МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
«КРЫМСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ имени В. И. ВЕРНАДСКОГО»
(ФГАОУ ВО «КФУ им. В. И. Вернадского»)
Таврическая академия (структурное подразделение)
Факультет математики и информатики
Кафедра информатики

Выявление закономерностей во множествах данных на основе моделей дисперсионного анализа

Выпускная квалификационная работа

Обучающегося 6 курса

Направления подготовки 01.04.02.Прикладная математика и информатика

Форма обучения очная

Научный руководитель,
доцент кафедры информатики
кандидат физико-математических наук

К ЗАЩИТЕ ДОПУСКАЮ:

Заведующий кафедрой,
доктор физико-математических наук,
профессор

Симферополь, 2016

Оглавление

Введение....4
1.1. Общая задача дисперсионного анализа....6
1.2. Классический дисперсионный анализ Фишера....8
1.3. Общая линейная модель....11
1.4. Одномерный дисперсионный анализ....14
1.4.1. Однофакторный *ANOVA*-анализ....14
1.4.2. Апостериорные критерии и контрасты....18
1.4.3. Однофакторный ранговый анализ....21
1.4.4. Двухфакторный *ANOVA*-анализ....23
1.4.5. Дисперсионный анализ с повторными измерениями....25
1.5. Планирование эксперимента....26
2. Многомерный дисперсионный анализ....29
2.1. Многомерная обобщенная линейная модель....29
2.2. Критерии проверки многомерной гипотезы....31
2.3. Отклонение от основных предположений....33
3. Задачи исследований с применением модели дисперсионного анализа....34
3.1. Точный тест Фишера....34
3.2. Проверка значимости модели регрессии....36
3.3. Модели дисперсионного анализа в исследовании сводной таблицы успеваемости....37
3.3.1. Одномерная однофакторная модель анализа итогов промежуточной аттестации....38
3.3.2. Одномерная двухфакторная модель анализа итогов промежуточной аттестации....43
3.3.3. Многомерная двухфакторная модель анализа итогов промежуточной аттестации....44
3.4. Анализ данных опроса об удовлетворенности программой....45
3.4.1. Одномерная модель удовлетворенности....47
3.4.2. Одномерная модель полезности....48
3.4.3. Многомерная модель....49
Заключение....52
Список использованной литературы....53

Введение

Многомерный статистический анализ (МСА) представляет собой самостоятельное направление статистической теории. Интерес к нему связан с широкими возможностями в отображении и моделировании реальных объектов и явлений, имеющих многопризнаковую природу. Многопризнаковая природа различных классов объектов ведет к тому, что их анализ и отображение в моделях возможны при учете всех наиболее существенных характеристик.

Многомерный статистический анализ является логическим развитием традиционной статистики. Однако при анализе многомерных данных используют и методы, иногда не имеющие четкой теоретической трактовки в терминах проверки гипотез, построения доверительных интервалов и т.д.

В основе многомерного анализа лежит предварительная обработка данных, включающая описательную статистику, визуальные средства анализа и формулировку гипотез о законах распределения и статистических закономерностях. Главной же задачей является анализ структуры выборочных совокупностей и выявление закономерностей.

Одним из мощных и широко распространенных методов является *дисперсионный анализ*. Более того, он занимает особое место в статистических исследованиях, поскольку практически все модели анализа данных опираются на изучение вариации признаков и численные оценки этой вариации (дисперсии). В задачах поиска закономерностей в массивах данных дисперсионный анализ применяется для исследования влияния одной или нескольких качественных переменных (факторов) на одну (и более) количественную переменную (отклик).

Несмотря на глубокую проработанность теоретических обоснований и широкое использование в прикладных задачах, методы дисперсионного анализа остаются *актуальными* в современных исследованиях, использующих процедуры статистического анализа больших массивов данных.

Целью выпускной квалификационной работы является изучение методов и моделей дисперсионного анализа и их применение в анализе многомерных данных.

Задачи работы включают:

- систематическое изложение основных положений, задач и методов дисперсионного анализа и их использования на различных этапах статистической обработки данных;
- изучение процедур одномерного и многомерного дисперсионного анализа и их иллюстрация с помощью специально разработанных примеров;

- исследование прикладных задач выявления закономерностей в наборах данных с использованием моделей дисперсионного анализа и интерпретация полученных результатов.

Выпускная квалификационная работа состоит из трех разделов. В первом разделе дано систематическое изложение моделей и методов дисперсионного анализа: приведены общая задача дисперсионного анализа и общая линейная модель. Среди задач одномерного дисперсионного анализа рассмотрены подробно однофакторная и двухфакторная модели, модель анализа с повторными измерениями и основы однофакторного рангового анализа.

Второй раздел посвящен задачам многомерного дисперсионного анализа: приводится многомерная обобщенная линейная модель и критерии проверки многомерной гипотезы.

В третьем разделе рассмотрен ряд задач исследования с использованием различных моделей дисперсионного анализа на множестве данных об успеваемости обучающихся, а также по результатам опроса обучающихся для выявления факторов, влияющих на удовлетворенность содержанием программы обучения.

Это самостоятельное исследование, результаты которого могут рассматриваться как иллюстрирующий материал по применению моделей дисперсионного анализа для выявления закономерностей в наборах данных.

1. Модели и методы дисперсионного анализа

1.1. Общая задача дисперсионного анализа

Становление многомерного статистического анализа (МСА) как науки с собственной теоретической базой относится к началу XX века и связано с работами К. Пирсона и Ч. Спирмена, посвященными факторному анализу. Первоначально потребность в специальных методах обработки многомерных данных возникла в аналитической практике в области психологии, биологии, позднее – в медицине, военной промышленности, технике и экономике. Период становления и активного развития фундаментальных исследований в многомерном анализе пришелся на 40 - 60-е годы. Наибольший вклад в развитие теории МСА внесли С. Барт, Г. Кайзер, Т. Келли, Д. Максвелл, С. Р. Рао, Л. Р. Такер, Л. Л. Терстоун, Р. Хорст (теория и алгоритмы факторного анализа), Л. Гутман, Дж. Б. Краскал, Р. В. Хемминг (многомерное шкалирование), А. Барр, Р. Фишер, Т. В. Хейк (дискриминантный анализ), Р. Льюис, Р. Трионон, Р. Сокал, Р. Сибсон, У. Уильямс, Т. Танимото, М. Жамбю (кластерный анализ), С. Райт (путевой анализ) и другие ученые.

Дисперсионный анализ занимает особое место в статистических исследованиях. Практически все модели анализа данных опираются на анализ вариации признаков и численные оценки этой вариации – суммы квадратов отклонений и дисперсии (средние суммы квадратов).

Метод дисперсионного анализа связан с исследованием значимости (существенности) различия средних в разных группах наблюдений и проверкой соответствующих гипотез. И поскольку количественные оценки такого различия основаны на сравнении сумм квадратов отклонений от средних и выборочных дисперсий, то отсюда происходит название метода. Нулевая гипотеза формулируется как гипотеза о равенстве средних в нескольких группах наблюдений. Ее проверка основана на декомпозиции общей дисперсии выборки на отдельные компоненты, обусловленные влиянием факторов.

Сочетание *ANOVA* (*Analysis of Variances*) означает анализ вариаций (дисперсий). Эта аббревиатура была введена Тьюки (*J.W. Tukey*). Для этой группы методов характерно наличие одной зависимой переменной, на вариацию которой оказывают влияние одна или несколько независимых переменных – факторов. При этом факторы имеют неколичественный характер, и предполагается возможность их изменения в нескольких уровнях. Зависимая переменная обязана быть количественной и нормально распределенной.

Именно этот тип анализа был представлен в работах Р. А. Фишера, предложившего в 1925 году фундаментальную концепцию дисперсионного анализа. Первоначально это была группа статистических методов, разрабатываемых в связи с некоторыми экспериментальными

задачами в различных областях сельскохозяйственных и биологических исследований. Однако в силу универсальности применяемого аппарата область применения стала необычайно широкой. Дисперсионный анализ использует свойство аддитивности дисперсии случайной величины, обусловленной действием независимых факторов. Фишер впервые определил дисперсионный анализ как «*отделение дисперсии, приписываемой одной группе причин от дисперсии, приписываемой другим группам причин*» [11].

Для многомерного статистического анализа характерно сочетание *MANOVA* (*Multicriterian Analysis of Variances*). В этом случае предполагается наличие группы зависимых переменных, варьирующих под влиянием ряда факторов. В этом случае возрастает сложность вычислений, но содержание и основные принципы не меняются: проверяются *гипотезы о равенстве средних* в подгруппах, соответствующих плану эксперимента. Многомерный дисперсионный анализ отличается большой эффективностью, особенно в сочетании с ковариационным и дискриминантным анализом.

Общая задача дисперсионного анализа, согласно [2], включает несколько факторов, каждый из которых имеет не менее двух уровней. Целью анализа является влияние уровней факторов на вариацию изучаемой переменной. В основе каждой задачи лежат план эксперимента и модель дисперсионного анализа.

План эксперимента (правило выбора комбинаций факторов для фиксации значений зависимой переменной) позволяет при сравнительно малом числе измерений оценить влияние каждого из факторов на измеряемый количественный показатель. При пассивных наблюдениях под планом подразумевается надлежащий отбор групп из большой совокупности данных.

Модель дисперсионного анализа есть математическое соотношение, представляющее каждую переменную в виде суммы среднего значения и ошибки; среднее значение представляется в виде суммы генерального среднего и «эффекта» от каждого фактора и комбинаций факторов. Возникающие статистические задачи связаны с оценкой этих эффектов и проверкой гипотез о них.

В отличие от моделей множественной регрессии, независимые переменные (факторы) в дисперсионном анализе не обязаны быть количественными и нормально распределенными. Это требование налагается только на зависимые переменные при дополнительном условии однородности дисперсий внутри групп. При их нарушении используют непараметрические методы.

Приведем далее ряд примеров постановки исследовательских задач из различных областей, ориентированные на использование методов дисперсионного анализа.

В промышленности и технике [7]:

- влияние на относительные деформации ползучести бетона факторов влажности среды, водоцементного отношения, вида цемента, времени действия нагрузки и других;
- влияние на прочность металлического сплава состава смесей;

В сельскохозяйственной области:

- анализ влияния типов участков, сортов семян, внесенных удобрений на урожайность культур;
- анализ влияния технической оснащенности и агрономической культуры на количественные и качественные показатели в садоводстве и виноградарстве;

В психологии [10]:

- оценка влияния стажа работы на структуру личностных ценностей педагога (опросник ценностей Shalom H. Schwartz);
- оценка влияния клубных предпочтений на уровень ситуативной и личностной тревожности футбольного болельщика (шкала State-Trait Anxiety Inventory (STAI) Charles D. Spielberger);
- оценка влияния типа вирусного заболевания на психическое состояние человека (тест дифференцированной самооценки функционального состояния).

В биологии:

- влияние мест и среды обитания на характерные признаки видов организмов;

В медицине

- влияние методов подсчета числа лейкоцитов на количественный результат анализа крови;

1.2. Классический дисперсионный анализ Фишера

Рональд Эйлмер Фишер (*Ronald Aylmer Fisher*, 1890 – 1962) – английский теоретик математической статистики, биолог и генетик, удостоенный рыцарского звания.

По данным опроса среди ученых по поводу того, кто сильнее всего повлиял на биологию XX века, Р. А. Фишер был признан первым; в этом отношении его даже сравнивают с Чарльзом Дарвином, оказавшим наибольшее влияние на биологию XIX века. За годы работы на Ротамстедской опытной станции Фишер убедился в том, что традиционно используемый биологами статистический аппарат не соответствует условиям и потребностям исследований, и перенес акценты на критерии существенности и методы анализа малых выборок, предложил методы упрощения обработки статистических данных, заложил основы дисперсионного анализа.

По влиянию и вкладу в математическую статистику Р. Фишера сравнивают с Гауссом, поскольку он ввел в статистику такие методы и инструменты анализа, которые вывели ее на уровень фундаментальной

науки. В 1925 году он опубликовал свою первую книгу о статистических методах для научных работников (*Statistical Method For Research Worker*), которая стала стандартным справочным пособием для ученых во многих дисциплинах. В 1935 году была издана книга по планированию экспериментов (*The Design of Experiments*), которая также пользовалась заслуженным успехом. Что касается первой из упомянутых книг, то она неоднократно переиздавалась во всем мире, и только в 1958 году была издана на русском языке (12-е переработанное издание 1954 года) с весьма критичным предисловием о «буржуазной статистике» [12].

Классическая схема дисперсионного анализа для одной зависимой переменной и одного фактора, разработанная Фишером [12], предполагает наличие $n = k n'$ наблюдений зависимой нормально распределенной переменной *X*, разделенных на *k* групп (классов) в соответствии с *k* уровнями (градациями) фактора, по n' наблюдений в каждой группе. Рассматривается сумма квадратов отклонений значений переменной *X* от общего среднего $\bar{x}$, которая раскладывается на две доли – сумму квадратов отклонений групповых средних $\bar{x}_p \ (p = 1,2,\ldots,k)$ от общего среднего $\bar{x}$ и сумму квадратов отклонений наблюдений внутри групп от групповых средних $\bar{x}_p$:

$$\sum_{1}^{kn'} (x - \bar{x})^2 = k \sum_{1}^{n'} (\bar{x}_p - \bar{x})^2 + \sum_{1}^{kn'} (x - \bar{x}_p)^2 .$$

Это соотношение систематизирует таблица 1.1, которая задает формат итоговой таблицы дисперсионного анализа по схеме Фишера, используемый во всех современных прикладных пакетах.

Таблица 1.1. Таблица результатов дисперсионного анализа.

	Степени свободы	Суммы квадратов	Средние квадраты (дисперсии)
Внутри классов	$n'(k-1)$	$\sum_{1}^{kn'} (x - \bar{x}_p)^2$	$\frac{\sum_{1}^{kn'} (x - \bar{x}_p)^2}{n'(k-1)}$
Между классами	$n' - 1$	$k \sum_{1}^{n'} (\bar{x}_p - \bar{x})^2$	$\frac{k \sum_{1}^{n'} (\bar{x}_p - \bar{x})^2}{n' - 1}$
Итого	$n'k - 1$	$\sum_{1}^{kn'} (x - \bar{x})^2$	$\frac{\sum_{1}^{kn'} (x - \bar{x})^2}{n'k - 1}$

Проверка значимости различия сумм квадратов – компонент общей суммы квадратов выполняется на основе стандартного нормального

распределения. Статистика *z* формируется как половина разности логарифмов компонент общей суммы квадратов.

Фишер предложил способ получения независимых оценок двух дисперсий (межклассовой и внутриклассовой) и для проверки значимости различия использовал отношение дисперсий (средних квадратов). Статистика *F*, равная отношению межклассовой дисперсии к внутриклассовой, имеет *F*-распределение с соответствующими степенями свободы. Если рассчитанное по выборочным данным значение *F*-отношения превышает критическое значение с заданным уровнем значимости α, то нулевая гипотеза отклоняется.

Доказательство того, что величина F, равная отношению дисперсий, подчиняется F-распределению, вытекает из теоремы Кокрена [2, стр. 231].

Фишер сформулировал принципы многофакторного дисперсионного анализа и разработал связанные с ним задачи и методы организации активного эксперимента. В [11, стр. 191] приводится классическая задача «Дисперсионный анализ в полевом опыте», в которой собранные данные об урожае картофеля исследовались на наличие влияния типа внесенных удобрений на урожайность картофеля различных сортов: «Участок … был разделен на 36 делянок, на которых было размещено 12 сортов картофеля; три делянки каждого сорта были без всякой системы, т. е. в рандомизированном порядке разбросаны по всей площади участка. Каждая делянка была разделена на три части, одна из которых не получила никакого добавочного удобрения, в то время как две другие получили соответственно сульфат калия и хлорид калия».

На языке планирования это означает организацию полного двухфакторного эксперимента, в котором один фактор имеет 12 уровней, а второй 3 уровня. Итогом анализа стало заключение: «В данном опыте не обнаруживается различие в отзывчивости сортов на удобрение».

В [11] Фишер обосновал применимость дисперсионного анализа как самостоятельного статистического метода и привел его различные приложения в регрессионном и ковариационном анализе, технике планирования эксперимента, установлении различия групп и др.

Так, в регрессионном анализе для определения формы уравнения регрессии составляется таблица дисперсионного анализа, где источниками вариации зависимой переменной считаются вариация независимых переменных (выраженная суммой квадратов отклонений предсказанных регрессией значений от среднего, или *функцией регрессии*) и случайная вариация (сумма квадратов предсказанных и наблюдаемых значений зависимой переменной, или *отклонение от функции регрессии*). Отношение этих сумм квадратов выражает коэффициент детерминации, а его квадратный корень – корреляционное отношение. Существенность этих показателей обосновывает форму регрессии.

1.3. Общая линейная модель

Будем рассматривать *общую линейную модель* [2] как теоретическую основу дисперсионного, а также и регрессионного анализа. Пусть $Y_1,\ldots,Y_n$ – случайные величины, а $y_1,\ldots,y_n$ – их реализации (наблюдения). Предположим, что среднее значение *i*-й случайной величины линейно зависит от *p* параметров $\theta_1,\ldots,\theta_p$, так что

$$E(Y_i)=\theta_1 x_{1i}+\ldots+\theta_p x_{pi}+\varepsilon_i,\ i=1,\ldots,n, \qquad (1.1)$$

где $x_{1i},\ldots,x_{1p}$ – известные постоянные, определяемые в дисперсионном анализе матрицей плана, а в регрессионном – матрицей наблюдений. Каждое наблюдаемое значение можно записать в виде суммы

$$y_i=\theta_1 x_{1i}+\ldots+\theta_p x_{pi}+\varepsilon_i,\ i=1,\ldots,n,$$

где $\varepsilon_i,\ldots,\varepsilon_p$ – ошибки.

Общая линейная модель задается соотношениями (1.1) при дополнительных предположениях об ошибках:

$$E(\varepsilon_i)=0;\ V(\varepsilon_i)=\sigma^2;\ \operatorname{cov}(\varepsilon_i,\varepsilon_j)=0,\ i,j=1,\ldots,n,\ i\neq j, \qquad (1.2)$$

т. е. ошибки некоррелированы, имеют нулевые средние и одинаковые дисперсии.

Отметим, что при $x_{1i}\equiv 1$ (1.1) и (1.2) представляют модель множественной линейной регрессии.

Оценки $\hat{\theta}_1,\ldots,\hat{\theta}_p$ параметров $\theta_1,\ldots,\theta_p$ определяются методом наименьших квадратов путем минимизации суммы квадратов

$$SS=\sum_{i=1}^{n}\left(y_i-\theta_1 x_{1i}-\ldots-\theta_p x_{pi}\right)^2.$$

Значения оценок $\theta_1,\ldots,\theta_p$ линейно зависят от наблюдений. Если эти оценки определены однозначно (в противном случае налагаются дополнительные условия), то оценка наименьших квадратов для любой линейной функции параметров является той же линейной функцией от оценок самих параметров. В этом случае справедлива теорема Гаусса-Маркова.

Теорема. *Для общей линейной модели* (1.1), (1.2) *МНК-оценка* $\sum_{i=1}^{p} c_i \hat{\theta}_i$ *величины* $\sum_{i=1}^{p} c_i \theta_i$ *является несмещенной и обладает наименьшей дисперсией среди всех несмещенных оценок суммы* $\sum_{i=1}^{p} c_i \theta_i$, *линейных по* $y_1, \ldots, y_n$.

Помимо оценок параметров $\theta_1, \ldots, \theta_p$ должна быть получена оценка дисперсии ошибок σ^2 как средний квадрат ошибки:

$$MS_R \equiv s^2 = SS_R / \nu_R .$$

Здесь SS_R есть *остаточная сумма квадратов*:

$$SS_R = \sum_{i=1}^{n} \left(y_i - \hat{\theta}_1 x_{1i} - \ldots - \hat{\theta}_p x_{pi} \right)^2 ,$$

а ν_R есть *число степеней свободы*, которое определяется в дальнейшем для всех рассматриваемых случаев так, чтобы оценка дисперсии была несмещенной.

Общую линейную модель (1.1) с ограничениями (1.2) можно записать в матричных обозначениях следующим образом:

$$\tilde{y} = X' \tilde{\theta} + \tilde{\varepsilon}; \;\; E(\tilde{\varepsilon}) = 0, \; \mathrm{cov}(\tilde{\varepsilon}) = \sigma^2 I , \qquad (1.3)$$

где $\tilde{y} = (y_1, \ldots, y_n)'$, $\tilde{\theta} = (\theta_1, \ldots, \theta_p)$, $\tilde{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$,

$$X' = \begin{pmatrix} x_{11} & \ldots & x_{p1} \\ \ldots & \ldots & \ldots \\ x_{1n} & \ldots & x_{pn} \end{pmatrix}$$ – матрица плана.

Тогда минимизируемая сумма квадратов представляется в виде

$$SS = \left(\tilde{y} - X' \tilde{\theta} \right)' \left(\tilde{y} - X' \tilde{\theta} \right),$$

а оценки наименьших квадратов находятся как решения нормальных уравнений

$$\left(X X^{'}\right)\widetilde{\theta}=X\widetilde{y},$$

В случае, когда ранг *r* матрицы $X^{'}$ равен *p*, а значит, матрица $\left(X X^{'}\right)$ невырожденная это решение единственно: $\widetilde{\theta}=\left(X X^{'}\right)^{-1}X\widetilde{y}$ (это случай полного ранга). Ковариационная матрица для $\widetilde{\theta}$ вычисляется как $\operatorname{cov}\left(\widetilde{\theta}\right)=\sigma^{2}\left(X X^{'}\right)^{-1}$.

В случае неполного ранга при $r<p$ оценки неединственны. Остаточное число степеней свободы $\nu_{R}=n-r$, остаточная сумма квадратов принимает вид $SS_{R}=\left(\widetilde{y}-X^{'}\hat{\theta}\right)^{'}\left(\widetilde{y}-X^{'}\hat{\theta}\right)$, а *остаточный средний квадрат* равен значению

$$MS_{R}=\frac{1}{n-p}\left(\widetilde{y}-X^{'}\hat{\theta}\right)^{'}\left(\widetilde{y}-X^{'}\hat{\theta}\right).$$

В большинстве задач дисперсионного анализа *проверяемую гипотезу* можно записать в виде $H_{0}:\theta_{i_{1}}=\ldots=\theta_{i_{m}}=0$; это означает, что *m* из *p* параметров равны нулю. Если считать, что это последние *m* параметров, получим $H_{0}:\theta_{p-m+1}=\ldots=\theta_{p}=0$. При принятии гипотезы образуется усеченная модель

$$y_{i}=\theta_{1}x_{1i}+\ldots+\theta_{p-m}x_{p-m,i}+\varepsilon_{i},\ i=1,\ldots,n$$

Остаточная сумма квадратов при принятии гипотезы H_{0} есть

$$SS_{R}^{'}=\sum_{i=1}^{n}\left(y_{i}-\hat{\theta}_{1}x_{1i}-\ldots-\hat{\theta}_{p-m}x_{p-m,i}\right)^{2},$$

а $\nu_{R}^{'}$ – соответствующее число степеней свободы.

В этом случае статистика критерия правдоподобия имеет вид

$$F=\frac{(SS_{R}^{'}-SS_{R})/(\nu_{R}^{'}-\nu_{R})}{SS_{R}/\nu_{R}}. \qquad (1.4)$$

Величина $SS_H = SS'_R - SS_R$ называется *гипотетической суммой квадратов*, а $\nu_H = \nu'_R - \nu_R$ *– гипотетическим числом степеней свободы*. Статистика критерия *F* называется *F*-отношением.

Сумма SS'_R служит мерой того, насколько хорошо усеченная модель согласуется с наблюдениями. Сумма SS_H показывает, насколько хуже наблюдения аппроксимируются усеченной моделью, а *F* есть мера потерь при принятии гипотезы H_0. Чем больше *F*, тем хуже модель, поэтому при больших значениях *F* нужно отклонить гипотезу H_0.

Известно, что величина *F* распределена по закону Фишера $F(\nu'_R - \nu_R; \nu_R)$. Все рассматриваемые далее модели дисперсионного анализа можно записать в виде линейной модели.

1.4. Одномерный дисперсионный анализ

Одномерная модель дисперсионного анализа (*ANOVA*) предполагает изучение вариации *одной* зависимой переменной на нескольких подмножествах (группах) выборочной совокупности. Эта вариация может быть обусловлена вариацией одного или нескольких, возможно, неизвестных факторов.

1.4.1. Однофакторный *ANOVA*-анализ

Изучается влияние *одного* фактора на вариацию *одной* зависимой переменной *Y*. Такой тип анализа принято называть однофакторным *ANOVA* или *One-way ANOVA.*

Совокупность всех наблюдений разбивают на *I* групп, соответствующих уровням фактора, и вычисляют для каждой группы среднее значение и выборочную дисперсию.

Цель однофакторного *ANOVA*-анализа:

- установить, оказывает ли вариация фактора *A* (возможно, качественного) существенное влияние на средние значения в группах зависимой переменной *Y*;
- найти оптимальный уровень фактора.

Задача о влиянии фактора A, изменяющегося на уровнях $A_1,\ldots,A_I$, на зависимую переменную *Y* может быть сформулирована как задача установления *однородности* нескольких *совокупностей*: зависимые переменные $Y_1,\ldots,Y_n$ представлены выборками $\{y_i\}^1,\ldots,\{y_i\}^I$ и требуется сравнить их средние, т. е. проверить гипотезу: $H_0 = \mu_1 = \ldots = \mu_I$. В случае

ее принятия все выборки можно считать относящимися к одной совокупности.

Эта же задача может быть представлена и как *задача однофакторной классификации*: совокупность значений переменной *Y* согласно критерию *A* может быть разбита на *I* классов (критерий в данном случае есть способ выбора, процесс измерения). Здесь также можно сформулировать гипотезу о равенстве средних значений по классам: $H_0=\mu_1=\ldots=\mu_I$. В случае ее отклонения классификация считается значимой.

Рассмотрим одну из форм однофакторной модели дисперсионного анализа.

Пусть имеется *I* групп в исходной выборке. Обозначим через μ_i средние в *i*-й группе (*i*=1,...,*I*), считая каждую группу нормально распределенной с одной и той же дисперсией, $N(\mu_i,\sigma^2)$.

Пусть y_{ij} есть *j*-е наблюдение в *i*-й группе, которое можно выразить через групповое среднее:

$$y_{ij}=\mu_i+e_{ij},\ j=1,\ldots,J_i,i=1,\ldots,I,$$

где ошибки e_{ij} независимы и нормально распределены по закону $N(0,\sigma^2)$.

Затем выразим среднее μ_i в виде суммы *генерального среднего* μ и *дифференциальных* (*главных*) *эффектов* α_i, определяемых для каждой группы.

Определим $\mu=\frac{1}{n}\sum_{i=1}^{I}J_i\mu_i$, где $n=\sum_{i=1}^{I}J_i$, и $\alpha_i=\mu_i-\mu$.

Однофакторная модель дисперсионного анализа представляется в виде:

$$y_{ij}=\mu+\alpha_i+e_{ij},\ j=1,\ldots,J_i,i=1,\ldots,I. \qquad (1.5)$$

Существует два типа моделей:

- *модель с фиксированными эффектами* (*модель* I) – это модель вида (1.5), когда рассматриваются группы, соответствующие именно данным уровням фактора; другими словами, уровни фактора выбираются исследователем;

- *модель со случайными эффектами* (*модель* II) – если подгруппы, соответствующие различным уровням фактора, выбираются случайно из большого числа подгрупп, т. е. уровни фактора и соответствующие им переменные рассматриваются как случайные.

Если факторов более одного и одни из них соответствуют модели I, а другие модели II, говорят о *смешанной модели*.

Отметим, что формулы, задающие суммы квадратов и числа степеней свободы одинаковы для моделей всех типов.

В модели с фиксированными эффектами проверяется гипотеза

$$H_0=\mu_1=\ldots=\mu_I=\mu \quad \text{или} \quad H_0=\alpha_1=\ldots=\alpha_I=0.$$

Оценки μ, $\hat{\mu}_i$, $\hat{\alpha}_i$ вычисляются с помощью метода наименьших квадратов минимизацией суммы $SS=\sum_{i=1}^{I}\sum_{j=1}^{J_i}\left(y_{ij}-\mu-\alpha_i\right)^2$ и имеют вид:

$$\hat{\mu}=\overline{y}=\frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J_i}y_{ij};\quad \hat{\alpha}_i=\frac{1}{J_i}\sum_{j=1}^{J_i}y_{ij}-\overline{y}=\hat{\mu}_i-\hat{\mu},\ i=1,\ldots,I.$$

Несмещенная оценка дисперсии вычисляется в виде

$$MS_R=\frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{J_i}\left(y_{ij}-\overline{y}_i\right)^2.$$

Рассмотрим формирование компонент дисперсии на примере данных наблюдений, представленных таблицей 1.2.

Пусть I – число уровней фактора A; J_i – число испытаний (наблюдений) в i-й группе; $n=\sum_{i=1}^{I}J_i$ – общее число наблюдений; y_{ij} – значение отклика Y в j-м испытании на i-м уровне фактора (в i-й группе).

Таблица 1.2. План эксперимента.

№ исп.	Уровни фактора (группы)				
	A_1	…	A_i	…	A_I
1 … j … J_i	y_{11} y_{1J_1}	… … …	… … y_{ij} … …	… … …	y_{I1} y_{IJ_I}
	$\overline{y}_1$	…	$\overline{y}_i$	…	$\overline{y}_I$

В итоговой строке записаны групповые средние $\overline{y}_i = \frac{1}{J_i}\sum_{j=1}^{J_i} y_{ij}$; общая средняя находится по формуле: $\overline{y} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J_i} y_{ij}$.

Проверка основной гипотезы $H_0 = \mu_1 = \ldots = \mu_I = \mu$ может быть проведена на основе статистических критериев (стандартного нормального распределения при известных генеральных дисперсиях или *t*-критерия Стьюдента – при неизвестных).

Другой способ, который и выражает суть дисперсионного анализа, состоит в разложении общей суммы квадратов отклонений на компоненты.

Общая сумма квадратов (*TOTAL*)

$$SS_T = \sum_{i=1}^{I}\sum_{j=1}^{J_i} (y_{ij} - \overline{y})^2$$

является общей оценкой вариации переменной *Y*;

Для отклонения гипотезы H_0 и признания различия необходимо, чтобы вариация зависимой переменной *Y* внутри групп была значительно меньше межгрупповой вариации. Эти вариации количественно оценивают следующие суммы квадратов:

– *межгрупповая сумма квадратов* (*EFFECT*)

$$SS_E = \sum_{i=1}^{I} J_i (\overline{y}_i - \overline{y})^2$$

обусловлена различием значений зависимой переменной *Y* в группах, т. е. отклонением групповых средних от общей средней; межгрупповая сумма квадратов связана с *дифференциальным эффектом*;

– *внутригрупповая сумма квадратов* или *SS* остатка (*RESIDUAL*),

$$SS_R = \sum_{i=1}^{I}\sum_{j=1}^{J_i} (y_{ij} - \overline{y}_i)^2$$

обусловлена различием внутри групп, не связанным с эффектом фактора, т. к. его уровень зафиксирован для каждой группы.

При этом справедливо равенство, которое называют *основным тождеством дисперсионного анализа* [3]:

$$SS_T = SS_E + SS_R .$$

Таким образом, логика дисперсионного анализа состоит в разбиении общей суммы квадратов отклонений на компоненты, одна из которых обусловлена различием групповых средних (эффект фактора), а другая обусловлена случайной ошибкой.

Разделив суммы квадратов отклонений на соответствующие степени свободы, получают дисперсии (средние суммы квадратов отклонений, *MS*):

$$MS_E = \frac{SS_E}{I-1};\ \ MS_R == \frac{SS_R}{n-I};\ \ MS_T = \frac{SS_T}{n-1}.$$

Все три дисперсии являются оценками общей генеральной дисперсии σ^2 для совокупности $\{y_{ij}\}_n$. Однако оценка MS_E зависит от истинности предположения H_0: при нарушении H_0 межгрупповая дисперсия имеет тенденцию к возрастанию и определяет большую долю в общей дисперсии.

При использовании статистических пакетов модуль дисперсионного анализа формирует таблицу вида 1.3 результатов однофакторного дисперсионного анализа.

Таблица 1.3. Таблица одно факторного дисперсионного анализа.

Эффекты (источник вариации)	Суммы квадратов отклонений SS	Степени свободы ν	Дисперсии (средние квадраты) MS	F-отношения
Между уровнями фактора A (межгрупповая)	$SS_E = \sum_{i=1}^{I} J_i(\overline{y}_i - \overline{y})^2$	$\nu_E = I-1$	$MS_E = \frac{SS_E}{I-1}$	$F_A = \frac{MS_E}{MS_R}$
Внутри уровней (внутригрупповая)	$SS_R = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(y_{ij} - \overline{y}_i)^2$	$\nu_R = n-I$	$MS_R == \frac{SS_R}{n-I}$	
Общая	$SS_T = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(y_{ij} - \overline{y})^2$	$\nu_T = n-1$	$MS_T = \frac{SS_T}{n-1}$	

Для проверки основной гипотезы вычисляется *F*-отношение: $F=\frac{MS_E}{MS_E}$. Эта величина имеет *F*-распределение Фишера со степенями свободы $(I-1)$ и $(n-I)$.

Если при заданном уровне значимости α выполняется неравенство $F \geq F_{(\alpha;I-1;n-I)}$ (где $F_{(\alpha;I-1;n-I)}$ есть α-квантиль распределения Фишера), то гипотезу H_0 отклоняют, а в противном случае принимают.

Следовательно, при нарушении гипотезы H_0, т. е. при выполнении неравенства $F \geq F_{(\alpha;I-1;n-I)}$, влияние фактора на вариацию переменной *Y* считается статистически значимым с надежностью $\gamma=1-\alpha$.

1.4.2. Апостериорные критерии и контрасты

При принятии нулевой гипотезы все групповые средние считаются неразличимыми и оцениваются величиной общего среднего

$$\hat{\mu}=\overline{y}=\frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J_i} y_{ij};\ i=1,\ldots,I .$$

При отклонении нулевой гипотезы изучаются групповые средние

$$\hat{\mu}_i=\frac{1}{J_i}\sum_{j=1}^{J_i} y_{ij},\ i=1,\ldots,I .$$

и дифференциальные эффекты $\alpha_1,\ldots,\alpha_I$. Их оценками являются значения

$$\hat{\alpha}_i=\frac{1}{J_i}\sum_{j=1}^{J_i} y_{ij}-\overline{y}=\hat{\mu}_i-\hat{\mu},\ i=1,\ldots,I ,$$

имеющие нормальное распределение и статистически независимые от дисперсии остатков. Доверительные интервалы для $\hat{\mu}_i$ определяются по формулам

$$\hat{\mu}_i \pm t_{1-\alpha/2} v_R \sqrt{\frac{MS_R}{J_i}},\ i=1,\ldots,I .$$

При отклонении нулевой гипотезы имеется, по крайней мере, одна пара групповых средних, различие которых статистически значимо. Сам F-критерий не дает информации о том, какие именно из средних не равны.

Для установления того, какие из групповых средних можно считать наиболее значимо отличными от других средних, выполняются парные или множественные сравнения.

Установление значимости различий в парных сравнениях основано на *апостериорных критериях* (*Post Hoc*). Эта процедура проводится только после установления статистически достоверного результата дисперсионного анализа. В противном случае процедура некорректна.

Число апостериорных критериев очень велико, и наиболее часто используются следующие:

а) тест *LSD* (*Least Significant Difference*, наименьшей значимой разности) – совокупность *t-критериев* для всевозможных пар групп; однако этот критерий наиболее подвержен ошибкам;

б) критерий Бонферрони (*Bonferroni*) схож с критерием *LSD*, однако дает меньше ошибочных результатов за счет деления уровня значимости на число сравнений;

в) критерий Шеффе (*Scheffe*) использует *F*-критерий вместо *t*-критерия и потому точнее;

г) критерий Тьюки (*Tukey*), или HSD-критерий (*Honestly Significant Difference*, подлинно значимой разности) использует статистику Стьюдента и применяется при большом числе уровней факторов.

Полученные в процедурах критические значения определяют границы разностей средних по группам, превышение которых означает статистическую значимость различий. Наибольшая значимая разность средних указывает на соответствующие им группы, наиболее (а возможно, наименее) интересные для дальнейшего анализа.

Для установления значимости различий групповых средних проверяется гипотеза $H_0: \mu_j - \mu_k = 0$ при альтернативе $H_0: \mu_j - \mu_k \neq 0$.

Для этого строится доверительный интервал $\hat{\mu}_i \pm t_{1-\alpha/2} v_R \sqrt{MS_R \left(\frac{1}{J_i} + \frac{1}{J_k} \right)}$.

Если этот интервал *не содержит нуля*, то гипотеза отклоняется.

В исследовании также представляют интерес сравнения *контрастов* (*Contrasts*) – линейных комбинаций вида

$$\sum_{i=1}^{I} c_i \mu_i \text{, где } \sum_{i=1}^{I} c_i = 0 .$$

Выдвигая гипотезы о контрастах, можно получить дополнительную информацию.

К контрастам применимы апостериорные критерии. Например, критерий Шеффе предполагает проверку гипотезы $H_0: \left(\sum_{i=1}^{I} c_i \mu_i = 0 \right)$

против альтернативы $H_1:\left(\sum_{i=1}^{I} c_i \mu_i \neq 0\right)$. Рассматривается доверительный интервал для линейной комбинации $\sum_{i=1}^{I} c_i \mu_i \pm S$, где $S^2=(I-1)MS_R F_{(1-\alpha,(I-1,n-I)}\sum_{i=1}^{I}\frac{c_i}{J_i}$. Если этот интервал не содержит 0, то H_0 отклоняется. Эта проверка выполняется для каждого контраста, представляющего интерес.

Контрасты позволяют сравнивать один уровень фактора (и соответствующую группу) с другим, один уровень фактора со всеми остальными, разбить все уровни фактора на две группы и сравнить их. При этом, в отличие от апостериорных критериев, не требуется предварительного получения статистически значимого результата дисперсионного анализа. Сравнение на основе контрастов сводится к применению модифицированного варианта *t*-критерия.

Коэффициенты в линейной комбинации $\sum_{i=1}^{I} c_i \mu_i$ подбираются произвольно, главное, чтобы их сумма равнялась единице. Уровням фактора из разных подгрупп соответствуют разные знаки, для уровней, не включенных в сравнение, задается 0. Так, если сравниваются контрасты для фактора с тремя уровнями, причем первый уровень противопоставляется по значимости второму и третьему, коэффициенты определяются как – 2, 1, 1. Если же сравниваются только первый и третий уровни, то задается последовательность 1, 0, –1.

1.4.3. Однофакторный ранговый анализ

При нарушении основных предположений о нормальном распределении зависимой переменной и однородности дисперсий в группах наблюдений. При нарушении предположения о нормальном распределении одним из подходов является подбор преобразования, которое сделало бы распределение зависимой переменной близким к нормальному. Другой подход ориентирован на использование непараметрических методов, из которых наиболее разработаны ранговые методы.

Для проверки нулевой гипотезы могут быть использованы критерии Краскала-Уоллиса, Джонкхиера и другие [11].

Критерий Краскала-Уоллиса (*Kruskal-Wallis test*) применяется, если переменная *Y* измеряется в порядковой шкале (баллы, экспертные оценки) или закон ее распределения неизвестен**.**

Вместо значений переменной y_{ij} из выборки объема N используются их ранги $p_{ij} \in \{1,\ldots,I\}$ $I \le N$; при этом вычисляются:

– средние ранги по группам $$R_j = \frac{1}{l_j}\sum_{i=1}^{l_j} p_{ij};$$

– общий средний ранг $$R = \frac{N+1}{2};$$

Затем вычисляется *статистика Краскела-Уоллиса*

$$H = \frac{12}{N(N+1)}\sum_{j=1}^{k} l_j\left(R_j - \frac{N+1}{2}\right)^2,$$

которая в случае выполнения нуль-гипотезы имеет χ^2 – распределение.

Критерий Краскела-Уоллиса состоит в следующем: если статистика $H \ge \chi^2_{(1-\alpha;p-1)}$, то гипотезу H_0 о равенстве средних в группах, соответствующих градациям фактора, отклоняют. Для отклонения гипотезы нужно, чтобы значения критерия были большими.

Если гипотеза H_0 отвергается, то для выделения заведомо различных выборок применяются *свободные от распределения методы множественного сравнения*. Для сравнения всевозможных пар выборок можно использовать *приближенный метод Данна* (*Dunn test*). Две выборки различаются параметром сдвига значимо на уровне значимости α, если

$$|R_i - R_j| \ge z_{\alpha/(I(I-1))}\left(\frac{N(N+1)}{12}\right)^{1/2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)^{1/2}$$

где R_i, R_j – средние ранги выборок, $z_{\alpha/(I(I-1))}$ – квантиль стандартного нормального распределения.

Критерий Джонкхиера используется, если заранее известно, что группы значений результатов (зависимой переменной) упорядочены по возрастанию влияния фактора. Статистика Джонкхиера опирается на парную статистику Манна-Уитни, полученную для сравнения двух выборок $\{x_i\}_n$ и $\{y_i\}_n$ имеет вид:

$$U = \sum_{i=1}^{n}\sum_{j=1}^{n}\phi(x_i, y_j),$$

$$\phi(x_i, y_j) = \begin{cases} 1, & \text{если } x_i < y_j; \\ 1/2, & \text{если } x_i = y_j; \\ 0, & \text{если } x_i > y_j. \end{cases}$$

где

В общем случае, когда сравнивается k групп, для каждой пары натуральных чисел u и v, где $1 \le u < v \le k$, по выборкам с номерами u и v составляется статистика Манна-Уитни:

$$U_{u,v} = \sum_{i=1}^{n_u} \sum_{j=1}^{n_v} \phi(x_{iu}, y_{jv}),$$

а затем определяется статистика Джонкхиера:

$$J = \sum_{1 \le u < v \le k} U_{u,v}.$$

Для отклонения гипотезы нужно, чтобы значения критерия были большими.

При небольших объемах выборок и небольшом числе групп k значение статистики J табулировано. Для больших выборок объема N действует нормальная аппроксимация: $J \sim N(MJ, DJ)$, где

$$MJ = \frac{1}{4}\left(N - \sum_{j=1}^{k} n_j^2\right), \quad DJ = \frac{1}{72}\left(N^2(2N+3) - \sum_{j=1}^{k} \left(n_j^2(2n_j+3)\right)\right).$$

1.4.4. Двухфакторный *ANOVA*-анализ

При наличии двух варьирующих факторов рассматриваются модели, используемые для анализа их дифференциальных эффектов.

Два фактора *A* и *B* называются *пересекающимися*, если наблюдениями представлены все возможные сочетания их уровней, т. е. для каждой комбинации план эксперимента имеется хотя бы одно наблюдение. Модель с двумя пересекающимися факторами называют *двухфакторным планом*, *двухфакторной классификацией* или *факторной моделью с двумя факторами*.

Другой известный тип отношения факторов – когда один из факторов, пусть это *B*, *группируется* фактором *A*, т. е. каждый уровень фактора *B* встречается в паре не более чем с одним уровнем фактора *A*. Модель с такими факторами называется *двухфакторной моделью с группировкой* или *двухфакторной иерархической моделью*.

Рассмотрим случай влияния на зависимую переменную *Y* двух пересекающихся факторов: фактора *A*, изменяющегося на *I* уровнях, и фактора *B*, имеющего *J* уровней. На каждой паре сочетаний выполняется одинаковое число *K* испытаний. В предположении, что оба фактора соответствуют модели I, *двухфакторный план с фиксированными эффектами* задается соотношениями

$$y_{ijk}=\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}+e_{ijk},\ i=1,\ldots,I,\ j=1,\ldots,J,\ k=1,\ldots,K \quad (1.6)$$

Здесь μ – генеральное среднее, α_i есть *i*-й *дифференциальный* (*главный*) *эффект* фактора *A*, β_j есть *j*-й *дифференциальный* (*главный*) *эффект* фактора *B*, а величина $(\alpha\beta)_{ij}$ называется (*двухфакторным*) *взаимодействием i*-го уровня фактора *A* и *j*-го уровня фактора *B*. Ошибки e_{ijk} независимы и нормально распределены по закону $N(0,\sigma^2)$.

Модель, в которой все $(\alpha\beta)_{ij}$ при всех *i*, *j* равны нулю, называется *аддитивной* или моделью с главными эффектами (*Main Effects ANOVA*).

Для оценки параметров модели (1.6) вводится дополнительное предположение о равенстве нулю всех сумм эффектов

$$\sum_{i=1}^{I}\alpha_i=0;\ \sum_{j=1}^{J}\beta_j=0;\ \sum_{i=1}^{I}(\alpha\beta)_{ij}=0,\ j=1,\ldots,J;\ \sum_{j=1}^{J}(\alpha\beta)_{ij}=0,\ i=1,\ldots,J.$$

Тогда получают следующие оценки метода наименьших квадратов:

$$\hat{\mu}=\bar{y}=\frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}y_{ijk};$$

$$\hat{\alpha}_i=\frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K}y_{ijk}-\bar{y},\ i=1,\ldots,I;\ \hat{\beta}_j=\frac{1}{IK}\sum_{i=1}^{I}\sum_{k=1}^{K}y_{ijk}-\bar{y},\ j=1,\ldots,J;$$

$$(\hat{\alpha}\hat{\beta})_{ij}=\frac{1}{K}\sum_{k=1}^{K}y_{ijk}-\frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K}y_{ijk}-\frac{1}{IK}\sum_{i=1}^{I}\sum_{k=1}^{K}y_{ijk}+\bar{y},\ i=1,\ldots,I,\ j=1,\ldots,J.$$

Разложение суммы квадратов можно представить в таком виде:

$$SS_T=SS_A+SS_B+SS_{AB}+SS_R.$$

Суммы квадратов отклонений в правой части этого выражения соответствуют главным эффектам факторов и их взаимодействию. Таблица итогов дисперсионного анализа примет следующий вид (табл. 1.4):

Таблица 1.4. Таблица двухфакторного дисперсионного анализа.

Эффекты	Суммы квадратов отклонений SS	Дисперсии S^2	F-отношения
A	$SS_A = JK\sum_i \left(\bar{y}_{i\cdot ¿} - \bar{y}\right)^2$	$S_A^2 = \frac{SS_A}{I-1}$	$F_A = \frac{S_A^2}{S_R^2}$
B	$SS_B = IK\sum_j ¿¿¿$	$S_B^2 = \frac{SS_B}{J-1}$	$F_B = \frac{S_B^2}{S_R^2}$
Взаимодействие AB	$SS_{AB} = K\sum_i\sum_j ¿¿¿¿$	$S_{AB}^2 = \frac{SS_{AB}}{(I-1)(J-1)}$	$F_{AB} = \frac{S_{AB}^2}{S_R^2}$
Остаток (ошибка)	$SS_R = \sum\sum\sum ¿¿¿¿¿$	$S_R^2 = \frac{SS_R}{IJ(K-1)}$	
Общая	$SS_T = \sum\sum\sum \left(y_{ijk} - \bar{y}\right)^2$	$\bar{S}_T^2 = \frac{SS_T}{IJK-1}$	

Здесь средние значения обозначены как

$$\bar{y} = \bar{y}_{...} = \frac{1}{IJK}\sum_i\sum_j\sum_k y_{ijk};$$

$$\bar{y}_{i..} = \frac{1}{JK}\sum_j\sum_k y_{ijk};\quad \bar{y}_{.j.} = \frac{1}{IK}\sum_i\sum_k y_{ijk};\quad \bar{y}_{ij\cdot ¿} = \frac{1}{K}\sum_k y_{ijk}\ ¿.$$

Вычисленные наблюдаемые значения F-отношения сравниваются с критическими значениями $F_{(\alpha;\nu_1;\nu_2)}$ при соответствующих степенях свободы. Если для некоторого эффекта $F_¿ > F_{(\alpha;\nu_1;\nu_2)}$, то для него гипотеза H_0 отклоняется, и эффект признается статистически значимым.

Для иллюстрации эффектов взаимодействия используются категорированные *графики средних* (рис 1.1). Они показывают, как влияет (и влияет ли) переход из одной группы в другую по фактору A на изменение среднего значения Y в группах по фактору B.

Приведенная иллюстрация относится к случаю, когда изменение уровня фактора A от $A1$ до $A2$ приводит к уменьшению среднего значения Y в группе $B1$ и к увеличению в группе $B2$.

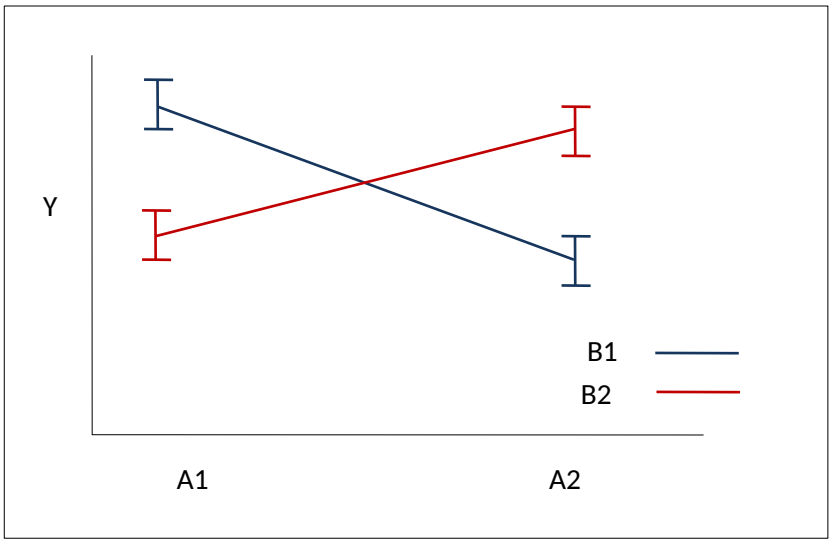

Рис. 1.1. График средних.

В данном пункте была рассмотрена одна из наиболее простых и распространенных моделей двухфакторного анализа. Обширный класс моделей для случая группируемых факторов, а также факторов с фиксированными, случайными или смешанными эффектами, для повторяемых планов и т. п. представлен в [2].

Остановимся вкратце на одной из таких моделей.

1.4.5. Дисперсионный анализ с повторными измерениями

Рассмотрим еще один тип дисперсионного анализа – анализ с повторными измерениями (*Repeated Measures ANOVA*), или, другими словами, анализ связанных выборок. В приведенной выше схеме двухфакторного дисперсионного анализа предполагается, что на каждом из *IJ* сочетаний уровней факторов проводится наблюдение за *K* объектами – образуются группы из *K* значений одной переменной. Эти группы наблюдений различаются между собой и разница между ними описывается *межгрупповым фактором*.

Если на каждом из *IJ* сочетаний уровней факторов имеются повторные измерения одних и тех же переменных (при разных условиях или в разное время) *для одних и тех же объектов*, то говорят о наличии *фактора повторных измерений* (называемого также *внутригрупповым фактором*, поскольку для оценки его значимости вычисляется внутригрупповая сумма квадратов). Для исследования значимости факторов повторных измерений, имеющих более двух уровней, используются методы многомерного дисперсионного анализа (раздел 2).

Во многих случаях эксперимент требует включение в план и межгруппового фактора, и фактора повторных измерений. Пусть например, анализируются итоги промежуточной аттестации знаний студентов двух различных групп в двух или трех семестрах изучения одной дисциплины. В роли межгруппового фактора могут выступать

академические группы. Оценки каждого студента по семестрам образуют внутригрупповой фактор (или фактор с повторными измерениями).

Интерпретация главных эффектов и взаимодействий для межгрупповых факторов и факторов повторных измерений совпадает, и оба типа факторов могут, очевидно, взаимодействовать между собой (например, у студентов первой группы повышаются оценки знаний по семестрам, а у студентов второй группы снижаются).

1.5. Планирование эксперимента

Одной из проблем многофакторного дисперсионного анализа является реализация программы наблюдений. Уже в случае двух факторов *полный факторный эксперимент* (ПФЭ) включает *IJK* наблюдений, и его проведение становится затруднительным.

В таком случае используются методы *планирования эксперимента* (*Experiment design*), предполагающие построение таких *неполных планов*, которые обеспечивают достижение цели анализа. Например, в случае наличия трех факторов, изменяющихся на четырех уровнях каждый, полный план включает $64=4^3$ эксперимента (наблюдения), а план типа латинского квадрата (табл. 1.5) сокращает их число до 16.

Здесь факторы 1 и 2 представлены уровнями 1, 2, 3, 4, а фактор 3 – уровнями *A*, *B*, *C*, *D*. Латинский квадрат относится к типу *гнездовых планов* – в них наблюдения «вложены» в ячейки плана («гнезда»), соответствующие сочетаниям уровней факторов.

Таблица 1.5. Латинский квадрат.

фактор 1 фактор 2	1	2	3	4
1	A	B	C	D
2	D	A	B	C
3	C	D	A	B
4	B	C	D	A

Другой способ сокращения числа экспериментов связан с построением дробных реплик при небольшой потере информации. Дробные реплики – это части матрицы плана (1/2, 1/4, 1/8 части плана), которые строят по специальным правилам.

Методы планирования эксперимента использовал Р. Фишер, разрабатывая основные вопросы дисперсионного анализа. Большая заслуга в дальнейшем развитии идей и методов планирования эксперимента принадлежит Боксу (*Box G. E. P.*, 1960), его ученикам и

последователям Уилсону (*Wilson K. B.),* Хантеру (*Hunter J. S.*), Бенкену (*Behnken D. W.*) и др. Математическая теория планирования эксперимента представлена в [8].

Экспериментальные методы и планирование эксперимента широко применяются в промышленности для оптимизации производственных процессов. Целью этих методов является поиск оптимальных уровней факторов, определяющих течение процесса производства. Многие из них зарекомендовали себя и в задачах планирования прибыли в бизнесе, управления финансовыми потоками в банковском деле и многих других. В научных приложениях методы дисперсионного анализа используются для выяснения реальной природы взаимодействий, проявляющейся во взаимодействии факторов высших порядков.

К числу основных типов планов, обычно используемых в исследованиях, относятся:

- планы $2^{(k-p)}$ (двухуровневые многофакторные планы), где k – число факторов, p – количество взаимодействий, не учитываемых в плане; применимы для факторов с двумя уровнями в предположении линейности изменений зависимой переменной;
- планы $3^{(k-p)}$ (трехуровневые многофакторные планы), где k – число факторов, p – количество взаимодействий, не учитываемых в плане; уровни факторов (0,1,2) удовлетворяют фундаментальному тождеству $(A+B+C)\bmod 3=0$;
- смешанные двух- и трехуровневые планы;
- отсеивающие планы для большого числа факторов;
- центральные композиционные планы (или планы поверхности отклика); зависимая переменная представляется линейной комбинацией факторов, их квадратичных эффектов и парных взаимодействий; уровни факторов выбираются произвольно;
- планы на латинских квадратах;
- робастные планы Тагучи;
- планы для смесей на симплексе и др.

2. Многомерный дисперсионный анализ

2.1. Многомерная обобщенная линейная модель

Методы *многомерного дисперсионного анализа MANOVA* (*Multcriterian Analysis of Variances*) используются, когда имеется несколько зависимых переменных $Y_1,\ldots,Y_p$. В этом случае возрастает сложность вычислений, но содержание и основные принципы не меняются: проверяются *гипотезы о равенстве средних* в подгруппах, соответствующих плану эксперимента.

При сравнении *p*-мерного дисперсионного анализа с одномерным, выполненным для *p* признаков по отдельности может оказаться, что результаты будут отличаться. Это объясняется тем, что в многомерных моделях учитываются все связи зависимых переменных, и их корреляция играет существенную роль.

Рассмотрим обобщенную линейную модель. В соответствии с планом эксперимента формируется *матрица плана X*, размер и вид которой определяется числом факторов, количеством их уровней и принятым кодированием.

В соответствии с [2] предположим, что для каждого из *n* объектов измеряются *p* переменных. Пусть вектор $Y_i=(y_{i1},\ldots,y_{in})'$ соответствует *n* независимым измерениям *i*-й переменной.

Для каждого такого вектора можно записать одномерную линейную модель:

$$E(Y_i)=X'\beta_i+\varepsilon_i, \tag{2.1}$$

где $(X')^{n\times m}$ – матрица плана ранга $r\le m<n$, $\beta_i=(\beta_{i1},\ldots,\beta_{im})'$ – вектор из *m* неизвестных параметров, специфичных для каждой переменной.

Далее *p* линейных моделей, задаваемых формулами (2.1), взятые вместе, составляют *многомерную обобщенную линейную модель* с одинаковой для всех переменных матрицей плана и различными векторами β_i. Зависимость переменных выражается формулой

$$\operatorname{cov}(Y_i,Y_j)=\sigma_{ij}I,\ \ i,j=1,\ldots,p \tag{2.2}$$

где σ_{ij} – ковариация между *i*-й и *j*-й переменными, $p\le n-r$, $m<n$.

Модель, задаваемую соотношениями (2.1) и (2.2), можно представить в виде

$$Y = X^{'}\beta + \varepsilon, \qquad (2.3)$$

где $Y^{n\times p} = \begin{pmatrix} y_{11} & y_{21} & \dots & y_{p1} \\ y_{12} & y_{22} & \dots & y_{p2} \\ \dots & \dots & \dots & \dots \\ y_{1n} & y_{2n} & \dots & y_{pn} \end{pmatrix}$ – матрица значений откликов,

$(X^{'})^{n\times m}$ – матрица плана ранга $r \le m < n$,

$\beta^{m\times p} = \begin{pmatrix} \beta_{11} & \beta_{21} & \dots & \beta_{p1} \\ \beta_{12} & \beta_{22} & \dots & \beta_{p2} \\ \dots & \dots & \dots & \dots \\ \beta_{1m} & \beta_{2m} & \dots & \beta_{pm} \end{pmatrix}$ – матрица неизвестных параметров,

$\varepsilon^{n\times p} = \|\varepsilon_{ij}\|$ – матрица случайных ошибок.

В качестве примера [2] модели с n=100, p=16, (n индивидов, p зависимых переменных-откликов) и одним фактором с m=3 уровнями запишем ее представление в виде (2.3):

$$\begin{pmatrix} y_{11} & \dots & y_{16,1} \\ \dots & \dots & \dots \\ \underline{y_{1,25}} & \dots & \underline{y_{16,25}} \\ y_{1,26} & \dots & y_{16,26} \\ \dots & \dots & \dots \\ \underline{y_{1,70}} & \dots & \underline{y_{16,70}} \\ y_{1,71} & \dots & y_{16,71} \\ \dots & \dots & \dots \\ y_{1,100} & \dots & y_{16,100} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ \dots & \dots & \dots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \dots & \dots & \dots \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ \dots & \dots & \dots \\ 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} \mu_1 & \mu_2 & \dots & \mu_{16} \\ \beta_{11} & \beta_{21} & \dots & \beta_{16,1} \\ \beta_{12} & \beta_{22} & \dots & \beta_{16,2} \end{pmatrix} + \varepsilon$$

Оценка параметров β_i модели (2.3) производится, как и в случае одномерной модели, на основе метода наименьших квадратов, что приводит к системе нормальных уравнений

$$(X^{'}X)\beta_i = XY_i, \quad i = 1, \dots, p.$$

Величина $R_0(i,i)=Y_i^{'}Y_i-(Y_i^{'}X^{'})\beta_i$ называется *остаточной суммой квадратов,* величина $R_0(i,j)=Y_i^{'}Y_j-(X^{'}Y_j^{'})\beta_i,\ i,j=1,\ldots,p$ – *остаточной суммой произведений.* На их основе определяются несмещенные оценки дисперсий: $\sigma_{ii}=\frac{R_0(i,i)}{n-r}$, $\sigma_{ij}=\frac{R_0(i,j)}{n-r}$.

Матрица

$$R_0^{p\times p}=\begin{pmatrix} R_0(1,1) & \ldots & R_0(1,p) \\ \ldots & \ldots & \ldots \\ R_0(p,1) & \ldots & R_0(p,p) \end{pmatrix}$$

называется *матрицей остаточных сумм квадратов и произведений*.

В обобщенной линейной модели требуется одновременно проверить *гипотезы*

$$H_0: H^{'}\beta_i=\xi_i,\ \ i=1,\ldots,p$$

где матрица $(H^{'})^{s\times m}$ имеет ранг $s\le r$ и векторы $\xi_i^{s\times 1}$ предполагаются заданными. Если гипотеза H_0 выполняется, то получается *модель с ограничениями*, на основе которой можно найти оценки $\beta_i^{¿}$ параметров β_i и матрицу R_1 остаточных сумм квадратов и произведений. Матрица R_1-R_0 называется матрицей сумм квадратов и произведений, *обусловленных отклонением от гипотезы*.

Разложение $R_1=R_0+(R_1-R_0)$ является обобщением одномерного дисперсионного анализа. Таким образом, отклонение от гипотезы H_0 может быть определено сравнением матриц R_0 и R_1-R_0.

2.2. Критерии проверки многомерной гипотезы

С. Н. Рой (*Roy S. N*, 1965) показал, что статистика критерия для проверки гипотезы H_0 основана на корнях $\lambda_1,\ldots,\lambda_p$ характеристического уравнения $|R_0-\lambda R_1|=0$. Известно [1], что из p корней этого уравнения лишь $t=\min(p,s)$ больше нуля, остальные равны нулю.

Для проверки гипотезы могут быть использованы различные функции, зависящие от корней. В числе критериев наиболее часто используются следующие:

1) Критерий наибольшего характеристического корня

$$V = \lambda_1 = \max\left(\lambda_1, \ldots, \lambda_p\right) \tag{2.4}$$

предложен Роем на основе эвристического метода, поскольку этот корень отражает максимальное отклонение от гипотезы.

2) Λ-*критерий Уилкса* (критерий отношения правдоподобия):

$$\Lambda = \lambda_1 \cdot \ldots \cdot \lambda_p = \frac{|R_0|}{|R_1|}. \tag{2.5}$$

Многомерный Λ-критерий Уилкса (*Wilks S. S.*, 1961), основан на сравнении матрицы R_0 остаточных сумм квадратов и произведений и матрицы R_1 сумм квадратов и произведений, *обусловленных отклонением от гипотезы.*

λ-критерий Уилкса можно переписать в другой форме, если положить $S_E = R_0$ и $S_H = R_1 - R_0$, а именно: $\Lambda = \frac{|S_E|}{|S_H + S_E|}$.

Статистика Λ при некоторых сочетаниях параметров имеет *F*-распределение. В общем случае Λ имеет приближенно χ^2-распределение с *ps* степенями свободы [1]. Таким образом, для определения критических значений можно использовать таблицу χ^2-распределения.

Лучшая аппроксимация для Λ-критерия предложена Рао (*Rao S. R.)* и сводится к *F*-распределению.

3) *Критерий следа Лоули и Хотеллинга* (*Lawley D. N., Hotelling H.*, 1951)

$$T^2 = Sp\left(\left(R_1 - R_0\right)\left(R_0\right)^{-1}\right) = \sum_{i=1}^{t} \lambda_i .$$

Этот критерий считается удобным для практических вычислений, так как не требует нахождения собственных значений. Здесь также можно при определенном сочетании параметров использовать χ^2-распределение или *F*-распределение.

Таким образом, для всех критериев существуют трудности расчета критических значений. И хотя для ряда сочетаний параметров получены таблицы критических значений, используемые на практике, имеющиеся таблицы критических значений $V_\alpha, \Lambda_\alpha, T_\alpha$ не являются исчерпывающими.

Есть также достаточно хорошие аппроксимации распределений статистик, например, *F*-аппроксимация, которую и используют в практической работе. Детально вопросы свойств и приближений критериев рассмотрены в [1].

Если общий многомерный критерий значим, делают заключение о значимости соответствующего эффекта. Однако возникает вопрос о влиянии этого эффекта только на отдельную зависимую переменную или на всю их совокупность. Для оценки влияния главного эффекта или взаимодействия исследуют одномерные *F*-критерии по зависимым переменным, т. е. исследуются зависимые переменные, вносящие вклад в значимость многомерного критерия.

Одним из приемов анализа является суммирование нескольких независимых переменных и применение схемы одномерного дисперсионного анализа, Однако суммирование не всегда имеет смысл, а кроме того надо учитывать усреднение ошибок.

2.3. Отклонение от основных предположений

Многомерный дисперсионный анализ основывается на ряде предпосылок, включающих многомерное нормальное распределение векторов наблюдений, равенство ковариационных матриц, стохастическую независимость векторов наблюдений, отсутствие пропущенных данных. Эти предпосылки часто нарушаются на практике. И возникает вопрос об устойчивости критериев значимости.

Известно, что *F*-критерий устойчив к отклонению от нормальности распределения и в некоторой степени к нарушению равенства дисперсий. Для многомерных критериев исследования устойчивости не имеют окончательного характера. Но невыполнение предпосылок может привести лишь к тому, что изменится уровень значимости. А это означает, что многомерный анализ имеет смысл, так как он уплотняет экспериментальные данные.

На практике в большинстве случаев многомерный дисперсионный анализ применяют без детального исследования многомерного нормального распределения. Гораздо важнее проверять предпосылку о равенстве ковариационных матриц. В случае ее нарушения выводы ненадежны. И наиболее высокие требования предъявляются к независимости векторов наблюдений.

3. Задачи исследований с применением модели дисперсионного анализа

Изложенный в разделах 1 и 2 теоретический аппарат дисперсионного анализа позволяет проводить разнообразные исследования в различных областях. В многочисленных монографиях, учебных пособиях и практических руководствах, посвященных анализу данных, приводится большое количество иллюстрирующих примеров, а также описания и подходы к решению комплексных исследовательских задач. Надежность результатов исследований и выводы опираются на строгие обоснования и обеспечиваются выполнением основных предположений.

Далее будут рассмотрены задачи анализа наблюдений и фиксируемых результатов в области образовательной деятельности.

3.1. Точный тест Фишера

Точный тест Фишера (*Fisher's exact test*) – это непараметрический тест статистической значимости, используемый в анализе категориальных данных, когда размеры выборки малы и данные не уравновешены. Для больших уравновешенных выборок применяется тест хи-квадрат.

Тест Фишера [15] обычно используется для исследования значимости взаимосвязи между двумя переменными в факторной таблице размерности 2×2 (таблице сопряженности признаков). Величина вероятности P теста вычисляется, как если бы значения на границах таблицы известны.

Пусть, например, необходимо установить различие в отношениях к некоторому событию или объекту («удовлетворен», «не удовлетворен») в двух категориях опрашиваемых (категория 1, категория 2). В общем виде таблица сопряженности имеет следующее представление:

Таблица 3.1. Таблица сопряженности. Общий вид.

	«удовлетворен»	«не удовлетворен»	Всего
Категория 1	a	b	$a+b$
Категория 2	c	d	$c+d$
Всего	$a+c$	$b+d$	n

Значения a, b, c, d есть клеточные частоты, значения итоговых столбца и строки – так называемые маргинальные (граничные) значения.

Проверяемая нулевая гипотеза состоит в предположении об отсутствии различия в удовлетворенности событием (или объектом) в двух категориях опрашиваемых, т. е. соотношение между удовлетворенными и неудовлетворенными в обеих группах не отличается.

Фишер показал, что вероятность получения любого такого набора величин a, b, c, d определяется гипергеометрическим распределением:

$$P=\binom{a+b}{a}\binom{c+d}{c}/\binom{n}{a+c}=\frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}, \qquad (3.1)$$

где $\binom{m}{k}$ есть биномиальные коэффициенты. Формула (3.1) дает точную вероятность наблюдения любого специфического набора данных, при условии *заданных* маргинальных итогов, общего итога и нулевой гипотезы.

В одностороннем тесте Фишера выбирается наименьшее из чисел a, b, c, d и моделируются все возможные таблицы сопряженности с клеточными частотами от 0 до найденного наименьшего числа в соответствующей клетке, так чтобы маргинальные итоги не менялись, т. е. соответственно пересчитываются все клеточные чистоты. Для каждой такой таблицы по формуле (3.1) определяется вероятность, которая, если она не превышает вероятность исходной конфигурации таблицы, заносится в сумму вероятностей.

В двустороннем тесте Фишера определяется еще и наибольшее значение среди a, b, c, d и моделируются таблицы сопряженности с частотами в соответствующей клетке от найденного максимума до маргинального итога строки с пересчетом других клеточных частот. Вероятности (3.1), не превышающие исходной, суммируются.

Количество таблиц зависит от распределения клеточных частот, вот почему тест Фишера применяется для неуравновешенных (перекошенных») таблиц сопряженности. Большинство современных статистических пакетов вычисляет значение тестов Фишера, в некоторых случаях даже там, где приближение хи-квадрат также было бы приемлемым.

Пример 3.1. Рассмотрим в качестве примера исследование удовлетворенности магистерской программой по результатам опроса обучающихся одного курса, разделенных на две категории – совмещающих учебу с работой и не работающих.

Таблица 3.2. Таблица сопряженности. Общий вид.

	«удовлетворен»	«не удовлетворен»	Всего
Работает	*11*	*1*	*12*
Не работает	*8*	*1*	*9*
Всего	*19*	*2*	*21*

Проверяемая нулевая гипотеза состоит в предположении об отсутствии различия в удовлетворенности магистерской программой в двух категориях опрашиваемых, т. е. соотношение между

удовлетворенными и неудовлетворенными в обеих группах не отличается. Тест Фишера (односторонний), выполненный с использованием тест-калькулятора пакета *STATISTICA*, определил вероятность P=0,6857. Это означает принятие нуль-гипотезы.

3.2. Проверка значимости модели регрессии

Традиционной областью применения дисперсионного анализа является парный и множественный регрессионный анализ. Эта область применения обусловлена общей линейной моделью (1.1), представленной в п. 1.3. Различие заключается в типах независимых переменных: в модели регрессии все переменные количественные и нормально распределенные, а в дисперсионном анализе роль независимых переменных играют неколичественные факторы, изменяющиеся на присущих им уровнях (градациях). Зависимая переменная в обоих случаях (за исключением непараметрических задач анализа) должна быть количественной и нормально распределенной.

Рассмотрим уравнение регрессии $y=f(\tilde{x},\tilde{\beta})+\varepsilon$, где $\tilde{x}=(x_1,\ldots,x_k)$ – независимые переменные, представленные в n наблюдениях матрицей $X=\|x_{ij}\|_{n\times k}$, а $\tilde{\beta}=(\beta_1,\ldots,\beta_p)$ – вектор параметров, оценки которых $B=(b_1,\ldots,b_p)$ для линейной функции f вычисляются на основе метода наименьших квадратов: $B=(X^TX)^{-1}(X^TY)$, где $Y=\|y_i\|_n$ – вектор наблюдений зависимой переменной.

Для проверки значимости уравнения регрессии формулируется нулевая гипотеза о том, что функция f не соответствует описанию связи зависимой переменной с независимыми. В случае, например, парной линейной регрессии $y=\beta_0+\beta_1x+\varepsilon$ нулевая гипотеза имеет вид $H_0:\beta_1=0$. Проверка гипотезы использует F-отношение, которое вычисляется после заполнения таблицы дисперсионного анализа вида 3.3.

Таблица 3.3. Таблица дисперсионного анализа в модели регрессии.

Источник вариации	Суммы квадратов отклонений SS	Степени свободы ν	Дисперсии (средние квадраты) MS	F-отношения
Функция регрессии	$SS_E=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$	$\nu_E=p-1$	$MS_E=\frac{SS_E}{p-1}$	$F=\frac{MS_E}{MS_R}$
Отклонение от функции регрессии	$SS_R=\sum_{i=1}^{n}(\hat{y}_i-y_i)^2$	$\nu_R=n-p$	$MS_R==\frac{SS_R}{n-p}$	
Общая	$SS_T=\sum_{i=1}^{n}(y_i-\bar{y})^2$	$\nu_T=n-1$	$MS_T=\frac{SS_T}{n-1}$	

В приведенной таблице сумма квадратов, объясненная функцией регрессии $SS_E=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$, представляет собой отклонения предсказанных значений $\hat{y}_i=f(\tilde{x}_i,B)\,(i=1,\ldots,n)$ от наблюдаемого среднего $\bar{y}$. Сумма $SS_E=\sum_{i=1}^{n}(\hat{y}_i-y_i)^2$ определяется случайными отклонениями наблюдений от предсказаний.

Отношение дисперсий (средних квадратов) $F=\frac{MS_E}{MS_E}$ *F*-распределение Фишера со степенями свободы $(p-1)$ и $(n-p)$.

Если при заданном уровне значимости α выполняется неравенство $F\geq F_{(\alpha;p-1;n-p)}$ (где $F_{(\alpha;p-1;n-p)}$ есть α-квантиль распределения Фишера), то нулевую гипотезу отклоняют, а в противном случае принимают.

Следовательно, при нарушении гипотезы, т. е. при выполнении неравенства $F\geq F_{(\alpha;I-1;n-I)}$, уравнение регрессии считается статистически значимым с надежностью $\gamma=1-\alpha$.

3.3. Модели дисперсионного анализа в исследовании сводной таблицы успеваемости

В качестве примера прикладного исследования на основе моделей дисперсионного анализа изучаются данные из сводной ведомости успеваемости студентов факультета по итогам промежуточной аттестации.

Генеральная совокупность (весь контингент бакалавриата) разбита на группы по направлениям подготовки (Математика (1), Прикладная математика (2), Прикладная математика и информатика (3)) и курсам (1, 2, 3, 4). Оцениваемые отклики в модели – относительное значение успеваемости (*ABS*, %), относительное значение качества (*QUO*, %). Статистические данные по двум промежуточным аттестациям (зимняя и летняя сессии) представлены в соответствующих таблицах.

Цель исследования: выяснить, имеются ли существенные различия в средних показателях успеваемости в выделенных группах. В качестве возможных факторов, определяющих различие средних, рассматриваются направления подготовки (т. е. различные категории обучающихся) и курсы.

3.3.1. Одномерная однофакторная модель анализа итогов промежуточной аттестации

Цель анализа: установить влияние факторов *Direction* (1, 2, 3) и *Course* (1,2,3,4) на показатель абсолютной успеваемости *ABS*, а затем на показатель качества успеваемости. Таким образом, будут рассмотрены четыре модели анализа с одним фактором и одной зависимой переменной.

В таблице 3.4 приведены данные по итогам одной промежуточной аттестации (зимняя сессия).

Таблица 3.4. Сводные данные успеваемости в зимнюю сессию.

№	Direction	Course	ABS	QUA
1	М	1	**72,46**	**27,54**
2	М	2	**40,38**	**19,23**
3	М	3	**54,9**	**27,45**
4	М	4	**34,48**	**10,34**
5	ПМ	1	**45,24**	**26,19**
6	ПМ	2	**61,9**	**33,33**
7	ПМ	3	**76,47**	**41,18**
8	ПМ	4	**50**	**50**
9	ПМИ	1	**81,82**	**52,73**
10	ПМИ	2	**62,5**	**50**
11	ПМИ	3	**52,94**	**38,24**
12	ПМИ	4	**57,14**	**50**

а) Проверка нулевой гипотезы $H_0: \mu_1 = \mu_2 = \mu_3$ (отсутствие различия средних показателей абсолютной успеваемости *ABS* по группам фактора *Direction* (1, 2, 3) – направление подготовки). Гистограмма распределения переменной *ABS* представлена на рисунке 3.1.

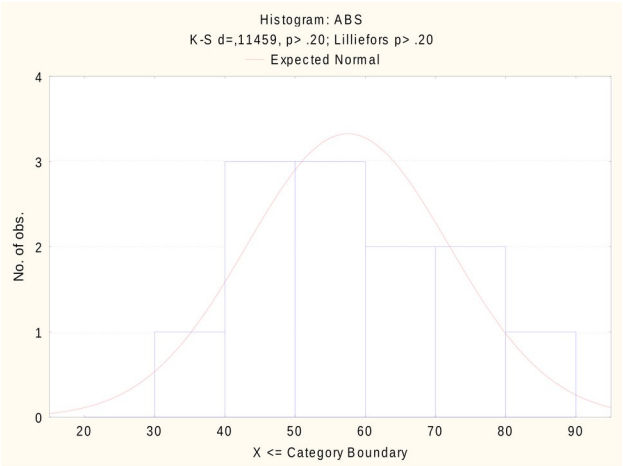

Рис. 3.1. Гистограмма переменной ABS.

Выборочные оценки средних показателей абсолютной успеваемости по трем группам составляют:

$$\bar{y}_1 = ABS(1) = 50,56;\ \bar{y}_2 = ABS(2) = 58,40;\ \bar{y}_3 = ABS(3) = 63,6,$$

при равных дисперсиях $s_1^2 = 286,62;\ s_2^2 = 194,04;\ s_3^2 = 162,82$. График средних представлен на рисунке 3.2. Наблюдаемое различие средних показывает наиболее высокую абсолютную успеваемость по направлению ПМИ(3).

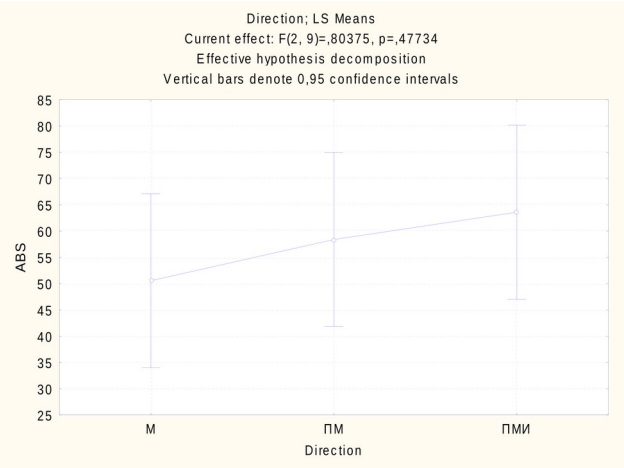

Рис. 3.2. График средних показателей абсолютной успеваемости *ABS* по уровням фактора *Direction* (1,2,3).

В таблице результатов однофакторного дисперсионного анализа (табл. 3.5) представлены компоненты общей суммы квадратов отклонений и *F*-отношение.

Таблица 3.5. Результаты однофакторного дисперсионного анализа по фактору *Direction*.

Источник вариации	Суммы квадратов, *SS*	Степени свободы, ν	Средние квадраты, *MS*	*F*	*p*
Между группами	345,03	2	172,51	0,8038	0,4773
Внутри групп	1931,71	9	214,63		
Всего	2276,74	11			

Поскольку вероятность *F*-отношения $p > 0,05$, нулевая гипотеза принимается, т. е. *различие средних показателей абсолютной успеваемости по направлениям подготовки не является существенным*.

б) Проверка нулевой гипотезы $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ (отсутствие различия средних показателей абсолютной успеваемости *ABS* по группам фактора *Course* (1, 2, 3, 4) – курс). График средних показателей абсолютной успеваемости *ABS* представлен на рисунке 3.3.

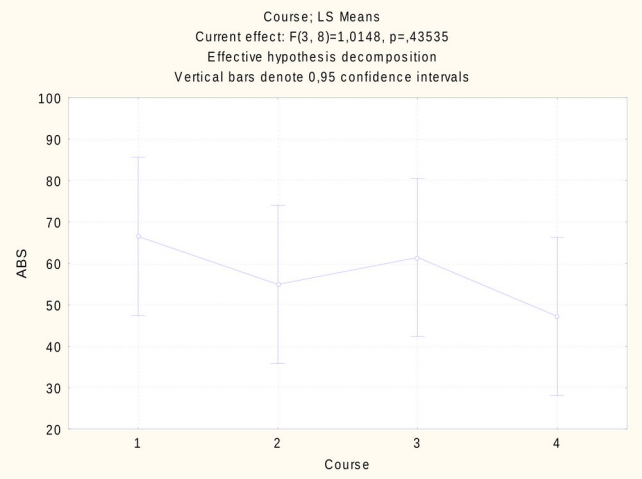

Рис. 3.3. График средних показателей абсолютной успеваемости *ABS* по уровням фактора *Course* (1,2,3,4).

Выборочные оценки средних показателей абсолютной успеваемости по четырем курсам составляют:

$\bar{y}_1 = ABS(1) = 66,51;\ \bar{y}_2 = ABS(2) = 54,93;\ \bar{y}_3 = ABS(3) = 61,43;\ \bar{y}_4 = ABS(4) = 47,21$

при дисперсиях $s_1^2 = 361,02;\ s_2^2 = 158,76;\ s_{31}^2 = 170,46;\ s_4^2 = 134,09$.

В таблице результатов однофакторного дисперсионного анализа (табл. 3.6) представлены компоненты общей суммы квадратов отклонений и *F*-отношение.

Таблица 3.6. Результаты однофакторного дисперсионного анализа по фактору *Course*.

Источник вариации	Суммы квадратов, *SS*	Степени свободы, ν	Средние квадраты, *MS*	*F*	*p*
Между группами	627,55	3	209,19	1,0148	0,4353
Внутри групп	1649,16	8	206,15		
Всего	2276,74	11			

Поскольку вероятность *F*-отношения $p > 0,05$, нулевая гипотеза принимается, т. е. *различие средних показателей абсолютной успеваемости по направлениям подготовки не является существенным*.

в) Проверка нулевой гипотезы $H_0: \mu_1 = \mu_2 = \mu_3$ (отсутствие различия средних показателей качества успеваемости *QUA* по группам фактора *Direction* (1, 2, 3) – направление подготовки). График средних приведен на рисунке 3.5.

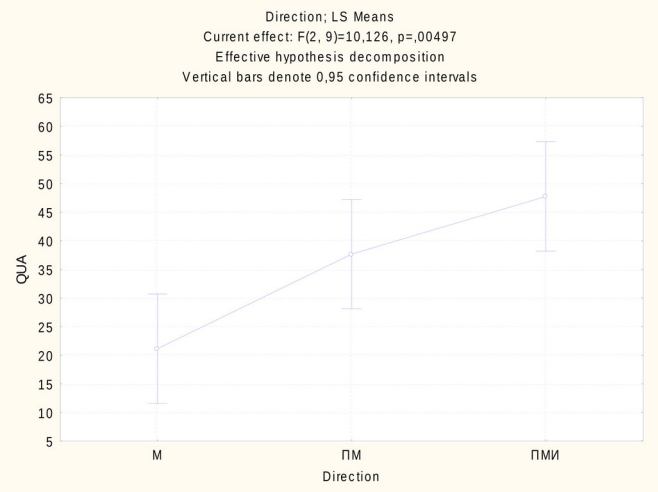

Рис. 3.5. График средних показателей качества успеваемости *QUA* по уровням фактора *Direction* (1,2,3).

Выборочные оценки средних показателей качества успеваемости по трем направлениям составляют:

$$\bar{y}_1 = QUA(1) = 21,14;\ \bar{y}_2 = QUA(2) = 37,67;\ \bar{y}_3 = QUA(3) = 47,74$$

при дисперсиях $s_1^2 = 66,91;\ s_2^2 = 104,85;\ s_{31}^2 = 41,73.$

Результаты дисперсионного анализа представлены в итоговой таблице 3.7.

Таблица 3.7. Результаты однофакторного дисперсионного анализа по фактору *Direction.*

Источник вариации	Суммы квадратов, *SS*	Степени свободы, ν	Средние квадраты, *MS*	*F*	*p*
Между группами	1443,27	2	721,64	10,1258	0,0049
Внутри групп	641,40	9	71,21		
Всего	2084,68	11			

Оказалось, что в данном случае вероятность *F*-отношения $p < 0,05$, и нулевая гипотеза отклоняется, т. е. *различие средних показателей качества успеваемости по направлениям подготовки является существенным*.

г) Проверка нулевой гипотезы $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ (отсутствие различия средних показателей качества успеваемости *QUA* по группам фактора *Course* (1, 2, 3, 4) – курс). График средних показателей абсолютной успеваемости *ABS* представлен на рисунке 3.6.

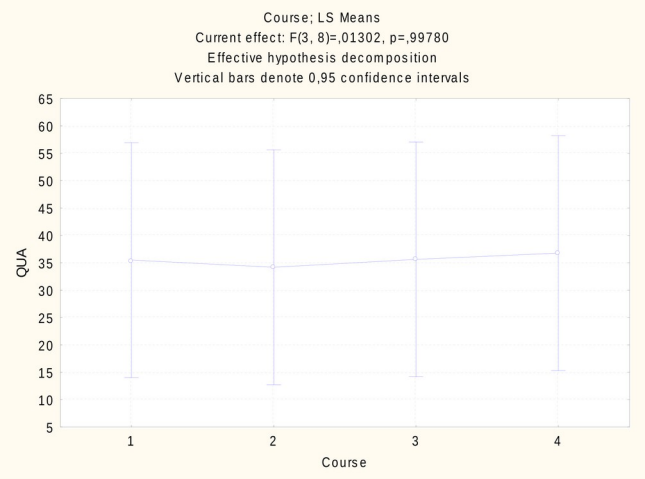

Рис. 3.6. График средних показателей качества успеваемости *QUA* по уровням фактора *Course* (1,2,3,4).

Выборочные оценки средних показателей качества успеваемости по четырем курсам составляют:

$\bar{y}_1 = QUA(1) = 35,48;\ \bar{y}_2 = QUA(2) = 34,18;\ \bar{y}_3 = QUA(3) = 35,62;\ \bar{y}_4 = QUA(4) = 36,78$

при дисперсиях $s_1^2 = 223,20;\ s_2^2 = 237,16;\ s_{31}^2 = 52,27;\ s_4^2 = 523,95$.

Результаты дисперсионного анализа представлены в итоговой таблице 3.8.

Таблица 3.8. Результаты однофакторного дисперсионного анализа по фактору *Course*.

Источник вариации	Суммы квадратов, *SS*	Степени свободы, ν	Средние квадраты, *MS*	*F*	*p*
Между группами	10,13	3	3,38	1,0148	0,9977
Внутри групп	2074,54	8	259,32		
Всего	2084,68	11			

Поскольку вероятность *F*-отношения $p > 0,05$, нулевая гипотеза принимается, т. е. *различие средних показателей качества успеваемости по направлениям подготовки не является существенным*.

Вывод по моделям 3.3.1. В целом по итогам одной промежуточной аттестации не выявлено существенного различия показателей абсолютной успеваемости как по трем различным направлениям подготовки (и

связанных с ними категориям обучающихся), так и по четырем курсам обучения. Для показателей качества успеваемости выявлено существенное различие по направлениям подготовки, причем наилучший показатель качества определен для направления ПМИ(3); по курсам существенных различий качества успеваемости нет.

3.3.2. Одномерная двухфакторная модель анализа итогов промежуточной аттестации

Цель анализа: установить влияние двух факторов *Direction* (1, 2, 3) и *Course* (1,2,3,4) и их взаимодействия на показатель абсолютной успеваемости *ABS.* В таблице 3.9 приведены данные по итогам двух промежуточных аттестаций (зимняя и летняя сессии).

Таблица 3.9. Сводные данные успеваемости за 2 сессии.

№	Direction	Course	ABS	QUO
1	М	1	**72,46**	**27,54**
2	М	1	**48,53**	**32,35**
3	М	2	**40,38**	**19,23**
4	М	2	**40**	**26**
5	М	3	**54,9**	**27,45**
6	М	3	**38**	**36**
7	М	4	**34,48**	**10,34**
8	М	4	**78,57**	**10,71**
9	ПМ	1	**45,24**	**26,19**
10	ПМ	1	**38,1**	**21,43**
11	ПМ	2	**61,9**	**33,33**
12	ПМ	2	**54,55**	**50**
13	ПМ	3	**76,47**	**41,18**
14	ПМ	3	**47,06**	**35,29**
15	ПМ	4	**50**	**50**
16	ПМ	4	**83,33**	**0**
17	ПМИ	1	**81,82**	**52,73**
18	ПМИ	1	**40,74**	**25,93**
19	ПМИ	2	**62,5**	**50**
20	ПМИ	2	**54,84**	**38,71**
21	ПМИ	3	**52,94**	**38,24**
22	ПМИ	3	**51,43**	**28,57**
23	ПМИ	4	**57,14**	**50**
24	ПМИ	4	**85,71**	**21,43**

а) Одномерная модель с зависимой переменной *ABS.* Наблюдения образуют полный факторный план 3х4 (первый фактор Направление, *Direction*, изменяется на 3 уровнях, фактор Курс, *Course,* – на 4 уровнях).

График средних по двум факторам представлен на рисунке 3.7.

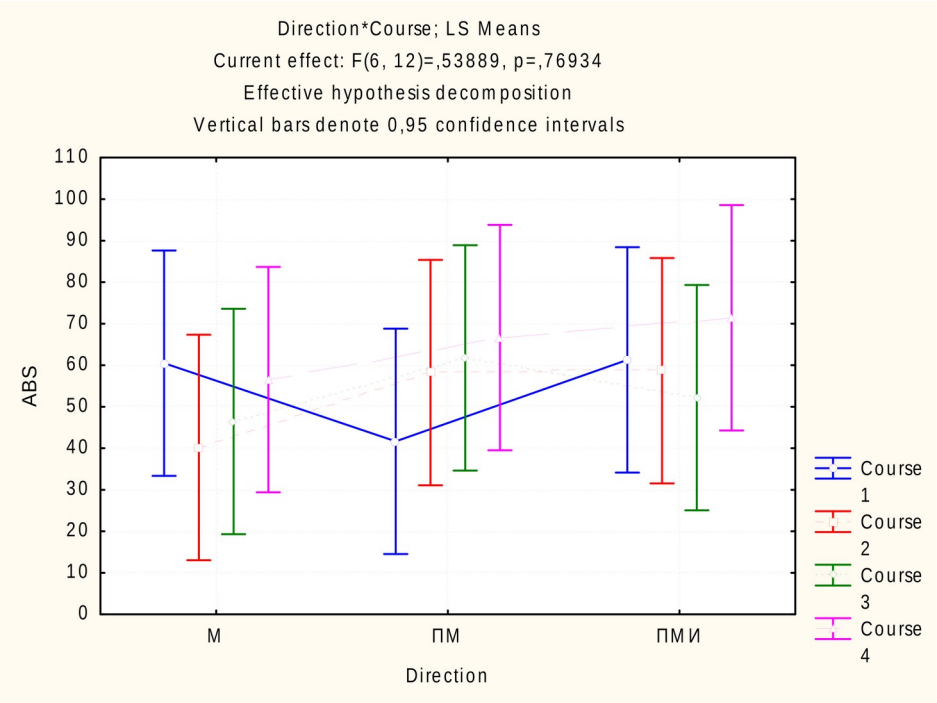

Рис. 3.7. График средних показателей абсолютной успеваемости ABS по градациям факторов *Direction* (1,2,3) и *Course* (1,2,3,4).

Оценки дифференциальных эффектов и взаимодействия приведены в таблице 3.10.

Таблица 3.10. Результаты двухфакторного анализа по факторам *Direction* и *Course*.

Эффекты	Суммы квадратов, *SS*	Степени свободы, *ν*	Средние квадраты, *MS*	*F*	*p*
Direction	405,41	2	202,71	0,6531	0,5379
Course	601,91	3	200,64	0,6465	0,5999
*Direction*Course*	1003,40	6	167,23	0,5389	0,7693
Error	3723,97	12	310,33		

Гипотеза о равенстве средних принимается по значениям *F*-критерия, поскольку все вероятности превышают 0,05.

Вывод по модели 3.3.2: показатели абсолютной успеваемости и качества не показывают статистически значимой (существенной) зависимости от факторов направления подготовки (*Direction*) и курса обучения (*Course*) и их взаимодействий.

3.3.3. Многомерная двухфакторная модель анализа итогов промежуточной аттестации

Цель анализа: установить влияние факторов *Direction*, *Course* на совокупность показателей абсолютной успеваемости и ее качества по промежуточного контроля.

Для данных таблицы 3.9 итоговой успеваемости можно формально выполнить анализ на основе многомерной модели с двумя зависимыми переменными (*ABS*, *QUA*) и двумя факторами (*Direction*, *Course*). Этот тип анализа может показать влияние факторов на группу переменных в целом.

Переменные *ABS* и *QUA*, как установлено по данным таблицы 3.9, не коррелируют между собой (r=0,0).

Таблица 3.11 итогов многомерного анализа включает значения Λ-критерия Уилкса по каждому фактору и их взаимодействию.

Таблица 3.11. Результаты многомерного анализа по факторам *Direction* и *Course*.

Эффекты	*Test*	значение	*F*	Степени свободы *эффектов*	Степени свободы *ошибки*	*p*
Direction	*Wilks*	0,6254	1,4545	4	22	0,2498
Course	*Wilks*	0,7555	0,5515	6	22	0,7635
*Direction*Course*	*Wilks*	0,5831	0,5675	12	22	0,8446

Значение Λ-критерия Уилкса и вероятностей *F*-распределения не отклоняют гипотезу об отсутствии различий средних в группах, что согласуется с результатами и интерпретацией одномерного анализа (пп. 3.3.1 и 3.3.2).

Вывод по модели 3.3.3: совокупность средних значений успеваемости и качества не показывает существенного различия по группам направлений подготовки и курсам обучения по программам бакалавриата.

3.4. Анализ данных опроса об удовлетворенности программой

В качестве прикладного исследования на основе моделей одномерного и многомерного дисперсионного анализа рассмотрена задача изучения оценки содержания магистерской программы обучающимися.

Был проведен анкетный опрос обучающихся выпускного курса магистратуры по направлению 01.04.02. Прикладная математика и информатика. Всего удалось опросить 21 человека.

В анкете первая группа вопросов (четыре вопроса) относится к статусу обучающегося:

1) совмещение учебы с работой (переменная *WORK*(0, 1, 2));
2) средний балл успеваемости (переменная *BALL*(3, 4, 5));
3) относительное количество пропусков занятий (переменная *OUT*(3, 4, 5), где «3» соответствует пропуску от 25 до 50% занятий, «4» – от 50 до 75%, «5» – свыше 75%);

4) магистерская программа (переменная *PROG*(1, 2)), где «1» есть программа «Системное программирование и информационное моделирование» (СПИМ), «2» – «Теоретическая информатика и кибернетика» (ТИК).

Эти переменные в моделях анализа считаются независимыми, т. е. играют роль факторов с соответствующим числом уровней.

Вторая группа вопросов (шесть вопросов) предлагает оценить в пятибалльной шкале ряд показателей, представленных в таблице 3.12. Там же заданы имена соответствующих зависимых переменных – откликов.

Таблица 3.12. Вопросы анкеты.

№	Содержание вопроса	Переменная
5)	Удовлетворенность содержанием магистерской программы	*SATISFAC*
6)	Качество преподавания непрофильных дисциплин	*QN*
7)	Качество преподавания профильных дисциплин	*QP*
8)	Полезность дисциплин программы для выполнения ВКР	*UT_FQW*
9)	Полезность дисциплин программы для Вашей работы (настоящей или будущей)	*UT_W*
10)	Общая оценка факультета	*FMI*

Цель анализа: установить влияние статуса обучающегося, представленного переменными-факторами *WORK*(0, 1, 2), *BALL*(3, 4, 5), *PROG*(1, 2), *OUT*(3, 4, 5).

В корреляционной матрице количественных переменных, представленной в таблице 3.13, среди значимых на уровне 0,05 можно отметить корреляции переменной *FMI* с другими переменными модели.

Таблица 3.13. Корреляционная матрица.

	Correlations (Анализ анкет.st Marked correlations are signif N=21 (Casewise deletion of m							
Varia	BA	OU	SAT	QI	QI	UT_	UT	FM
BALL	1,0	-0,	0,1	0,1	0,3	0,2	0,2	0,1
OUT	-0,	1,0	0,0	0,1	0,3	0,3	-0,	0,0
SATI	0,1	0,0	1,0	0,5	0,2	0,2	0,0	0,4
QN	0,1	0,1	0,5	1,0	0,5	0,5	0,3	0,7
QP	0,3	0,3	0,2	0,5	1,0	0,5	-0,	0,5
UT_F	0,2	0,3	0,2	0,5	0,5	1,0	0,5	0,6
UT_W	0,2	-0,	0,0	0,3	-0,	0,5	1,0	0,4
FMI	0,1	0,0	0,4	0,7	0,5	0,6	0,4	1,0

Средние значения и стандартные отклонения по всем переменным модели, за исключением переменной *PROG*, приведены в таблице 3.14.

Таблица 3.14. Дескриптивная статистика.

	Descriptive Statistics (A				
Variat	Vali	Mea	Minir	Maxi	Std.
BALL	21	4,0	3	5	0,7
OUT	21	52,3	25	75	22,2
SATI	21	3,9	3	5	0,6
QN	21	4,1	2	5	0,6
QP	21	4,7	4	5	0,4
UT_F	21	4,1	3	5	0,8
UT_V	21	4,0	2	5	0,9
FMI	21	4,5	3	5	0,6

3.4.1. Одномерная модель удовлетворенности для зависимой переменной *SATISFAC* (удовлетворенность содержанием программы) с четырьмя факторами и оценкой главных эффектов (*Main effects ANOVA*). Средние значения по градациям факторов представлены в таблице 3.15.

Таблица 3.15. Групповые средние для переменной *SATISFAC*.

	Descriptive Statistics (Анализ анкет.sta						
Effe	Leve Fact	N	SATIS Mear	SATIS Std.D	SATIS Std.E	SATIS -95,0	SATIS +95,0
Tota		2	3,952	0,66	0,14	3,64	4,25
WO	0	9	3,556	0,52	0,17	3,15	3,96
WO	1	3	4,667	0,57	0,33	3,23	6,10
WO	2	9	4,111	0,60	0,20	3,64	4,57
BAL	3	5	3,800	0,44	0,20	3,24	4,35
BAL	4	1	4,000	0,81	0,25	3,41	4,58
BAL	5	6	4,000	0,63	0,25	3,33	4,66
OU	25	7	3,857	0,69	0,26	3,21	4,49
OU	50	5	4,000	0,70	0,31	3,12	4,87
OU	75	9	4,000	0,70	0,23	3,45	4,54
PR	СПи	1	4,000	0,63	0,19	3,57	4,42
PR	ТИ	1	3,900	0,73	0,23	3,37	4,42

Таблица средних отражает некоторое различие значений переменной *SATISFAC* в группах, соответствующих уровням факторов, и позволяет сделать предположения о влиянии факторов на средние оценки удовлетворенности программой. Так, наибольшее различие средних наблюдается в группах по фактору *WORK*. Итоги дисперсионного анализа и разложение общей суммы квадратов представлены в таблице 3.16.

Таблица 3.16. Результаты одномерного анализа с четырьмя факторами (главные эффекты) для переменной *SATISFAC*.

	Univariate Results for Each DV Sigma-restricted parameterizat Effective hypothesis decompos				
Effe	Deg Free	SATI SS	SATIS MS	SATIS F	SATIS p
Inte	1	232,1	232,3	561,5	0,00
WO	2	3,24	1,62(	3,91	0,04
BAL	2	0,28	0,14(	0,33	0,71
OU	2	0,00	0,00:	0,00	0,99
PRC	1	0,00	0,00(	0,01	0,90
Errc	13	5,37	0,41:		
Tota	20	8,95			

Как видно из таблицы итоговой *существенным* является только различие средних по группам фактора *WORK*.

Межгрупповые различия по другим факторам *не являются существенными*, так как все вероятности *F*-отношений превышают 0,05; гипотеза о равенстве средних для них принимается.

Вывод по модели 3.4.1: на оценку удовлетворенности содержанием магистерской программы оказывает только фактор *WORK*(0,1,2) совмещения учебы с работой.

3.4.2. Одномерная модель полезности для зависимой переменной *UT_W* (полезность дисциплин программы для работы) с четырьмя факторами и оценкой главных эффектов (*Main effects ANOVA*). Средние значения по градациям факторов представлены в таблице 3.17.

Таблица 3.17. Групповые средние для переменной *UT_W*.

	Descriptive Statistics (Анализ анкет.sta)						
Effec	Level Factc	N	UT_\ Mea	UT_\ Std.D	UT_\ Std.E	UT_\ -95,0	UT_\ +95,C
Tota		21	4,00(	0,94	0,20	3,56	4,43
WOF	0	9	3,88	0,92	0,30	3,17	4,60
WOF	1	3	3,33	0,57	0,33	1,89	4,76
WOF	2	9	4,33	1,00(	0,33	3,56	5,10
BAL	3	5	3,40(	1,14(	0,50	1,98	4,81
BAL	4	10	4,20(	0,91	0,29(	3,54	4,85
BAL	5	6	4,16	0,75	0,30	3,37	4,95
OUT	25	7	4,00(	1,00(	0,37	3,07	4,92
OUT	50	5	4,00(	0,70	0,31	3,12	4,87
OUT	75	9	4,00(	1,11	0,37	3,14(	4,85
PRC	СПИ	11	4,45	0,82(	0,24	3,90	5,00
PRC	ТИК	10	3,50(	0,84	0,26	2,89	4,10

Итоги дисперсионного анализа и разложение общей суммы квадратов представлены в таблице 3.18.

Таблица 3.18. Результаты одномерного анализа с четырьмя факторами (главные эффекты) для переменной *UT_W*.

Effect	Univariate Results for Each D Sigma-restricted parameteriz Effective hypothesis decomp				
	Degr. Freec	UT_ SS	UT_ MS	UT_ F	UT_ p
Interc	1	203,(	203,(	293,:	0,0(
WOR	2	1,0!	0,5:	0,7!	0,4!
BALL	2	1,6!	0,8-	1,2:	0,3:
OUT	2	1,3(	0,6!	0,9!	0,3!
PRO(	1	4,2'	4,2'	6,1'	0,0:
Error	1	9,0(	0,6!		
Total	2	18,0			

Здесь значимое различие средних имеется только по группам фактора *PROG*: обучающиеся по программе СПИМ дают более высокую среднюю оценку полезности дисциплин программы для работы.

По другим зависимым переменных существенных различий средних по группам факторов не установлено.

Вывод по модели 3.4.2: на оценку полезности дисциплин программы для практической работы оказывает только фактор *PROG*(СПИМ, ТИК) программы обучения.

3.4.3. Многомерная модель для зависимых переменных *SATISFAC* (удовлетворенность содержанием программы) и *UT_W* (полезность дисциплин программы для работы). Из числа независимых переменных (факторов) в модель включены *WORK*(0,1,2) (совмещение учебы с работой) и *BALL*(3,4,5) (балл успеваемости) и их взаимодействие.

Таблица средних значений зависимых переменных (табл. 3.19) по группам уровней факторов включает те же значения, которые были получены в одномерных тестах (табл. 3.15, 3.17).

Для сочетаний уровней факторов *WORK* BALL* также представлены групповые средние. Так, наивысшая оценка удовлетворенности программой (5 баллов) соответствует сочетанию 1*4 (группа работающих не по специальности с успеваемостью 4); наивысшая оценка полезности дисциплин программы для работы (4,75) соответствует сочетанию 2*4 (группа работающих по специальности с успеваемостью 4).

Таблица 3.19. Средние значения *SATISFAC* и *UT_W* в многомерном анализе

Effect	Descriptive Statistics (Анализ анкет.sta)						
	Leve of Factc	Leve of Fact	N	SATISF Mear	SATISF Std.De	UT_W Mean	UT_W Std.De
Total			2:	3,9523	0,669(	4,000(	0,948(
WORK	(		9	3,5555	0,527(	3,888	0,927
WORK	1		3	4,6666	0,577	3,333	0,577
WORK	2		9	4,1111	0,600	4,333	1,000(
BALL	3		5	3,800(	0,447	3,400(	1,140
BALL	4		1(	4,000(	0,816	4,200(	0,918
BALL	5		6	4,000(	0,632	4,166	0,752
WORK*BA	(	3	2	4,000(	0,000(	4,000(	1,414
WORK*BA	(	4	4	3,250(	0,500(	4,000(	1,154
WORK*BA	(	5	3	3,666	0,577	3,666	0,577
WORK*BA	1	3	1	4,000(		3,000(	
WORK*BA	1	4	2	5,000(	0,000(	3,500(	0,707
WORK*BA	2	3	2	3,500(	0,707	3,000(	1,414
WORK*BA	2	4	4	4,250(	0,500(	4,750(	0,500(
WORK*BA	2	5	3	4,333	0,577	4,666	0,577

На рисунке 3.8 представлены графики средних переменной *SATISFAC* по группам взаимодействующих факторов *WORK* BALL*. Эти графики позволяют визуально оценить различие средних.

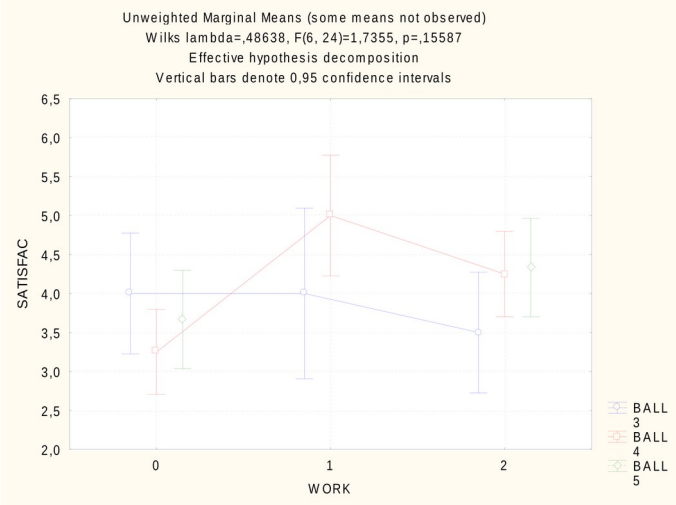

Рис. 3.8. Графики средних по переменной *SATISFAC*.

Итоговая таблица многомерного анализа (табл. 3.20), содержащая значения Λ-Уилкса, не показывает существенного различия средних (все вероятности превышают 0,05), в то время как одномерный анализ показал (табл. 3.16) существенность различия средних по группам фактора *WORK*.

Таблица 3.20. Результаты многомерного анализа с двумя факторами для переменных *SATISFAC* и *UT_W*.

Effect	Multivariate Tests of Sigma-restricted par Effective hypothesis					
	Te	Val	F	Ef d1	Er d1	p
Interc	W	1,0(		(		
WOR	W	0,8:	1,4(	:	1	0,2:
BALL	W	0,7	1,7	:	1	0,2:
WOR	W	0,4:	1,7:	(	2	0,1:

Выводы по модели 3.4.3: Многомерный дисперсионный анализ не установил существенности различия средних по совокупности переменных удовлетворенности программой обучения и полезности содержания дисциплин по группам факторов совмещения работы с учебой и балла успеваемости.

Заключение

В выпускной квалификационной работе приведено систематическое изложение основ дисперсионного анализа как метода выявления закономерностей в массивах данных различной природы.

В теоретической части работы приведена общая линейная модель для одномерного анализа *ANOVA* и обобщенная линейная модель многомерного анализа *MANOVA*. Приведены программы статистических процедур проверки одномерных и многомерных гипотез, перечислены основные критерии, представлены структуры итоговых таблиц анализа и правила получения статистических выводов, что представляет интерес для практических задач анализа данных.

В прикладной части работы представлены этапы и результаты исследований фактического материала, касающегося образовательной области. Первое исследование имело целью анализ итогов промежуточной аттестации (показатели успеваемости и качества) по различным направлениям подготовки и четырем курсам программы бакалавриата. Примененные модели одномерного однофакторного и двухфакторного анализа в целом не выявили существенности различия средних показателей успеваемости и качества по группам факторов направления и курса.

Второе исследование проведено по итогам опроса обучающихся по программам магистратуры с целью выявления средних оценок удовлетворенности содержанием обучения. Подготовленные анкеты опроса и его результаты послужили статистическим материалом для анализа. Результаты применения двух одномерных и многомерной модели анализа в целом не выявили существенного различия оценок удовлетворенности программой обучения по категориям обучающихся, совмещающих учебу с работой, как и по показателям их успеваемости.

Сценарии анализа данных и полученные результаты могут использоваться в учебном процессе как примеры проведения исследований с использованием моделей дисперсионного анализа.

Список использованной литературы

1. Аренс, Х. Многомерный дисперсионный анализ. Пер. с немецкого В. М. Ивановой и Ю. Н. Тюрина / Х. Аренс, Ю. Лёйтер. -– М.: Финансы и статистиска, 1985. -– 230 с.
2. Афифи, А. Статистический анализ: Подход с использованием ЭВМ. Пер с англ. /А. Афифи, С. Эйзен. – М. : Мир, 1982. -– 488 с.
3. Бендат, Дж. Прикладной анализ случайных данных / Дж. Бендат, А. Пирсол — М.: Мир,1989. — 540 с.
4. Боровиков, В. П. Искусство анализа данных. 2-е издание / В. П. Боровиков — СПб.: ПИТЕР, 2005.
5. Вуколов Э. А. Основы статистического анализа. Практикум по статистическим методам и исследованию операций с использованием пакетов STATISTICA и EXCEL.: учебное пособие. / Э. А. Вуколов. -– М. : ФОРУМ, 2008. -– 464 с.
6. Крамер, Г. Математические методы статистики. Пер. с английского под ред. А. Н. Колмогорова / Г. Крамер. -– М. : Мир, 1975. -– 625 с.
7. Львовский, Е. Н. Статистические методы построения эмпирических формул: Учеб. пособие для втузов. / Е. Н. Львовский. -– М. : Высшая школа, 1988. -– 239 с.
8. Математическая теория планирования эксперимента. / Под ред. С. М. Ермакова. -– М. : Наука, 1983. -– 392 с.
9. Многомерный статистический анализ в экономике: Учеб. пособие для вузов / Под ред. проф. В. Н. Тамашевича. — М. : ЮНИТИ-ДАНА, 1999. — 598 с.
10. Слепко, Ю. Н.. Анализ данных и интерпретация результатов психологического исследования: учебное пособие / Ю. Н. Слепко, Т. В. Ледовская. — Ярославль: Изд-во ЯГПУ, 2013. — 136 с.
11. Тюрин, Ю. Н. Анализ данных на компьютере: учеб. пособие для студентов вузов / Ю. Н. Тюрин, А. А. Макаров. – Изд. 4-е, перераб. – М. : Форум, 2008. -– 366 с.
12. Фишер, Р. Э. Статистические методы для исследователей / Р. Э. Фишер. — М. : Госстатиздат, 1958. — 267 с.
13. Шеффе Г. Дисперсионный анализ, пер. с англ./ Г. Шеффе. — М., 1963.
14. Электронный ресурс http://www.machinelearning.ru/wiki/index.php?