{i=1}^{k1} \sum_{j=1}^{k2} c_{ij} = 0, c_{ij} = c_{A_i} c_{B_j}, A_i, B_j$ - i-й уровень фактора A и j-й уровень фактора B соответственно.

Нулевая гипотеза имеет вид: $H_0 : Lk_{AB} = 0$ и проверяется с помощью критерия Фишера. Статистика критерия вычисляется по формуле

$$F_{Lk_{AB}} = \frac{MS_{Lk_{AB}}}{MS_{Err}},$$

где MS_{Err} - дисперсия, возникающая вследствие случайной ошибки, а $MS_{Lk_{rs}}$ вычисляется на основании следующей величины:

$$SS_{Lk_{AB}} = \frac{n\hat{Lk}_{AB}^2}{\sum_{i=1}^{k1} \sum_{j=1}^{k2} c_i j^2},$$

где n - общее количество наблюдений[11].

1.5. Критерий Тьюки

Ещё одним способом выяснить какие именно группы имеют существенное различие средних, является критерий Тьюки[2]. Этот критерий позволяет попарно сравнить все группы. Имеется k выборок равного объёма n из нормально распределённой совокупности:

$x_{11}, \ldots, x_{1n},$

$x_{21}, \ldots, x_{2n},$

$\ldots$

$x_{k1}, \ldots, x_{kn}.$

Проверяется гипотеза о статистической неразличимости средних:

$H_0 : \bar{\mu}_1 = \bar{\mu}_2 = \ldots = \bar{\mu}_k.$

Статистика критерия имеет вид:

$$M_{ij} = \frac{|\bar{x}_i - \bar{x}_j|}{s\sqrt{\frac{n}{2}}},$$

где s^2 является оценкой общей дисперсии с $\nu = k(n-1)$ степенями свободы, т.е.

$$s^2 = \frac{1}{k(n-1)} \sum_{j=1}^{k} \sum_{i=1}^{n} (x_{ij} - \bar{x})^2.$$

Если $M_{ij} < m_{\alpha,k*,\nu}$, то средние i-й и j-й выборок признаются не различающимися.

Здесь $k* = \frac{k(k-1)}{2}$, $m_{\alpha,k*,\nu}$ - верхняя критическая точка модуля "стъюдентизированного"максимума.

Таким образом, нулевая гипотеза равенства всех $j = 1, ..., k$ средних не отклоняется только тогда, когда все $\frac{k(k-1)}{2}$ пар средних удовлетворяют вышеприведенному неравенству.

Таблицы значений $m_{\alpha,k*,\nu}$ опубликованы в [12].

2 Двухфакторный дисперсионный анализ

2.1. Двухфакторный дисперсионный анализ с повторными измерениями

Определение 2.1: *Дисперсионный анализ с повторными измерениями* - это такой вид дисперсионного анализа, для которого на каждом уровне исследуемого фактора измерения производятся на одних и тех же субъектах. Рассмотрим его для случая двух факторов.

Полная сумма квадратов в этом случае разбивается следующим образом:

$S_{total} = S_A + S_B + S_{AB} + S_{WG}$,

где S_{WG}, в свою очередь, разбивается на S_{Err} и S_S.

S_S - внутригрупповая сумма квадратов для фактора субъекты, S_{Err} - сумма квадратов ошибки. Дело в том, что вследствие того, что на каждом уровне фактора измерения производятся на одних и тех же субъектах, эти субъекты могут рассматриваться в качестве отдельного фактора. Все дальнейшие вычисления производятся аналогично простому случаю, с той лишь разницей, что для вычисления внутригрупповой дисперсии вместо S_{WG} используется S_{Err} (в случае классического дисперсионного анализа S_{WG} совпадает с S_{Err}). Такая модель дисперсионного анализа имеет преимущество по сравнению с классической моделью: за счет уменьшения суммы квадратов ошибки увеличивается значение F-статистики, а это в свою очередь приводит к повышению мощности критерия для обнаружения значимых различий средних.[13][14]

Для проведения дисперсионного анализа с повторениями необходимо, чтобы помимо трех основных предположений выполнялось ещё и предположение *сферичности*.

Определение 2.2: *Сферичность* – свойство, согласно которому, дисперсии

разностей между различными уровнями фактора с повторными измерениями равны. Нулевая гипотеза о том, что набор выборок удовлетворяет условию сферичности, проверяется с помощью теста Моучли.[15] При нарушении условия сферичности вероятность ошибочного отклонения нулевой гипотезы становится больше, чем уровень значимости α, для устранения этого эффекта выполняется поправка статистики F-критерия по методу Гринхауса-Гейсера.[16]

2.2. Описание данных

Объектом исследования в этой работе являлась база данных сети аптек (название не разглашается в связи с коммерческой тайной), содержащая информацию о доходе по каждой единице товара в трех различных филиалах, детализированную по месяцам (были представлены данные за 2015 год).

Аптечная сеть включает в себя три филиала, один из которых расположен на первом этаже торгового центра (далее *филиал 3*), остальные две торговые точки расположены в независимых торговых помещениях (*филиал 1*, *филиал 2*).

Целью исследования было выяснить, оказывают ли время года и месторасположение конкретного филиала влияние на доход сети по основной категории товаров – лекарственным препаратам. Многофакторный дисперсионный анализ был выбран основным инструментом анализа. Математические расчеты и построение графиков проводились с помощью статистического пакета R. Так как измерения по каждому товару производятся на всех уровнях обоих факторов (для измерений на разных уровнях используются одни и те же наименования товаров), анализ проводился по модели двухфакторного дисперсионного анализа с повторениями.

2.3. Сбор и подготовка данных

Из генеральной совокупности товаров категории *лекарственные средства*, представленных в ассортименте каждого из филиалов, случайным образом было отобрано 100 наименований. Для каждого времени года был вычислен суммарный доход по каждому товару. Аналогичные вычисления были проведены для оставшихся двух филиалов сети. Таким образом, были сформированы данные, состоящие из двенадцати выборок (для всех сочетаний уровней факторов *филиал* и *сезон*) или групп. Для дальнейшей обработки данные были сформированы в таблицу, содержащую тринадцать столбцов (столбец «наименование» и двенадцать столбцов, каждый из которых представляет собой одну из вышеупомянутых групп). С помощью функции `read.delim` был произведен импорт данных в R.

2.4. Проверка данных

Перед проведением исследования необходимо убедиться, что подготовленные данные удовлетворяют основным положениям дисперсионного анализа, описанным в предыдущей главе.

1. Данные были отобраны из генеральной совокупности случайным образом, а значит каждая выборка имеет случайный и независимый характер.

2. Для проверки предположения о нормальном распределении зависимой переменной, которой является переменная "доход была использована функция `shapiro.test`, осуществляющая проверку нулевой гипотезы о нормальном распределении выборки по критерию Шапиро-Уилка. Так как для каждой из тестируемых групп p-значение (вероятность ошибки первого рода) превысило значение заданного уровня значи-

мости $\alpha = 0.05$, нулевая гипотеза была принята для каждой из двенадцати выборок.

3. Проверка гомоскедастичности групп была произведена с помощью функции `cochran.test`, являющейся программной реализацией критерия Кохрана. Вероятность ошибки первого рода, как и в предыдущем пункте, превысила заданное значение $\alpha = 0.05$, что позволило сделать выбор в пользу принятия нулевой гипотезы о равенстве дисперсий генеральных совокупностей, из которых тестируемые выборки были извлечены.

Убедившись, что данные удовлетворяют исходным положениям ANOVA, можно переходить к работе с ними.

2.5. Дисперсионный анализ

Из-за особенностей работы в R с данными, содержащими повторные измерения, таблицу нужно трансформировать в формат long (на одну строку приходится одно наблюдение). Для этого используется функция melt. Данные в новом формате сохранены в переменной `longData`. Новая таблица состоит из трех колонок: «наименование», «группы», содержащей имена исходных колонок, из которой взята информация о доходе, и «доход», представляющей собой колонку с данными о доходе по каждому наименованию. Можно заметить, что столбец «группы» содержит информацию как о сезоне, так и о филиале. Разделим эти признаки. Зная, что первые 400 строк содержат информацию о первом филиале, каждые 100 из которых - об одном из четырех сезонов, и что та же логика справедлива для следующих 800 строк, создадим колонки «сезон» и «филиал». Для этого используем функцию `gl`. Теперь можно переходить непосредственно к анализу. Прежде

всего, построим график boxplot, чтобы оценить данные графически. Результат можно увидеть ниже.

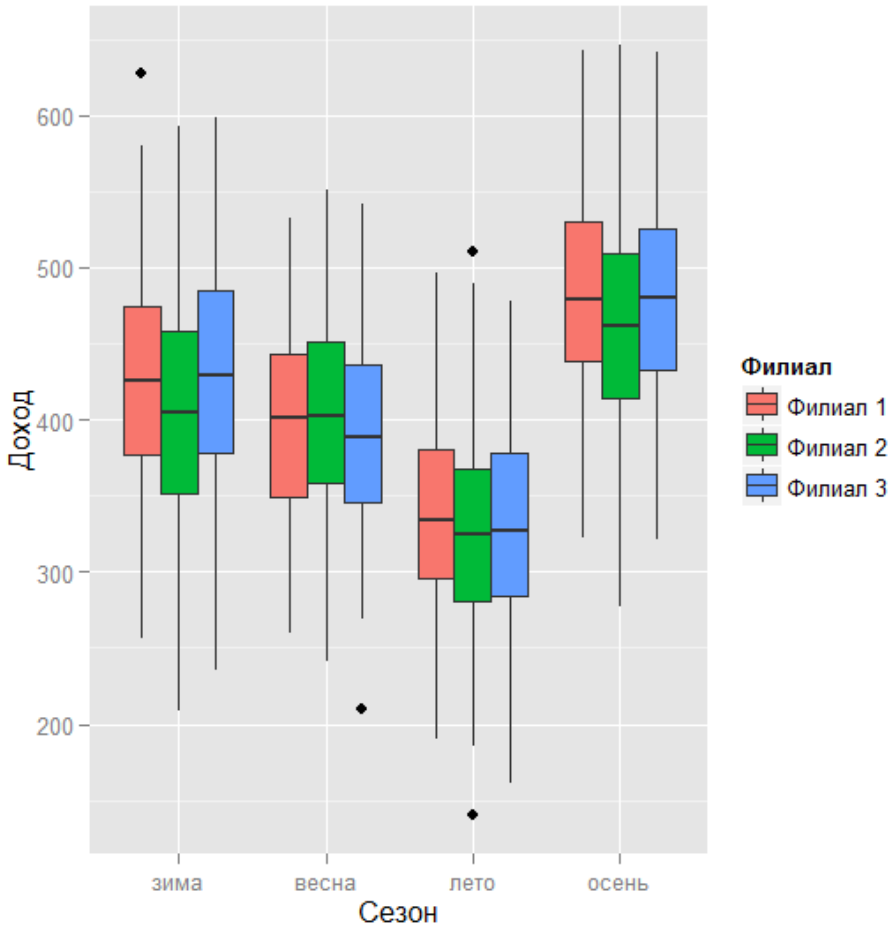

Рис. 1: график boxplot

Как видно по графику, некоторая тенденция к снижению размера дохода наблюдается в летний период, в осенний же период он, напротив, превышает аналогичное значение для остальных сезонов. Дальнейший анализ позволит выявить значимость этих различий, а также сделать выводы о наличии (отсутствии) значимых различий между размером дохода среди трех филиалов.

Проведем дисперсионный анализ с помощью функции `aov_ez` пакета `afex`. Применив к построенной модели функцию `summary`, выведем на экран полученный результат.

```
Univariate Type III Repeated-Measures ANOVA Assuming Sphericity

                   SS num Df Error SS den Df          F  Pr(>F)
(Intercept)  196501016      1   542914     99 35831.8168 < 2e-16 ***
сезон          3238953      3  1301470    297   246.3801 < 2e-16 ***
филиал           24575      2   964142    198     2.5234 0.08276 .
сезон:филиал     52797      6  2700122    594     1.9358 0.07302 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Mauchly Tests for Sphericity

             Test statistic p-value
сезон                0.95689 0.50620
филиал               0.97199 0.24856
сезон:филиал         0.84535 0.70061
```

Рис. 2: результат ANOVA

Анализ на значимость различий уровней факторов следует начать с оценки результатов проверки на сферичность с помощью теста Моучли. По результатам этого теста можно заключить, что свойство сферичности выполняется для каждого фактора и для их взаимодействия (p-значение$> \alpha = 0.05$), поэтому нет необходимости использовать исправленные p-значения. Посмотрев на результаты и оценив -значения (вероятности ошибки при отклонении нулевой гипотезы), можно увидеть для каких факторов и комбинаций нулевая гипотеза может быть принята. Такое заключение мы можем сделать для фактора *филиал* – согласно этому заключению, фактор *филиал* не имел значительного влияния на размер дохода. Гипотеза принимается и для взаимодействия *сезон-филиал* – это значит, что фактор *сезон* оказывал одинаковое влияние на размер дохода в каждом из трех филиалов. Анализ позволил выявить наличие значимых различий между уровнями фактора *сезон*.

Чтобы визуально оценить полученные результаты, построим график.

1. Сезон

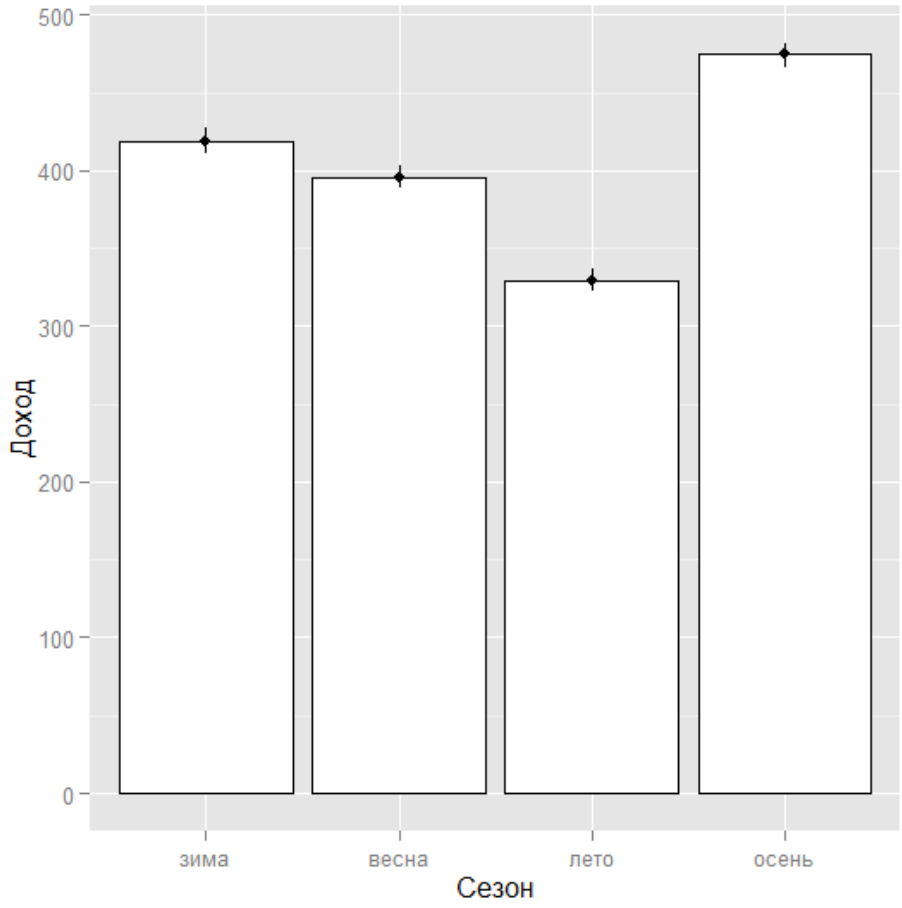

Рис. 3: сезон

Согласно результатам дисперсионного анализа, фактор *сезон* является значимым. Это значит, что среднее значение дохода по каждому сезону, приходящегося на каждый товар, различается хотя бы для двух времен года. Дисперсионный анализ проверяет нулевую гипотезу о наличии различия между средними значениями уровней фактора. Чтобы выяснить, где именно лежит это различие, необходимо воспользоваться методом контрастов.

Небезызвестным является тот факт, что в летний период уровень продаж в аптеках как правило снижается, поэтому целесообразно считать

значения уровня *лето* контрольной выборкой. С помощью первого контраста сравним значения дохода в летний период и значения дохода в остальные периоды года. Далее, можно заметить, что согласно построенному графику, среднее значение дохода в осенний период было несколько выше остальных, поэтому с помощью второго контраста сравним уровни зима и весна с уровнем осень. Третий контраст позволит сравнить между собой доход в зимний и весенний периоды. Создадим переменные для этих контрастов:

```
ZVOvsL<-c(1, 1,-3,1)
ZVvsO<-c(1, 1, 0,-2)
ZvsV<-c(1, -1, 0,0)
```

С помощью функции contrast проведем дополнительный анализ по методу контрастов.

contrast	estimate	SE	df	t.ratio	p.value
Все_лето	300.18641	13.239417	297	22.674	<.0001
зимавесна_осень	-134.60478	9.361682	297	-14.378	<.0001
зима_весна	23.12772	5.404969	297	4.279	<.0001

Рис. 4: метод контрастов

Как можно видеть на рисунке 4, p-значения для каждого из установленных контрастов оказалось достаточно малым, чтобы считать каждый из них значимым. Проводя интерпретацию результатов, полученных при исследовании фактора *сезон*, можно заключить, что доход в летний период существенно более низкий, чем в остальные времена года, доход в первом полугодии значительно ниже, чем осенью, а при сравнении зимнего и весеннего периода обнаруживается существенное превосходство значения среднего дохода зимой. Другими словами, уровень дохода исследуемого предприятия носит сезонный характер.

2. Филиал

По результатам дисперсионного анализа значимых различий между средними значениями дохода за единицу товара по трем филиалам выявлено не было, что абсолютно согласовывается с визуальным представлением.

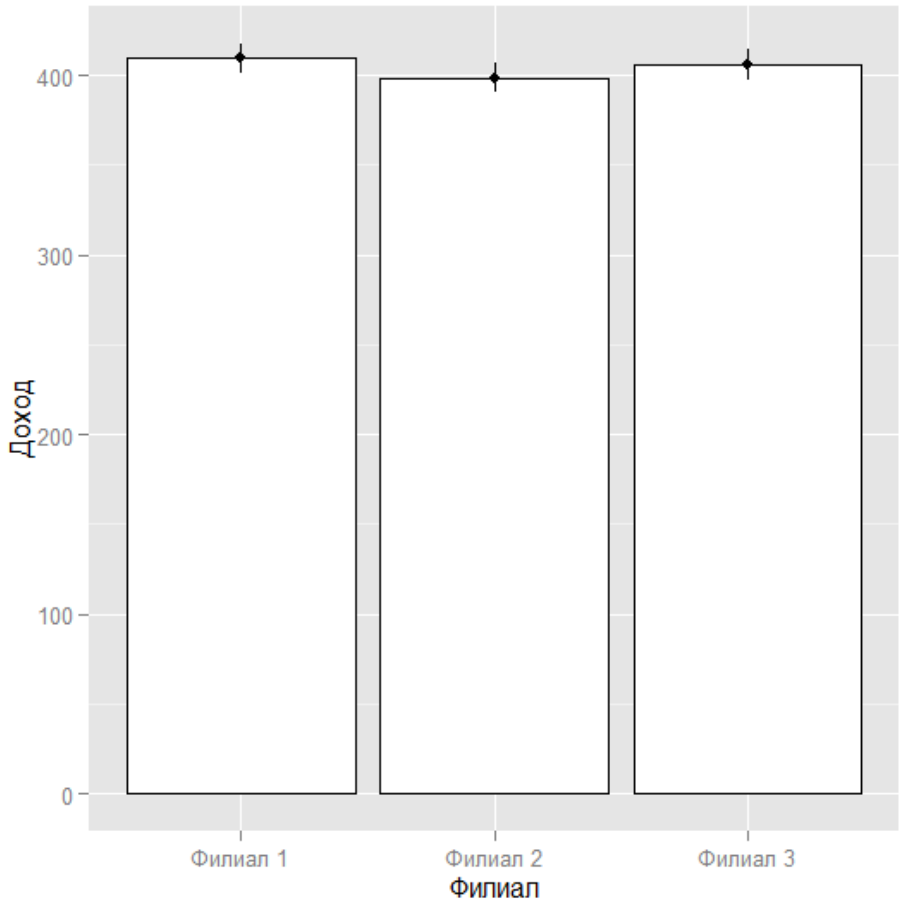

Рис. 5: филиал

Такой результат не требует проведения дополнительных исследований.

Интерпретировать полученный результат можно следующим образом: торговые точки приносят примерно одинаковый доход, независимо от их месторасположения. Примечательным является и тот факт, что различий между уровнем дохода от продаж в филиале, расположенном в

торговом центре и продаж в двух других торговых точках выявлено не было.

3. Сезон и филиал

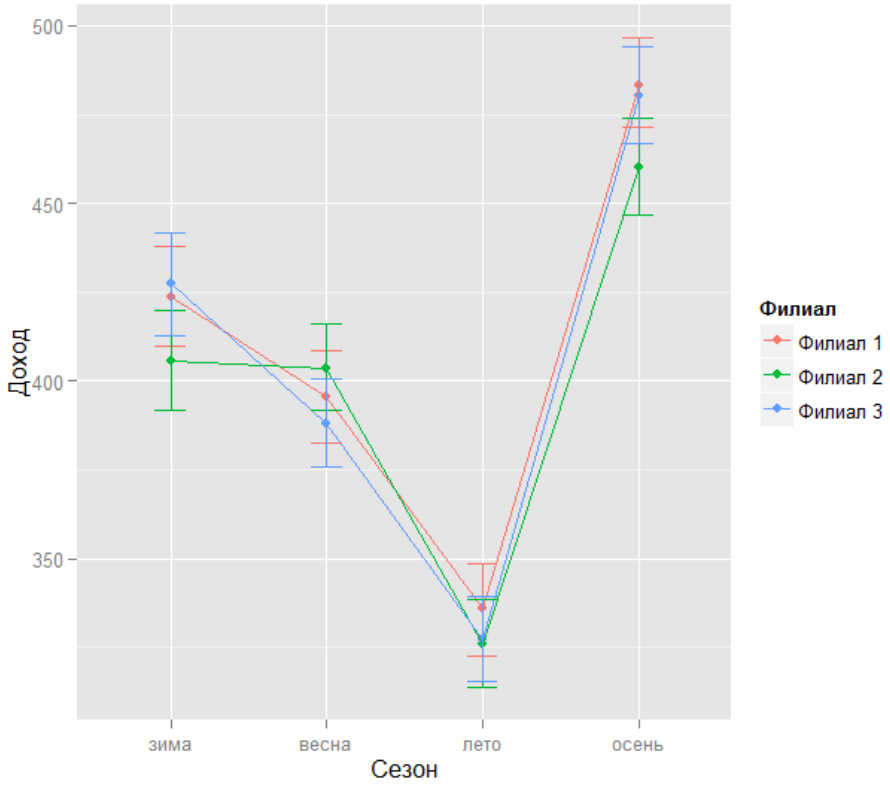

Рис. 6: эффект взаимодействия

Взаимодействие факторов *сезон* и *филиал*, согласно проведенному анализу, также не является значимым. Это означает, что влияние фактора сезон на всех уровнях фактора филиал проявляется одинаково. Другими словами, явление сезонности проявляется одинаково в каждом из трех филиалов.

2.6. Вывод

По результатам исследования можно заключить, что из двух исследуемых факторов, существенное влияние на уровень дохода оказывает лишь

один из них – сезон. Эффекта взаимодействия факторов выявлено не было, а значит явление сезонности проявляется одинаково в каждом из трех филиалов. Подводя итог, можно сказать, что средний доход, получаемый от продажи ряда товаров категории «лекарственные препараты», варьируется в зависимости от сезона, достигая своего максимума в осенний период и существенно снижаясь в летний. Уровни дохода в различных точках продажи не имеют значимых различий.

3 Трехфакторный дисперсионный анализ: смешанная модель

Определение 3.1: *Многофакторный дисперсионный анализ по смешанной модели* - это такой вид дисперсионного анализа, который включает в себя и межгрупповые (переменные, на разных уровнях которых измерения производятся на одних и тех же субъектах), и внутригрупповые переменные. Рассмотрим эту модель на примере двухфакторного анализа. Пусть A - межгрупповая переменная, а B - внутригрупповая. Полная сумма квадратов в этом случае разбивается следующим образом:

$$S_{total} = S_A + S_B + S_{AB} + S_{E_a} + S_{E_b} + S_{WG},$$

где:

$S_{total} = \sum_{i=1}^{a}\sum_{j=1}^{n}\sum_{k=1}^{t} y_{ijk}^2 - Nt\bar{y}_{...}^2$

$S_A = t\sum_{i=1}^{a} n\bar{y}_{i..}^2 - Nt\bar{y}_{...}^2$

$S_B = N\sum_{k=1}^{t} \bar{y}_{..k}^2 - Nt\bar{y}_{...}^2$

$S_{AB} = \sum_{i=1}^{a}\sum_{k=1}^{t} n\bar{y}_{i.k}^2 - Nt\bar{y}_{...}^2 - S_A - S_B$

$S_{E_a} = t\sum_{i=1}^{a}\sum_{j=1}^{n} \bar{y}_{ij.}^2 - t\sum_{i=1}^{a} n_i\bar{y}_{i..}^2$

$S_{E_b} = S_{total} - S_A - S_B - S_{AB} - S_{E_a}$

N - общее число наблюдений, a - количество уровней фактора A, t - количество уровней фактора B, n - количество наблюдений в каждой ячейке.[17] Дальнейшие вычисления производятся аналогично классической модели.

3.1. Планирование

Исследование, проведенное в рамках главы 1, можно расширить, добавив в него ещё один фактор. Таким фактором была выбрана категория то-

варов. Новый вариант исследования предполагает расширение списка анализируемых товаров за счет добавления к ним наименований из двух категорий, составляющих большую часть ассортимента сети аптек – *БАДы* (биологически активные добавки) и *предметы личной гигиены*. С появлением третьего фактора, добавилось ещё три взаимодействия, влияние которых также необходимо проверить. Новая независимая переменная – категория, не является внутригрупповой (измерения для каждого уровня этой переменной проводятся на разных товарах), таким образом, исследование принимает вид трехфакторного дисперсионного анализа, проводимого по смешанной модели (две внутригрупповых переменных и одна межгрупповая).

3.2. Сбор данных

Как и в предыдущей главе, из генеральной совокупности каждой из трех вышеупомянутых категорий товаров, представленных в ассортименте каждого из филиалов, случайным образом было отобрано 100 наименований. Для каждого времени года был вычислен суммарный доход по каждому товару. Такие манипуляции были проведены для каждого из трех филиалов. Таким образом, данные, подготовленные к обработке, представляли собой таблицу, содержащую четырнадцать столбцов (столбцы «наименование» и «категория», и двенадцать столбцов с информацией о доходе – по четыре столбца, в соответствии с количеством времен года, для каждого из трех филиалов).

3.3. Дисперсионный анализ

После проведения проверки полученных выборок на нормальность распределения генеральных совокупностей и на равенство дисперсий, можно переходить к непосредственной работе с данными. Как и в случае с двух-

факторным анализом, наличие повторных измерений обуславливает необходимость приведения таблицы в формат long с помощью функции melt. Данные в новом формате сохранены в переменной Data1. Новая таблица состоит из шести колонок: «наименование», «категория», имя колонки, из которой взята информация о доходе («группы»), колонка, содержащая данные о доходе («доход»), а также две колонки, содержащие имена уровней факторов сезон и филиал. Перед началом анализа построим и проанализируем график boxplot (рис. 7).

Глядя на график, можно заметить, что сезонность более всего выражена для категорий *БАДы* и *лекарственные средства*, в то время как уровень дохода по товарам категории *предметы личной гигиены* слабо колеблется в течение года. Дисперсионный анализ позволит решить стоит ли принять или отвергнуть это предположение, и сделать дальнейшие выводы.

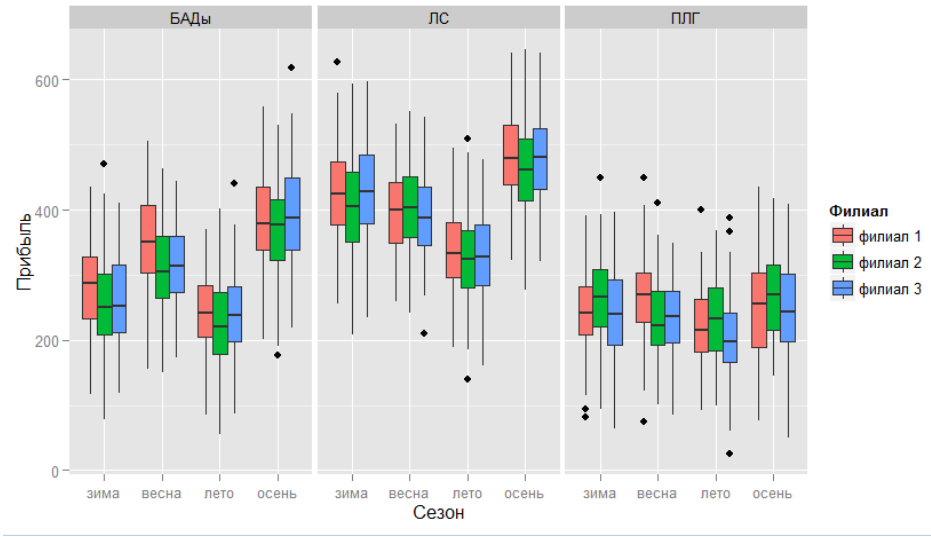

Рис. 7: график boxplot

Воспользовавшись функцией `aov_ez`,проведем дисперсионный анализ.

```
Univariate Type III Repeated-Measures ANOVA Assuming Sphericity

                            SS num Df Error SS den Df          F    Pr(>F)
(Intercept)          360694332      1  1411163    297 75913.4047 < 2.2e-16 ***
категория             16147102      2  1411163    297  1699.1970 < 2.2e-16 ***
филиал                   86278      2  2627596    594     9.7521 6.804e-05 ***
категория:филиал        134148      4  2627596    594     7.5814 5.830e-06 ***
сезон                  5592614      3  3813511    891   435.5583 < 2.2e-16 ***
категория:сезон        1887009      6  3813511    891    73.4810 < 2.2e-16 ***
филиал:сезон             76620      6  8310039   1782     2.7384 0.0118837 *
категория:филиал:сезон  170817     12  8310039   1782     3.0525 0.0002835 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Mauchly Tests for Sphericity

                      Test statistic p-value
филиал                       0.99848 0.79841
категория:филиал             0.99848 0.79841
сезон                        0.98967 0.68888
категория:сезон              0.98967 0.68888
филиал:сезон                 0.91609 0.17135
категория:филиал:сезон       0.91609 0.17135
```

Рис. 8: результат ANOVA

Анализ на значимость различий уровней факторов следует начать с оценки результатов проверки на сферичность с помощью теста Моучли. Обратившись к результатам теста Моучли, содержащимся в таблице проведенного дисперсионного анализа (рис. 8), можно заключить, что свойство сферичности выполняется для каждого фактора и для их взаимодействий (p-значение> 0.05).

Посмотрев на результаты и оценив p-значения, можно увидеть для каких факторов и комбинаций нулевая гипотеза может быть отвергнута. Такое заключение мы можем сделать для всех факторов и для всех взаимодействий факторов. Так как взаимодействие трех факторов является значимым, интерпретация главных эффектов факторов и взаимодействия пар факторов может оказаться недостоверной. Это значит, что основные выводы должны основываться именно на интерпретации эффекта взаимодействия трех факторов. Поэтому, стоит сразу перейти к рассмотрению этого главного эффекта.

Значимость взаимодействия трех факторов означает, что одно или несколько двойных взаимодействий значительно различаются вдоль уровней третьей переменной.

Для начала построим график, иллюстрирующий взаимодействие факторов *сезон* и *филиал* для каждого из уровней фактора *категория*. Можно предположить, что это взаимодействие оказывает большее влияние на третий уровень фактора *категория*, чем на два остальных. Об этом свидетельствует тот факт, что на первых двух графиках линии проходят практически параллельно, в то время как на третьем графике присутствуют пересекающиеся линии. Чтобы проверить выдвинутое предположение, проведем двухфакторный дисперсионный анализ отдельно для каждого уровня фактора *категория* При этом нужно отметить, что при вычислении статистики критерия Фишера в качестве значения внутригрупповой дисперсии необходимо использовать значение, вычисленное при проведении дисперсионного анализа для трех факторов, тем самым сохранив оценку внутригрупповой вариации признаков неизменной.

Для каждого полученного значения F-статистики может быть вычислено p-значение. Это позволит увидеть при каком уровне значимости нулевая гипотеза может быть принята. По результатам двухфакторного дисперсионного анализа были получены следующие значения:

1. БАДы:

   $F = 2.5031$

   $p = 0.0204$

2. Лекарственные средства:

   $F = 1.887$

   $p = 0.0796$

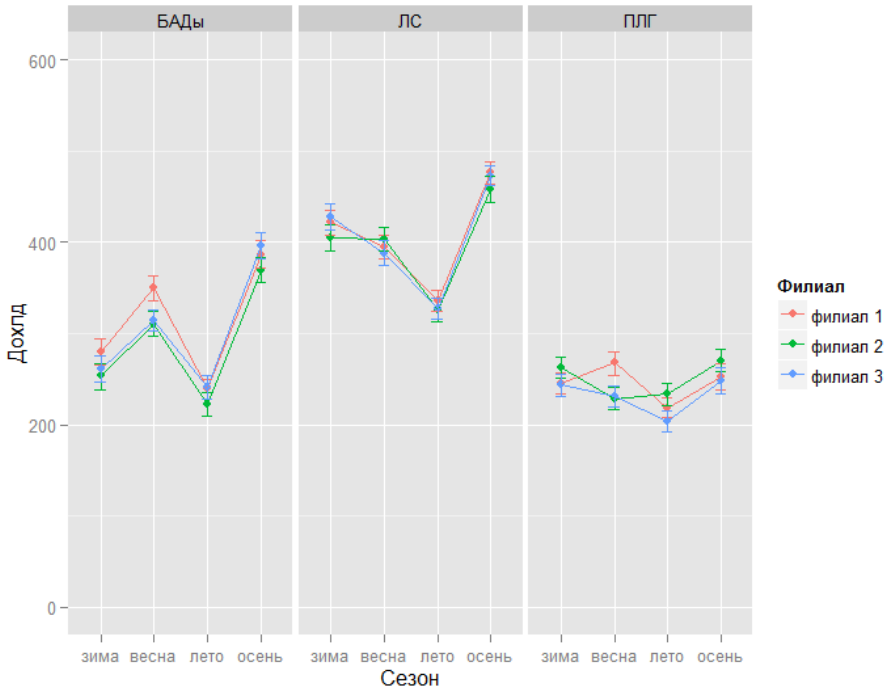

Рис. 9: двойные взаимодействия: сезон-филиал

3. Предметы личной гигиены:

   $F = 4.4533$

   $p = 0.00018$

Таким образом, нулевая гипотеза об отсутствии эффекта взаимодействия факторов *сезон* и *филиал* принимается только на уровне *лекарственные средства* при уровне значимости $\alpha = 0.05$. Это значит, что для получения более детальной информации необходимо изучить влияние этих факторов на зависимую переменную по отдельности. Так как суммы квадратов для каждого фактора уже были вычислены на предыдущем этап, остается только вычислить статистику F-теста, используя в качестве значения внутригрупповой дисперсии значение, вычисленное при проведении дисперсионного анализа для трех факторов.

1. Сезон:

   $F = 252.253$

$$p = 3.32 \cdot 10^{-188}$$

2. Филиал:

   $F = 2.77$

   $p = 0.011$

Для фактора *сезон* вероятность ошибки первого рода очень мала, что говорит о наличии сильного влияния этого фактора на зависимую переменную. Для фактора *филиал* можно отклонить нулевую гипотезу при $\alpha = 0.05$, но при $\alpha = 0.01$ она принимается. Чтобы выяснить где точно лежат установленные различия, воспользуемся методом Тьюки.

категория = ЛС:

contrast	estimate	SE	df	t.ratio	p.value
зима - весна	23.127722	5.341682	891	4.330	0.0001
зима - лето	89.191867	5.341682	891	16.697	<.0001
зима - осень	-55.738530	5.341682	891	-10.435	<.0001
весна - лето	66.064145	5.341682	891	12.368	<.0001
весна - осень	-78.866253	5.341682	891	-14.764	<.0001
лето - осень	-144.930397	5.341682	891	-27.132	<.0001

Рис. 10: фактор сезон: метод Тьюки

категория = ЛС:

contrast	estimate	SE	df	t.ratio	p.value
филиал.1 - филиал.2	10.924172	4.702958	594	2.323	0.0535
филиал.1 - филиал.3	3.833291	4.702958	594	0.815	0.6938
филиал.2 - филиал.3	-7.090880	4.702958	594	-1.508	0.2880

Рис. 11: фактор филиал: метод Тьюки

Как видно в таблице результатов метода, p-значение значительно меньше 0.05 для всех сочетаний уровней фактора *сезон*, что говорит о том, что все различия между уровнями фактора, проиллюстрированные на рис. 3.7, являются значимыми: средний доход по товарам категории *лекарственные средства* является наибольшим для осеннего периода и наименьшим для зимнего.

Согласно результатам применения метода Тьюки, фактор *филиал* не имеет значимых различий между уровнями (при $\alpha = 0.01$). Опираясь на этот факт, а также на результаты дисперсионного анализа (p-значение$= 0.011$), можно признать, что различия между уровнями фактора филиал не являются значимыми для товаров категории *лекарственные средства*.

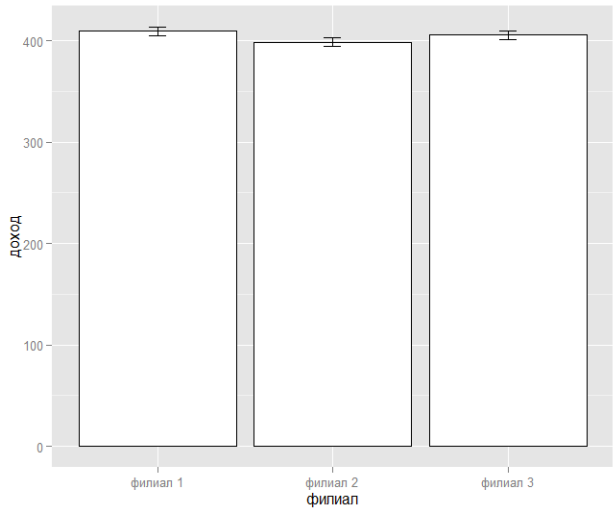

Рис. 12: лекарственные средства: сезон

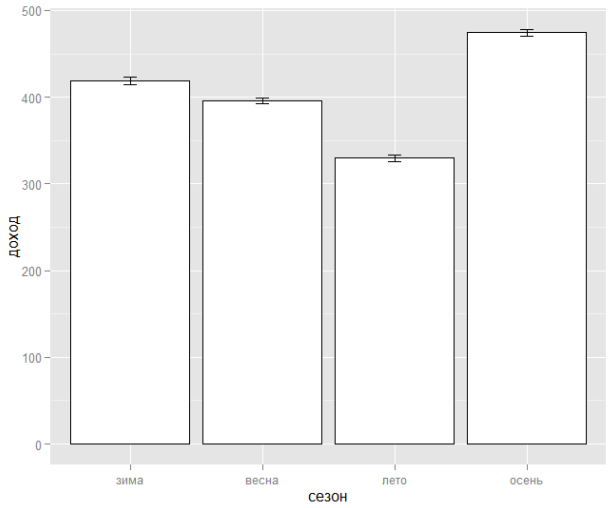

Рис. 13: лекарственные средства: филиал

Для проведения дальнейшего анализа на уровнях *БАДы* и *предметы личной гигиены* фактора *категория*, может быть применен метод контрастов.

1. БАДы:

   Посмотрев на график, можно предположить, что значимые различия имеются между разницей уровней *весна*, *лето* фактора *сезон* на первом уровне фактора *филиал*, и этой разницей на его третьем уровне. Такое же предположение можно сделать и для второго и третьего уровней фактора *филиал*. Зададим вектора, представляющие контрасты.

   t1=c(0,0,0,0, 0,-1,1,0, 0,1,-1,0, 0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)

   t2=c(0,-1,1,0, 0,0,0,0, 0,1,-1,0, 0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)

Каждый вектор представляет собой набор весовых коэффициентов: по 12 коэффициентов на каждый уровень фактора *категория*, разбивающиеся, в свою очередь, на три группы, представляющие собой коэффициенты уровней фактора *филиал*, и включающие по 4 коэффициета - по одному на каждый из уровней фактора *сезон*.

Применив команду `contrast`,получим таблицу, включающую в себя p-значения для каждого из контрастов. Значимое различие было обнаружено для разницы среднего дохода на уровнях *весна* и *лето* между уровнями *филиал 1* и *филиал 3* ($p - = 0,0053$). Это значит, что снижение дохода при смене весеннего периода на летний выражено значительно больше в первом филиале, чем в третьем (расположенном в торговом центре).

2. Предметы личной гигиены:

   По аналогии с тем, как это было сделано для уровня *БАДы*, зададим контрасты для уровня *предметы личной гигиены* (для тех сочетаний уровней двух факторов, на которых, на основании графического представления, может быть сделано предположение о наличии значимых различий).

   t1=c(0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0, 1,-1,0,0, -1,1,0,0, 0,0,0,0)
   t2=c(0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0, 1,-1,0,0, 0,0,0,0, -1,1,0,0)
   t3=c(0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0, 0,1,-1,0, 0,-1,1,0, 0,0,0,0)
   t4=c(0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0, 0,1,-1,0, 0,0,0,0, 0,-1,1,0)
   t5=c(0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0, 1,-1,0,0, -1,1,0,0)
   t6=c(0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0, 0,1,-1,0, 0,-1,1,0)

   Применим метод контрастов. Получим, что разница среднего дохода между уровнями *зима-весна* существенно различалась для первого и второго, а также для первого и третьего филиалов. Анализ также по-

казал, что снижение дохода при смене весеннего периода на летний выражено значительно больше в первом филиале, чем во втором; такой же вывод справедлив и для третьего и второго филиалов (во втором филиале средний доход, приходящийся на единицу товара, даже возрастает).

Следующим шагом проанализируем взаимодействие факторов *филиал* и *категория* вдоль уровней фактора *сезон*. Построим график.

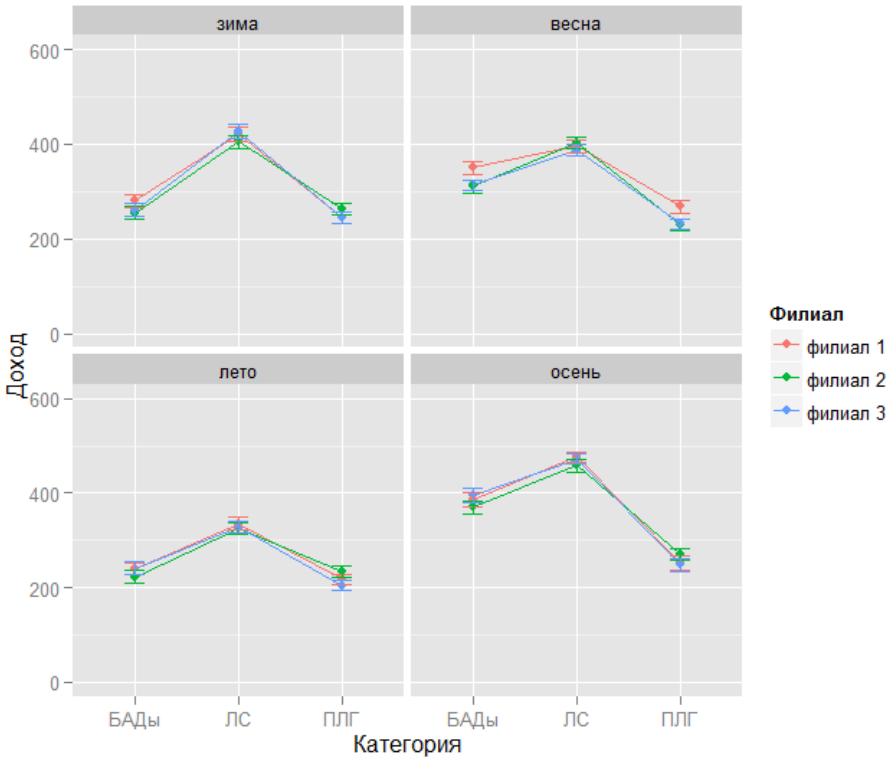

Рис. 14: двойные взаимодействия: филиал-категория

На каждом уровне фактора *сезон* проведем двухфакторный дисперсионный анализ для проверки гипотезы отсутствия эффекта взаимодействия факторов *филиал* и *категория*. Результаты представлены ниже:

1. Зима:

   $F = 4.2934$

   $p = 0.00197$

2. Весна:

   $F = 4.3338$

   $p = 0.00184$

3. Лето:

   $F = 3.5934$

   $p = 0.0067$

4. Осень:

   $F = 5.0146$

   $p = 0.00056$

По результатам анализа можно заключить, что нулевая гипотеза отклоняется на каждом уровне фактора *сезон*, то есть для каждого сезона справедливо утверждение о том, что влияние фактора *филиал* различно для товаров разных категорий.

Так как было установлено наличие эффекта взаимодействия, для получения более детальной информации необходима дальнейшая интерпретация. Чтобы выяснить, где именно лежат обнаруженные различия, воспользуемся методом контрастов. По графику сложно сделать предположение о том какие контрасты могут быть значимыми, поэтому имеет смысл применить метод для всех возможных комбинаций уровней факторов.

Не останавливаясь подробно на каждом контрасте, проведем общую интерпретацию полученного результата. Можно заметить, что разница между средним доходом, приносимым товарами категории *БАДы* и средним доходом по товарам категории *лекарственные средства*, не имела существенных различий среди всех филиалов. Это утверждение справедливо для всех сезонов, кроме весны. Также, разница между средним доходом по товарам категории *БАДы* и по товарам категории *предметы личной гигиены*

не имела существенных различий для первого и третьего филиалов - это справедливо для всех сезонов (доход от продажи товаров категории *БАДы* превышает доход от продажи товаров категории *предметы личной гигиены* в первом филиале примерно настолько же, насколько он превышает его в третьем филиале). Такое же утверждение справедливо и для товаров категорий *лекарственные средства* и *предметы личной гигиены*, для первого и третьего филиалов, для всех сезонов кроме весны. Весной же средний уровень дохода по БАДам и по предметам личной гигиены в первом филиале существенно превышает средний уровень дохода по этим же категориям в других филиалах. Таблицу с результатами применения метода контрастов можно найти в приложениях.

Теперь аналогичным образом проанализируем взаимодействие факторов *категория* и *сезон* вдоль уровней фактора *филиал*.

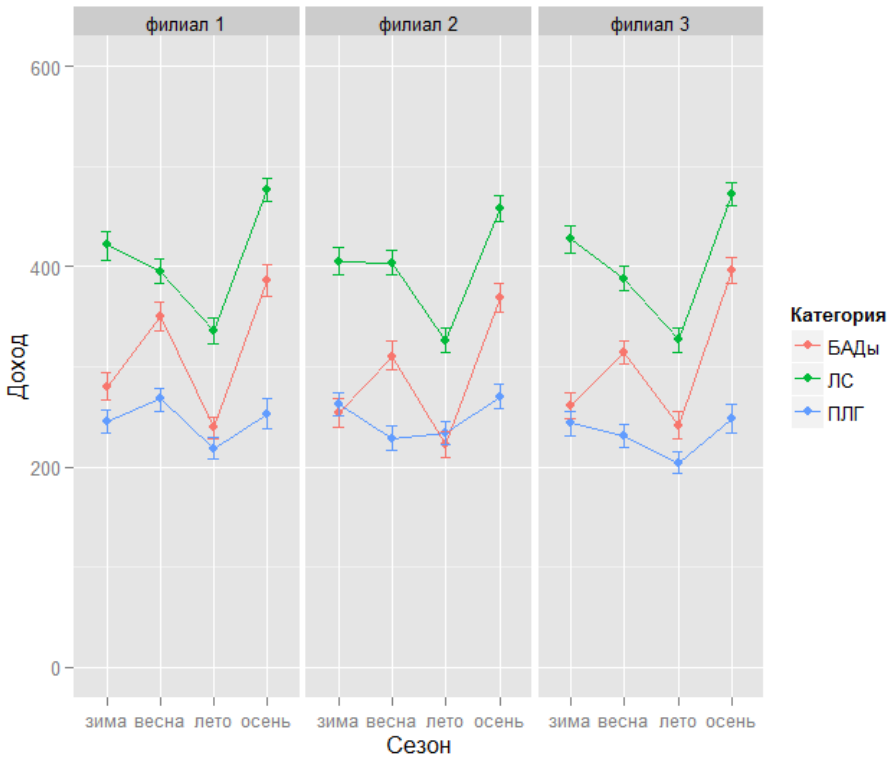

Рис. 15: двойные взаимодействия: категория-сезон

Проведем двухфакторный дисперсионный анализ для проверки нулевой гипотезы отсутствия эффекта взаимодействия факторов *категория* и *сезон* для каждого из уровней фактора *филиал*.

1. Филиал 1:

   $F = 27.1363$

   $p = 8.25 \cdot 10^{-30}$

2. Филиал 2:

   $F = 23.9141$

   $p = 2.55144 \cdot 10^{-26}$

3. Филиал 3:

   $F = 29.082$

   $p = 6.83 \cdot 10^{-30}$

Согласно полученным p-значениям, можно сказать, что анализируемое взаимодействие является значимым для всех уровней фактора филиал, то есть фактор сезонности по разному проявляется на разных уровнях фактора *категория*. Чтобы выяснить где именно лежат выявленные различия, воспользуемся методом контрастов.

Не будем подробно останавливаться на результатах для каждого уровня фактора филиал, сделаем лишь общие выводы (подробные результаты можно найти в Приложении). Можно заметить, что для всех филиалов незначимым оказался контраст *БАДы-лекарственные средства*, *лето-осень*; для двух филиалов (второй и третий) незначимым оказался контраст *БАДы-лекарственные средства*, *весна-лето*. Незначимость этих контрастов может быть обусловлена зависимостью уровня доходов товаров категорий *БАДы* и *лекарственные средства* от сезона, которая, в частности, проявляется в снижении дохода по этим категориям в летний период и его повышении

в осенний. Различия среднего дохода от продажи товаров разных категорий варьируются в зависимости от сезона - это наблюдается при сравнении практически всех сочетаний уровней факторов.

3.4. Вывод

По результатам трехфакторного дисперсионного анализа был сделан ряд выводов: во-первых, было установлено, что для категории *лекарственные средства* фактор сезонности одинаково проявляется во всех трех филиалах. Наибольший доход товары этой категории приносят в осенний период, далее следуют зимний и весенний периоды и на последнем месте стоит летний период. Для товаров оставшихся двух категорий было выявлено наличие эффекта взаимодействия факторов *сезон* и *филиал*. Дисперсионный анализ также позволил сделать вывод о том, что влияние месторасположения филиала по-разному проявляется для товаров разных категорий. Было выявлено наличие значимого взаимодействия факторов категория и сезон: установлено, что фактор сезон по-разному влияет на уровень дохода от продажи товаров разных категорий. С помощью метода контрастов было показано, что товары категорий *БАДы* и *лекарственные средства* имеют схожий характер колебаний уровня дохода в зависимости от сезона, в то время как доход от товаров категории *предметы личной гигиены* имеет слабовыраженную периодичность, как и предполагалось на этапе анализа графика.

Выводы

Проведение многофакторного дисперсионного анализа помогло выявить сильные и слабые стороны этого метода.

Достоинства многофакторного дисперсионного анализа:

1. Метод имеет множество различных форм (дисперсионный анализ с повторными измерениями, смешанная модель, дисперсионный анализ для ячеек с разным количеством измерений), что существенно расширяет варианты проведения исследований
2. Имеется возможность проверить влияние сразу нескольких факторов
3. Метод позволяет проверить наличие влияния взаимодействия факторов

Недостатки многофакторного дисперсионного анализа:

1. Метод чувствителен к нарушениям условий нормальности и гомоскедастичности
2. Чтобы выявить, на каких именно уровнях фактора находится различие, требуется применение дополнительных методов (метод контрастов или post-hoc тесты)

Заключение

В работе было проведено исследование доходности сети аптек: изучалось влияние различных факторов на доход предприятия. Анализ проводился с использованием инструментов статистического пакета R. Первая часть исследования проводилась по модели двухфакторного дисперсионного анализа с повторными измерениями: изучалось влияние факторов *сезон* и *филиал* на доход по товарам категории *лекарственные средства*. После проверки на выполнение основных предположений дисперсионного анализа, данные были импортированы в рабочую среду статистического пакета R и приведены к подходящему для работы формату. Перед началом анализа, на основании графического представления данных, были сделаны предположения о влиянии указанных факторов. Проведенный в дальнейшем дисперсионный анализ позволил подтвердить некоторые из них и опровергнуть остальные. Следующим этапом, к фактору *сезон*, на уровнях которого было обнаружено наличие существенного различия в средних значениях зависимой переменной, был применен метод контрастов, позволивший получить более подробную информацию, в частности, выяснить для каких именно уровней фактора наблюдается существенное различие средних. Таким образом, по результатам первой части исследования, были сделаны выводы о том, что уровень дохода фирмы от продажи товаров категории *лекарственные средства* носит сезонный характер: наблюдается существенный рост уровня дохода сети осенью и его спад в летний период. Кроме того, анализ показал, что доход от продажи товаров данной категории не имел существенных различий среди трех филиалов. Был также сделан вывод об отсутствии эффекта взаимодействия исследуемых факторов.

Вторая часть исследования проводилась по модели смешанного трех-

факторного дисперсионного анализа. Исследование было расширено, были добавлены товары других категорий (*БАДы*, *предметы личной гигиены*), и, вместе с тем, третий фактор - *категория*. Дальнейший анализ показал наличие значимого взаимодействия трех факторов. Вследствие этого, последующая работа была направлена на исследование и интерпретацию именно этого главного эффекта. По результатам применения метода дисперсионного анализа (двухфакторной модели) были выявлены уровни факторов, на которые двойные взаимодействия оказывают значимое влияние. Применения метода контрастов и метода Тьюки позволило уточнить полученную информацию и сформировать окончательные выводы, ознакомиться с которыми можно обратившись к заключению второй главы.

Проведенное исследование позволило выявить достоинства и недостатки метода дисперсионного анализа в различных его формах.

Список литературы

[1] Фишер Р. А. Статистические методы для исследователей. М.:Госстатиздат 1958

[2] Кобзарь А. И. Прикладная математическая статистика. Для инженеров и научных работников. М.:ФИЗМАТЛИТ, 2006.

[3] Шеффе Г. Дисперсионный анализ М.: Наука, 1980

[4] Крамер Г. Математические методы статистики, М.: Мир, 1975

[5] Field A., Miles J., Field Z. Discovering Statistics Using R, 2012

[6] Мастицкий С.Э., Шитиков В.К. (2014) Статистический анализ и визуализация данных с помощью R. – Электронная книга, адрес доступа: http://r-analytics.blogspot.com

[7] Nugzar Todua, Teona Dotchviri ANOVA in marketing research of consumer behavior of different categories in georgian market // Annals of the „Constantin Brancusi" University of Targu Jiu, Economy Series, Issue 1, volume I/2015

[8] Лемешко Б. Ю., Лемешко С. Б. Сравнительный анализ критериев проверки отклонения распределения от нормального закона // Метрология. 2005. № 2. С.3-23

[9] Горбунова А А., Лемешко Б. Ю., Лемешко С. Б. Критерии проверки гипотез об однородности дисперсий при наблюдаемых законах, отличных от нормального // Материалы X международной конференции "Актуальные проблемы электронного приборостроения" АПЭП-2010. Т.6, Новосибирск, 2010. – С.36-41.

[10] Буре В. М., Грауэр Л. В. Лекция 7. Проверка гипотез о равенстве параметров двух нормально распределенных генеральных совокупностей. Однофакторный дисперсионный анализ // Лекция курса "математическая статистика"Computer Science Center, 2013

[11] Keppel G., Wickens T. D. Design and Analysis: A Researcher's Handbook, 2004

[12] Half G. J., Hendricson R. W. A table of percentage pointsof the largest absolute value of k Student t variates and its applications// Biometrika. 1971. V. 58. P. 323-332.

[13] Левин Д. М., Стэфан Д., Кребиль Т. С., Беренсон М. Л. Статистика для менеджеров с использованием Microsoft Excel. М.:Вильямс 2005

[14] https://statistics.laerd.com/statistical-guides/repeated-measures-anova-statistical-guide.php

[15] Mauchly, J. W. Significance test for sphericity of a normal n-variate distribution // The Annals of Mathematical Statistics, 1940, 11, 204-209.

[16] Geisser, S., Greenhouse, S.W. An extension of Box's result on the use of F distribution in multivariate analysis // Annals of Mathematical Statistics 1958. P. 885–891

[17] Jones B., Nachtsheim C. J Split-Plot Designs: What, Why, and How // Journal of Quality Technology, 2009 Vol. 41, No. 4, October 2009

[18] http://www.real-statistics.com/

Приложения

Формулы для случаев двухфакторного и трехфакторного дисперсионного анализа

$$SS_T = SS_A + SS_B + SS_{AB} + SS_W$$

$\bar{x}$ - общее среднее,

$\bar{x}_i$ - среднее значение наблюдений уровня i фактора A (аналогично определяется $\bar{x}_j$),

$\bar{x}_{ij}$ - среднее значение наблюдений, лежащих на пересечении уровней i и j факторов A и B соответственно,

x_{ijk} - k-е наблюдение на пересечении уровней i и j факторов A и B соответственно,

r - количество уровней фактора A,

c - количество уровней фактора B,

m - количество наблюдений в ячейке,

n - общее число наблюдений

$SS_T = \sum_k \sum_j \sum_i (x_{ijk} - \bar{x})^2$	$df_T = n - 1$	$MS_T = SS_T / df_T$
$SS_A = mc \sum_i (\bar{x}_i - \bar{x})^2$	$df_A = r - 1$	$MS_A = SS_A / df_A$
$SS_B = mr \sum_j (\bar{x}_j - \bar{x})^2$	$df_B = c - 1$	$MS_B = SS_B / df_B$
$SS_{AB} = m \sum_j \sum_i (\bar{x}_{ij} - \bar{x}_i - \bar{x}_j + \bar{x})^2$	$df_{AB} = (r-1)(c-1)$	$MS_{AB} = SS_{AB} / df_{AB}$
$SS_W = \sum_k \sum_j \sum_i (x_{ijk} - \bar{x}_{ij})^2$	$df_W = n - rc$	$MS_W = SS_W / df_W$

$$SS_T = SS_A + SS_B + SS_C + SS_{AB} + SS_{AC} + SS_{BC} + SS_{ABC} + SS_W$$

a - количество уровней фактора A,

b - количество уровней фактора B,

c - количество уровней фактора B,

$\bar{x}_{ijk}$ - среднее значение наблюдений, лежащих на пересечении уровней i, j и k факторов A, B и C соответственно,

x_{ijkt} - t-е наблюдение на пересечении уровней i, j и k факторов A, B и C соответственно.[18]

$SS_T = \sum_l \sum_k \sum_j \sum_i (x_{ijkl} - \bar{x})^2$	$df_T = n - 1$	$MS_T = SS_T / df_T$
$SS_A = mbc \sum_i (\bar{x}_i - \bar{x})^2$	$df_A = a - 1$	$MS_A = SS_A / df_A$
$SS_B = mac \sum_j (\bar{x}_j - \bar{x})^2$	$df_B = b - 1$	$MS_B = SS_B / df_B$
$SS_C = mab \sum_k (\bar{x}_k - \bar{x})^2$	$df_C = c - 1$	$MS_C = SS_C / df_C$
$SS_{AB} = mc \sum_j \sum_i (\bar{x}_{ij} - \bar{x}_i - \bar{x}_j + \bar{x})^2$	$df_{AB} = (a-1)(b-1)$	$MS_{AB} = SS_{AB} / df_{AB}$
$SS_{AC} = mb \sum_k \sum_i (\bar{x}_{ik} - \bar{x}_i - \bar{x}_k + \bar{x})^2$	$df_{AC} = (a-1)(c-1)$	$MS_{AC} = SS_{AC} / df_{AC}$
$SS_{BC} = ma \sum_k \sum_j (\bar{x}_{jk} - \bar{x}_j - \bar{x}_k + \bar{x})^2$	$df_{BC} = (b-1)(c-1)$	$MS_{BC} = SS_{BC} / df_{BC}$
$SS_{ABC} = m \sum_k \sum_j \sum_i (\bar{x}_{ijk} - \bar{x}_{ij} - \bar{x}_{ik} - \bar{x}_{jk} + \bar{x}_i + \bar{x}_j + \bar{x}_k - \bar{x})^2$	$df_{ABC} = (a-1)(b-1)(c-1)$	$MS_{ABC} = SS_{ABC} / df_{ABC}$
$SS_W = \sum_l \sum_k \sum_j \sum_i (x_{ijkl} - \bar{x}_{ijk})^2$	$df_W = n - abc$	$MS_W = SS_W / df_W$

Результаты применения метода контрастов

Взаимодействие факторов *филиал* и *категория* вдоль уровней фактора *сезон*

В таблицах приняты следующие обозначения:

б - БАДы,

п - предметы личной гигиены,

л - лекарственные средства,

1, 2, 3 - номера филиалов

1. Зима

contrast	estimate	SE	df	t.ratio	p.value
бл_12	8.241283	13.56965	2374.79	0.607	0.5437
бл_23	14.321032	13.56965	2374.79	1.055	0.2914
бл_13	22.562315	13.56965	2374.79	1.663	0.0965
бп_12	44.285471	13.56965	2374.79	3.264	0.0011
бп_23	-27.278953	13.56965	2374.79	-2.010	0.0445
бп_13	17.006517	13.56965	2374.79	1.253	0.2102
лп_12	36.044188	13.56965	2374.79	2.656	0.0080
лп_23	-41.599985	13.56965	2374.79	-3.066	0.0022
лп_13	5.555797	13.56965	2374.79	0.409	0.6823

2. Весна

contrast	estimate	SE	df	t.ratio	p.value
бл_12	47.7434416	13.56965	2374.79	3.518	0.0004
бл_23	-19.0849007	13.56965	2374.79	-1.406	0.1597
бл_13	28.6585409	13.56965	2374.79	2.112	0.0348
бп_12	0.3325005	13.56965	2374.79	0.025	0.9805
бп_23	-1.0007212	13.56965	2374.79	-0.074	0.9412
бп_13	-0.6682206	13.56965	2374.79	-0.049	0.9607
лп_12	-47.4109411	13.56965	2374.79	-3.494	0.0005
лп_23	18.0841796	13.56965	2374.79	1.333	0.1828
лп_13	29.3267615	13.56965	2374.79	2.161	0.0308

3. Лето

contrast	estimate	SE	df	t.ratio	p.value
бл_12	6.139603	13.56965	2374.79	0.452	0.6510
бл_23	-16.891741	13.56965	2374.79	-1.245	0.2133
бл_13	-10.752137	13.56965	2374.79	-0.792	0.4282
бп_12	31.700897	13.56965	2374.79	2.336	0.0196
бп_23	-48.282438	13.56965	2374.79	-3.558	0.0004
бп_13	-16.581541	13.56965	2374.79	-1.222	0.2218
лп_12	25.561294	13.56965	2374.79	1.884	0.0597
лп_23	-31.390697	13.56965	2374.79	-2.313	0.0208
лп_13	5.829403	13.56965	2374.79	0.430	0.6675

4. Осень

contrast	estimate	SE	df	t.ratio	p.value
бл_12	-5.893222	13.56965	2374.79	-0.434	0.6641
бл_23	-8.364916	13.56965	2374.79	-0.616	0.5377
бл_13	-14.258137	13.56965	2374.79	-1.051	0.2935
бп_12	34.951803	13.56965	2374.79	2.576	0.0101
бп_23	-51.481294	13.56965	2374.79	-3.794	0.0002
бп_13	-16.529491	13.56965	2374.79	-1.218	0.2233
лп_12	40.845024	13.56965	2374.79	3.010	0.0026
лп_23	-43.116378	13.56965	2374.79	-3.177	0.0015
лп_13	2.271354	13.56965	2374.79	0.167	0.8671

Взаимодействие факторов *сезон* и *категория* вдоль уровней фактора *филиал*

з - зима,

в - весна,

л - лето,

о - осень

1. Филиал 1

contrast	estimate	SE	df	t.ratio	p.value
бл_зв	-98.4121436	13.46931	2668.76	-7.306	<.0001
бп_зв	-47.4259847	13.46931	2668.76	-3.521	0.0004
лп_зв	50.9861588	13.46931	2668.76	3.785	0.0002
бл_вл	51.7344763	13.46931	2668.76	3.841	0.0001
бп_вл	62.0682288	13.46931	2668.76	4.608	<.0001
лп_вл	10.3337525	13.46931	2668.76	0.767	0.4430
бл_ло	-0.2203993	13.46931	2668.76	-0.016	0.9869
бп_ло	-113.5164987	13.46931	2668.76	-8.428	<.0001
лп_ло	-113.2960994	13.46931	2668.76	-8.411	<.0001
бл_оз	-46.8980665	13.46931	2668.76	-3.482	0.0005
бп_оз	-98.8742546	13.46931	2668.76	-7.341	<.0001
лп_оз	-51.9761881	13.46931	2668.76	-3.859	0.0001

2. Филиал 2

contrast	estimate	SE	df	t.ratio	p.value
бл_зв	-58.90998	13.46931	2668.76	-4.374	<.0001
бп_зв	-91.37896	13.46931	2668.76	-6.784	<.0001
лп_зв	-32.46897	13.46931	2668.76	-2.411	0.0160
бл_вл	10.13064	13.46931	2668.76	0.752	0.4520
бп_вл	93.43663	13.46931	2668.76	6.937	<.0001
лп_вл	83.30599	13.46931	2668.76	6.185	<.0001
бл_ло	-12.25322	13.46931	2668.76	-0.910	0.3631
бп_ло	-110.26559	13.46931	2668.76	-8.186	<.0001
лп_ло	-98.01237	13.46931	2668.76	-7.277	<.0001
бл_оз	-61.03257	13.46931	2668.76	-4.531	<.0001
бп_оз	-108.20792	13.46931	2668.76	-8.034	<.0001
лп_оз	-47.17535	13.46931	2668.76	-3.502	0.0005

3. Филиал 3

contrast	estimate	SE	df	t.ratio	p.value
бл_зв	-92.315918	13.46931	2668.76	-6.854	<.0001
бп_зв	-65.100723	13.46931	2668.76	-4.833	<.0001
лп_зв	27.215195	13.46931	2668.76	2.021	0.0434
бл_вл	12.323798	13.46931	2668.76	0.915	0.3603
бп_вл	46.154909	13.46931	2668.76	3.427	0.0006
лп_вл	33.831111	13.46931	2668.76	2.512	0.0121
бл_ло	-3.726399	13.46931	2668.76	-0.277	0.7821
бп_ло	-113.464449	13.46931	2668.76	-8.424	<.0001
лп_ло	-109.738050	13.46931	2668.76	-8.147	<.0001
бл_оз	-83.718519	13.46931	2668.76	-6.216	<.0001
бп_оз	-132.410263	13.46931	2668.76	-9.831	<.0001
лп_оз	-48.691744	13.46931	2668.76	-3.615	0.0003

В таблицах представлены p-значения метода контрастов. Красным цветом выделены те контрасты, которые по результатам применения метода были признаны значимыми.

Программная реализация в статистическом пакете R

Двухфакторный дисперсионный анализ

```
#подключаем библиотеки
library(reshape2)
library(pastecs)
library(nlme)
library(ggplot2)
library(nortest)
#library(car)
library(afex)
library(GAD)

Data<-read.delim("D:/Users/Sasha/СПбГУ/Диплом/
Two-way ANOVA.txt", header = TRUE)

#проверка на выполнение исходных предположений
y1=Data$зима_19
y2=Data$зима_15
y3=Data$зима_B
y4=Data$весна_19
y5=Data$весна_15
y6=Data$весна_B
y7=Data$лето_19
y8=Data$лето_15
y9=Data$лето_B
y10=Data$осень_19
y11=Data$осень_15
y12=Data$осень_B


x=c(var(y1),var(y2),var(y3),var(y4),var(y5),var(y6),
var(y7),var(y8),var(y9),var(y10),var(y11),var(y12))
cochran.test(x, rep(100,12))
```

```
cochran.test(x, rep(100,12), inlying=TRUE)


#проверка на нормальность
shapiro.test(y1) #p>0.05 - нулевая гипотеза не отвергается
shapiro.test(y2)
shapiro.test(y3)
shapiro.test(y4)
shapiro.test(y5)
shapiro.test(y6)
shapiro.test(y7)
shapiro.test(y8)
shapiro.test(y9)
shapiro.test(y10)
shapiro.test(y11)
shapiro.test(y12)

longData <-melt(Data, id = "наименование",
measured = c( "зима_19","зима_15", "зима_B", "весна_19",
"весна_15", "весна_B", "лето_19", "лето_15","лето_B",
"осень_19", "осень_15", "осень_B"))
names(longData)<-c("наименование", "группы", "доход")

longData$сезон<-gl(4, 100, labels = c("зима", "весна",
"лето","осень"))
longData$филиал<-gl(3, 400, 1200, labels = c("Филиал 1",
"Филиал 2", "Филиал 3"))

p <- ggplot(longData, aes(factor(сезон), доход))
p + geom_boxplot(aes(fill = factor(филиал)))+xlab("Сезон")
+ylab("Доход")+guides(fill = guide_legend(title = "Филиал"))

fit_all <- aov_ez("наименование","доход",longData,
within=c("сезон","филиал"))
```

```
summary(fit_all)

#метод контрастов
#сезон
ref1 <- lsmeans(fit_all,specs = c("сезон"))
ZVOvsL<-c(1, 1,-3,1)
ZVvsO<-c(1, 1, 0,-2)
ZvsV<-c(1, -1, 0,0)
summary(contrast(ref1,list(Все_лето=ZVOvsL,зимавесна_осень=ZVvsO,
зима_весна=ZvsV)))
seasonBar <-ggplot(longData,aes(сезон, доход))
seasonBar +stat_summary(fun.y = mean, geom ="bar",
fill ="White",colour ="Black")+
stat_summary(fun.data = mean_cl_boot, geom ="pointrange")+
labs(x ="Сезон", y ="Доход")

#филиал
shopBar <-ggplot(longData,aes(филиал, доход))
shopBar +stat_summary(fun.y = mean, geom ="bar", fill ="White",
colour ="Black")+stat_summary(fun.data = mean_cl_boot,
geom ="pointrange")+
labs(x ="Филиал", y ="Доход")

#сезон:филиал
incomeInt <-ggplot(longData,aes(сезон, доход, colour = филиал))
incomeInt +stat_summary(fun.y = mean, geom ="point")+
stat_summary(fun.y = mean, geom ="line",aes(group= филиал))+
stat_summary(fun.data = mean_cl_boot, geom ="errorbar", width =0.2)+
labs(x ="Сезон",y ="Доход", colour ="Филиал")
```

Трехфакторный дисперсионный анализ

```
#подключаем библиотеки
library(reshape2)
library(pastecs)
library(ez)
library(nlme)
library(ggplot2)
library(Rmisc)
library(nortest)
library(car)
library(caret)
library(e1071)
library(afex)

Data<-read.delim("D:/Users/Sasha/СПбГУ/Диплом
/Three-way ANOVA.txt", header = TRUE)
Data=head(Data,n=300)
Data=Data[1:14]

y1=Data$зима_19[Data$категория=="БАДы"]
y2=Data$зима_15[Data$категория=="БАДы"]
y3=Data$зима_B[Data$категория=="БАДы"]
y4=Data$весна_19[Data$категория=="БАДы"]
y5=Data$весна_15[Data$категория=="БАДы"]
y6=Data$весна_B[Data$категория=="БАДы"]
y7=Data$лето_19[Data$категория=="БАДы"]
y8=Data$лето_15[Data$категория=="БАДы"]
y9=Data$лето_B[Data$категория=="БАДы"]
y10=Data$осень_19[Data$категория=="БАДы"]
y11=Data$осень_15[Data$категория=="БАДы"]
y12=Data$осень_B[Data$категория=="БАДы"]
y13=Data$зима_19[Data$категория=="ЛС"]
y14=Data$зима_15[Data$категория=="ЛС"]
y15=Data$зима_B[Data$категория=="ЛС"]
```

```
y16=Data$весна_19[Data$категория=="ЛС"]
y17=Data$весна_15[Data$категория=="ЛС"]
y18=Data$весна_В[Data$категория=="ЛС"]
y19=Data$лето_19[Data$категория=="ЛС"]
y20=Data$лето_15[Data$категория=="ЛС"]
y21=Data$лето_В[Data$категория=="ЛС"]
y22=Data$осень_19[Data$категория=="ЛС"]
y23=Data$осень_15[Data$категория=="ЛС"]
y24=Data$осень_В[Data$категория=="ЛС"]
y25=Data$зима_19[Data$категория=="ПЛГ"]
y26=Data$зима_15[Data$категория=="ПЛГ"]
y27=Data$зима_В[Data$категория=="ПЛГ"]
y28=Data$весна_19[Data$категория=="ПЛГ"]
y29=Data$весна_15[Data$категория=="ПЛГ"]
y30=Data$весна_В[Data$категория=="ПЛГ"]
y31=Data$лето_19[Data$категория=="ПЛГ"]
y32=Data$лето_15[Data$категория=="ПЛГ"]
y33=Data$лето_В[Data$категория=="ПЛГ"]
y34=Data$осень_19[Data$категория=="ПЛГ"]
y35=Data$осень_15[Data$категория=="ПЛГ"]
y36=Data$осень_В[Data$категория=="ПЛГ"]

#проверка на гомоскедастичность
x=c(var(y1),var(y2),var(y3),var(y4),var(y5),var(y6),
var(y7),var(y8),var(y9),var(y10),var(y11),var(y12),
var(y13),var(y14),var(y15),var(y16),var(y17),var(y18),
var(y19),var(y20),var(y21),var(y22),var(y23),var(y24),
var(y25),var(y26),var(y27),var(y28),var(y29),var(y30),
var(y31),var(y32),var(y33),var(y34),var(y35),var(y36))

cochran.test(x, rep(100,36))
#проверка на нормальность
shapiro.test(y1) #p>0.05 - нулевая гипотеза не отвергается
shapiro.test(y2)
```

```
shapiro.test(y3)
shapiro.test(y4)
shapiro.test(y5)
shapiro.test(y6)
shapiro.test(y7)
shapiro.test(y8)
shapiro.test(y9)
shapiro.test(y10)
shapiro.test(y11)
shapiro.test(y12)
shapiro.test(y13)
shapiro.test(y14)
shapiro.test(y15)
shapiro.test(y16)
shapiro.test(y17)
shapiro.test(y18)
shapiro.test(y19)
shapiro.test(y20)
shapiro.test(y21)
shapiro.test(y22)
shapiro.test(y23)
shapiro.test(y24)
shapiro.test(y25)
shapiro.test(y26)
shapiro.test(y27)
shapiro.test(y28)
shapiro.test(y29)
shapiro.test(y30)
shapiro.test(y31)
shapiro.test(y32)
shapiro.test(y33)
shapiro.test(y34)
shapiro.test(y35)
shapiro.test(y36)
```

```
Data1<-melt(Data, id = c("наименование","категория"),
measured = c("зима_19","весна_19", "лето_19", "осень_19",
"зима_15", "весна_15","лето_15","осень_15","зима_В",
"весна_В","лето_В","осень_В"))
names(Data1)<-c("наименование", "категория", "группы", "доход")
Data1$сезон<-gl(4, 300, 3600, labels
= c("зима", "весна", "лето", "осень"))
Data1$филиал<-gl(3, 1200, 3600, labels
= c("филиал 1", "филиал 2", "филиал 3"))

#рисуем график
p <- ggplot(Data1, aes(factor(сезон), доход))
p + geom_boxplot(aes(fill = factor(филиал)))+
 facet_grid(. ~ категория)+xlab("Сезон")+
 ylab("Доход")+
guides(fill = guide_legend(title = "Филиал"))

fit_all <- aov_ez("наименование","доход",Data1,
between=c("категория"),within=c("филиал","сезон"))
summary(fit_all)

#Главный эффект сезон&филиал&категория
g71<-ggplot(Data1,aes(сезон,доход,colour=филиал))
g71+stat_summary(fun.y=mean,geom="point")+
stat_summary(fun.y=mean,geom="line",aes(group=филиал))+
stat_summary(fun.data=mean_cl_boot,geom="errorbar",width=0.2)+
labs(x="Сезон",y="Доход",colour="Филиал")+
scale_y_continuous(limits=c(0,600))+facet_wrap(~категория)
#1 ф-с
Data2=Data1[order(Data1$категория),]#категория
Data2=Data1[Data1$категория=='БАДы',]#БАДы
fit_all_2 <- aov_ez("наименование","доход",Data2,
within=c("филиал", "сезон"))
```

```
summary(fit_all_2)
#2
Data2=Data1[Data1$категория=='ЛС',]#ЛС
fit_all_2 <- aov_ez("наименование","доход",Data2,
within=c("филиал", "сезон"))
summary(fit_all_2)

ref1 <- lsmeans(fit_all,~сезон|категория)
summary(contrast(ref1,method="pairwise"))
EfSeason=summarySE(Data2, measurevar="доход",
groupvars=c("сезон"))
g2<-ggplot(EfSeason, aes(x=сезон, y=доход))
g2 + geom_bar(position=position_dodge(),
stat="identity",colour="black", fill="white", size=.3)
+ geom_errorbar(aes(ymin=доход-se, ymax=доход+se), width=.1)
+expand_limits(y = c(0, 100))

ref1 <- lsmeans(fit_all,~филиал|категория)
summary(contrast(ref1,method="pairwise"))
EfShop=summarySE(Data2, measurevar="доход", groupvars=c("филиал"))
g3<-ggplot(EfShop, aes(x=филиал, y=доход))
g3 + geom_bar(position=position_dodge(),
stat="identity",colour="black", fill="white", size=.3)
+ geom_errorbar(aes(ymin=доход-se, ymax=доход+se), width=.1)
+expand_limits(y = c(0, 100))

#3
Data2=Data1[Data1$категория=='ПЛГ',]#ПЛГ
fit_all_2 <- aov_ez("наименование",
"доход",Data2,within=c("филиал", "сезон"))
summary(fit_all_2)

#контрасты
#ПЛГ
```

```
ref1 <- lsmeans(fit_all,specs = c("сезон","филиал","категория"))
t1=c(0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0,
1,-1,0,0, -1,1,0,0, 0,0,0,0)#з-в,1-2
t2=c(0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0,
1,-1,0,0, 0,0,0,0, -1,1,0,0)#з-в,1-3
t3=c(0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0,
0,1,-1,0, 0,-1,1,0, 0,0,0,0)#в-л,1-2
t4=c(0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0,
0,1,-1,0, 0,0,0,0, 0,-1,1,0)#в-л,1-3
t5=c(0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0, 1,-1,0,0, -1,1,0,0)#з-в,2-3
t6=c(0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0, 0,1,-1,0, 0,-1,1,0)#в-л,2-3
summary(contrast(ref1,list(з_в_1_2=t1,з_в_1_3=t2,в_л_1_2=t3,
в_л_1_3=t4,з_в_2_3=t5,в_л_2_3=t6)))

#БАДы
t1=c(0,0,0,0, 0,-1,1,0, 0,1,-1,0, 0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0)#в-л,2-3
t2=c(0,-1,1,0, 0,0,0,0, 0,1,-1,0, 0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0)#в-л,1-3
summary(contrast(ref1,list(в_л_2_3=t1, в_л_1_3=t2)))
#ЛС
Data2=Data1[Data1$категория=='ЛС',]
fit_all_2 <- aov_ez("наименование","доход",Data2,within="сезон")
summary(fit_all_2)

g72<-ggplot(Data1,aes(категория,доход,colour=филиал))
g72+stat_summary(fun.y=mean,geom="point")
+stat_summary(fun.y=mean,geom="line",aes(group=филиал))
+stat_summary(fun.data=mean_cl_boot,geom="errorbar",width=0.2)
+labs(x="Категория",y="Доход",colour="Филиал")
+scale_y_continuous(limits=c(0,600))+facet_wrap(~сезон)
#4 ф-к
```

```
Data2=Data1[Data1$сезон=='зима',]
fit_all_2 <- aov_ez("наименование","доход",
Data2,within="филиал", between="категория")
summary(fit_all_2)
#5
Data2=Data1[Data1$сезон=='весна',]
fit_all_2 <- aov_ez("наименование","доход",
Data2,within="филиал", between="категория")
summary(fit_all_2)
#6
Data2=Data1[Data1$сезон=='лето',]
fit_all_2 <- aov_ez("наименование","доход",
Data2,within="филиал", between="категория")
summary(fit_all_2)
#7
Data2=Data1[Data1$сезон=='осень',]
fit_all_2 <- aov_ez("наименование","доход",
Data2,within="филиал", between="категория")
summary(fit_all_2)
ref1 <- lsmeans(fit_all,specs = c("филиал","категория","сезон"))
#зима
t1=c(1,-1,0, -1,1,0, 0,0,0, 0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0)
t2=c(0,1,-1, 0,-1,1, 0,0,0, 0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0)
t3=c(1,0,-1, -1,0,1, 0,0,0, 0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0)
t4=c(1,-1,0, 0,0,0, -1,1,0, 0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0)
t5=c(0,1,-1, 0,0,0, 0,-1,1, 0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0)
t6=c(1,0,-1, 0,0,0, -1,0,1, 0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0)
t7=c(0,0,0, 1,-1,0, -1,1,0, 0,0,0,0,0,0,0,0,0,
```

```
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0)
t8=c(0,0,0, 0,1,-1, 0,-1,1, 0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0)
t9=c(0,0,0, -1,0,1, 1,0,-1, 0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0)
summary(contrast(ref1,list(бл_12=t1,бл_23=t2,бл_13=t3,
бп_12=t4,бп_23=t5,бп_13=t6,лп_12=t7,лп_23=t8,лп_13=t9)))
#весна
t1=c(0,0,0,0,0,0,0,0,0, 1,-1,0, -1,1,0, 0,0,0,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0)
t2=c(0,0,0,0,0,0,0,0,0, 0,1,-1, 0,-1,1, 0,0,0,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0)
t3=c(0,0,0,0,0,0,0,0,0, 1,0,-1, -1,0,1, 0,0,0,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0)
t4=c(0,0,0,0,0,0,0,0,0, 1,-1,0, 0,0,0, -1,1,0,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0)
t5=c(0,0,0,0,0,0,0,0,0, 0,1,-1, 0,0,0, 0,-1,1,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0)
t6=c(0,0,0,0,0,0,0,0,0, 1,0,-1, 0,0,0, -1,0,1,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0)
t7=c(0,0,0,0,0,0,0,0,0, 0,0,0, 1,-1,0, -1,1,0,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0)
t8=c(0,0,0,0,0,0,0,0,0, 0,0,0, 0,1,-1, 0,-1,1,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0)
t9=c(0,0,0,0,0,0,0,0,0, 0,0,0, -1,0,1, 1,0,-1,
0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0)
summary(contrast(ref1,list(бл_12=t1,бл_23=t2,бл_13=t3,
бп_12=t4,бп_23=t5,бп_13=t6,лп_12=t7,лп_23=t8,лп_13=t9)))
#лето
t1=c(0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,
1,-1,0, -1,1,0, 0,0,0, 0,0,0,0,0,0,0,0,0)
t2=c(0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,
0,1,-1, 0,-1,1, 0,0,0, 0,0,0,0,0,0,0,0,0)
t3=c(0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,
```

```
1,0,-1, -1,0,1, 0,0,0, 0,0,0,0,0,0,0,0,0)
t4=c(0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,
1,-1,0, 0,0,0, -1,1,0, 0,0,0,0,0,0,0,0,0)
t5=c(0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,
0,1,-1, 0,0,0, 0,-1,1, 0,0,0,0,0,0,0,0,0)
t6=c(0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,
1,0,-1, 0,0,0, -1,0,1, 0,0,0,0,0,0,0,0,0)
t7=c(0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,
0,0,0, 1,-1,0, -1,1,0, 0,0,0,0,0,0,0,0,0)
t8=c(0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,
0,0,0, 0,1,-1, 0,-1,1, 0,0,0,0,0,0,0,0,0)
t9=c(0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,
0,0,0, -1,0,1, 1,0,-1, 0,0,0,0,0,0,0,0,0)
summary(contrast(ref1,list(бл_12=t1,бл_23=t2,бл_13=t3,
бп_12=t4,бп_23=t5,бп_13=t6,лп_12=t7,лп_23=t8,лп_13=t9)))
#осень
t1=c(0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0, 1,-1,0, -1,1,0, 0,0,0)
t2=c(0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0, 0,1,-1, 0,-1,1, 0,0,0)
t3=c(0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0, 1,0,-1, -1,0,1, 0,0,0)
t4=c(0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0, 1,-1,0, 0,0,0, -1,1,0)
t5=c(0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0, 0,1,-1, 0,0,0, 0,-1,1)
t6=c(0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0, 1,0,-1, 0,0,0, -1,0,1)
t7=c(0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0, 0,0,0, 1,-1,0, -1,1,0)
t8=c(0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0, 0,0,0, 0,1,-1, 0,-1,1)
t9=c(0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0, 0,0,0, -1,0,1, 1,0,-1)
```

```
summary(contrast(ref1,list(бл_12=t1,бл_23=t2,бл_13=t3,
бп_12=t4,бп_23=t5,бп_13=t6,лп_12=t7,лп_23=t8,лп_13=t9)))

g73<-ggplot(Data1,aes(сезон,доход,colour=категория))
g73+stat_summary(fun.y=mean,geom="point")
+stat_summary(fun.y=mean,geom="line",aes(group=категория))
+stat_summary(fun.data=mean_cl_boot,geom="errorbar",width=0.2)
+labs(x="Сезон",y="Доход",colour="Категория")
+scale_y_continuous(limits=c(0,600))+facet_wrap(~филиал)
#8 к-с
Data2=Data1[Data1$филиал=='филиал 1',]
fit_all_2 <- aov_ez("наименование","доход",Data2,
within="сезон", between="категория")
summary(fit_all_2)
#9
Data2=Data1[Data1$филиал=='филиал 2',]
fit_all_2 <- aov_ez("наименование","доход",Data2,
within="сезон", between="категория")
summary(fit_all_2)
#10
Data2=Data1[Data1$филиал=='филиал 3',]
fit_all_2 <- aov_ez("наименование","доход",Data2,
within="сезон", between="категория")
summary(fit_all_2)
ref1 <- lsmeans(fit_all,specs = c("категория","сезон","филиал"))
#филиал 1
t1=c(1,-1,0, -1,1,0, 0,0,0, 0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t2=c(1,0,-1, -1,0,1, 0,0,0, 0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t3=c(0,1,-1, 0,-1,1, 0,0,0, 0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t4=c(0,0,0, 1,-1,0, -1,1,0, 0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
```

```
t5=c(0,0,0, 1,0,-1, -1,0,1, 0,0,0, 0,0,0,0,0,0,0,
0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t6=c(0,0,0, 0,1,-1, 0,-1,1, 0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t7=c(0,0,0, 0,0,0, 1,-1,0, -1,1,0,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t8=c(0,0,0, 0,0,0, 1,0,-1, -1,0,1,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t9=c(0,0,0, 0,0,0, 0,1,-1, 0,-1,1,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t10=c(1,-1,0, 0,0,0, 0,0,0, -1,1,0,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t11=c(1,0,-1, 0,0,0, 0,0,0, -1,0,1,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t12=c(0,1,-1, 0,0,0, 0,0,0, 0,-1,1,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
summary(contrast(ref1,list(бл_зв=t1,бп_зв=t2,лп_зв=t3,
бл_вл=t4,бп_вл=t5,лп_вл=t6,бл_ло=t7,бп_ло=t8,лп_ло=t9,
абл_оз=t10,бп_оз=t11,лп_оз=t12)))
t1=c(0,0,0,0,0,0,0,0,0,0,0,0,
1,-1,0, -1,1,0, 0,0,0, 0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t2=c(0,0,0,0,0,0,0,0,0,0,0,0,
1,0,-1, -1,0,1, 0,0,0, 0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t3=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,1,-1, 0,-1,1, 0,0,0, 0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t4=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0, 1,-1,0, -1,1,0, 0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t5=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0, 1,0,-1, -1,0,1, 0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t6=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0, 0,1,-1, 0,-1,1, 0,0,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t7=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0, 0,0,0, 1,-1,0, -1,1,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t8=c(0,0,0,0,0,0,0,0,0,0,0,0,
```

```
0,0,0, 0,0,0, 1,0,-1, -1,0,1, 0,0,0,0,0,0,0,0,0,0,0,0)
t9=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0, 0,0,0, 0,1,-1, 0,-1,1, 0,0,0,0,0,0,0,0,0,0,0,0)
t10=c(0,0,0,0,0,0,0,0,0,0,0,0,
1,-1,0, 0,0,0, 0,0,0, -1,1,0, 0,0,0,0,0,0,0,0,0,0,0,0)
t11=c(0,0,0,0,0,0,0,0,0,0,0,0,
1,0,-1, 0,0,0, 0,0,0, -1,0,1, 0,0,0,0,0,0,0,0,0,0,0,0)
t12=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,1,-1, 0,0,0, 0,0,0, 0,-1,1, 0,0,0,0,0,0,0,0,0,0,0,0)
summary(contrast(ref1,list(бл_зв=t1,бп_зв=t2,лп_зв=t3,
бл_вл=t4,бп_вл=t5,лп_вл=t6,бл_ло=t7,бп_ло=t8,лп_ло=t9,
бл_оз=t10,бп_оз=t11,лп_оз=t12)))
t1=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0, 1,-1,0, -1,1,0, 0,0,0, 0,0,0)
t2=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0, 1,0,-1, -1,0,1, 0,0,0, 0,0,0)
t3=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0, 0,1,-1, 0,-1,1, 0,0,0, 0,0,0)
t4=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0, 1,-1,0, -1,1,0, 0,0,0)
t5=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0, 1,0,-1, -1,0,1, 0,0,0)
t6=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0, 0,1,-1, 0,-1,1, 0,0,0)
t7=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0, 0,0,0, 1,-1,0, -1,1,0)
t8=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0, 0,0,0, 1,0,-1, -1,0,1)
t9=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0, 0,0,0, 0,1,-1, 0,-1,1)
t10=c(0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0, 1,-1,0, 0,0,0, 0,0,0, -1,1,0)
t11=c(0,0,0,0,0,0,0,0,0,0,0,0,
 0,0,0,0,0,0,0,0,0,0,0,0, 1,0,-1, 0,0,0, 0,0,0, -1,0,1)
```

```
t12=c(0,0,0,0,0,0,0,0,0,0,0,0,
 0,0,0,0,0,0,0,0,0,0,0,0, 0,1,-1, 0,0,0, 0,0,0, 0,-1,1)
summary(contrast(ref1,list(бл_зв=t1,бп_зв=t2,лп_зв=t3,
бл_вл=t4,бп_вл=t5,лп_вл=t6,бл_ло=t7,бп_ло=t8,лп_ло=t9,
бл_оз=t10,бп_оз=t11,лп_оз=t12)))
```