

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное
учреждение высшего образования
«СЕВАСТОПОЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Кафедра «Информационные технологии и компьютерные системы»

СТАТИСТИЧЕСКИЙ ПАКЕТ SPSS ДЛЯ СОЦИАЛЬНЫХ НАУК

Методические указания
для выполнения практических работ по дисциплине
для студентов очной формы обучения направления подготовки
39.03.01 «Социология»

В трех частях

Часть III

**Анализ взаимосвязей между данными
в статистическом пакете PSPP**

Севастополь
СевГУ
2025

УДК 316:004.428(076.5)
ББК 60.5:32.973-018.2я73
С78

Р е ц е н з е н т :
Е.Н. Машенко - доцент кафедры ИТиКС

Составитель: И. А. Балакирева

С78 **Статистический пакет SPSS для социальных наук** : методические указания к выполнению практических работ по дисциплине для студентов очной формы обучения направления подготовки 39.03.01 «Социология». – В 3-х частях. Часть III. Анализ взаимосвязей между данными в статистическом пакете PSPP / Севастопольский государственный университет, кафедра «Информационные технологии и компьютерные системы» ; сост. И. А. Балакирева. – Севастополь : СевГУ, 2025. – 63 с. – Текст : электронный.

Целью методических указаний является оказание помощи студентам в применении статистического пакета PSPP для анализа взаимосвязей между данными в социологических исследованиях.

УДК 316:004.428(076.5)
ББК 60.5:32.973-018.2я73

Методические указания рассмотрены и утверждены на заседании кафедры ИТиКС, протокол № 8 от 11 марта 2025 г.

© Балакирева И. А., сост., 2025
© ФГАОУ ВО «Севастопольский
государственный университет», 2025

СОДЕРЖАНИЕ

1	Практическая работа № 1. Исследование взаимосвязей между данными на основе корреляционного анализа в статистическом пакете PSPP.....	4
2	Практическая работа № 2. Исследование взаимосвязей между данными на основе однофакторного дисперсионного анализа в статистическом пакете PSPP.....	9
3	Практическая работа № 3. Непараметрические критерии в статистическом пакете PSPP.....	17
4	Практическая работа № 4. Линейный регрессионный анализ в статистическом пакете PSPP	27
5	Практическая работа № 5. Бинарная логистическая регрессия в статистическом пакете PSPP.....	39
6	Практическая работа № 6. Факторный анализ в статистическом пакете PSPP.....	44
	Приложение А. Таблица значений F-критерия Фишера при уровне значимости $\alpha=0,05$	57
	Приложение Б. Задания для самостоятельной работы.....	58

Практическая работа № 1

ИССЛЕДОВАНИЕ ВЗАИМОСВЯЗЕЙ МЕЖДУ ДАННЫМИ НА ОСНОВЕ КОРРЕЛЯЦИОННОГО АНАЛИЗА В СТАТИСТИЧЕСКОМ ПАКЕТЕ PSPP

Цель работы: научиться выявлять наличие взаимосвязи между данными на основе корреляционного анализа.

1. Анализ взаимосвязей между данными. Краткие теоретические сведения.

Исследуя явления в самых различных областях, статистика неизбежно сталкивается с зависимостями, как между количественными, так и качественным признаками.

Задача статистики – выявить такие зависимости и дать их количественную характеристику.

Взаимосвязанные признаки (переменные) делятся на 2 группы:

- независимые (факторные) - определяющие факторы, влияющие на изменение других признаков;
- зависимые (результативные) - результат влияния независимых признаков.

Зависимости между признаками могут быть двух видов:

- функциональные (детерминированные);
- статистические (недетерминированные, стохастические).

Зависимость между двумя признаками x и y называется функциональной, если определенному значению признака x однозначно соответствует значение признака y .

Зависимость называется статистической, если зависимый признак y определяется одновременным действием многих факторов и его значение при фиксированном значении x может варьироваться.

Статистические зависимости можно обнаружить только при массовом наблюдении. Исследование статистических зависимостей и есть суть анализа взаимосвязей.

Анализ взаимосвязей включает в себя:

- Корреляционный анализ
- Дисперсионный анализ
- Регрессионный анализ.

2 Корреляционный анализ.

Основные показатели КА – *коэффициент корреляции (количественная оценка)* и *диаграмма рассеяния (наглядность)*.

Если зависимость существует, то коэффициент корреляции позволяет также установить, является она прямой или обратной.

Прямая зависимость означает, что факторная и результативная переменная

возрастают или убывают одновременно.

Обратная зависимость присутствует тогда, когда факторная переменная возрастает, а результирующая - убывает и наоборот.

Коэффициент корреляции Пирсона (r -Пирсона) характеризует существование **линейной** связи между двумя величинами.

Условия применения корреляционного анализа:

1. Исследуемые переменные x и y должны быть распределены нормально.
2. Исследуемые переменные x и y должны быть измерены в числовой или порядковой шкале.
3. Количество значений в исследуемых переменных x и y должно быть одинаковым.

Коэффициент корреляции изменяется в диапазоне от -1 до +1. Считается, что если коэффициент корреляции по абсолютной величине больше 0,5, то связь между признаками заметная, иначе – связь отсутствует (рис. 1.1).

Сильная	Нет	Сильная
обратная	линейной	прямая
связь	связи	связь
-1	0	+1

Рисунок 1.1 – Характеристика связей между данными

Пример 1. Вычислить коэффициент корреляции Пирсона, построить диаграмму рассеяния и определить тип зависимости для двух признаков – возраста пациента и его артериального давления. Исходные данные приведены в таблице 1.1, файл **возраст_давление.sav**.

Таблица 1.1 - Результаты наблюдения пациентов.

Субъект	Возраст x_i	Давление y_i
1	43	128
2	48	120
3	56	135
4	61	143
5	67	141
6	70	152

Для построения диаграммы рассеяния необходимо выполнить команду **Graphs. Scatterplot.**

В результате имеем (рис. 1.2):

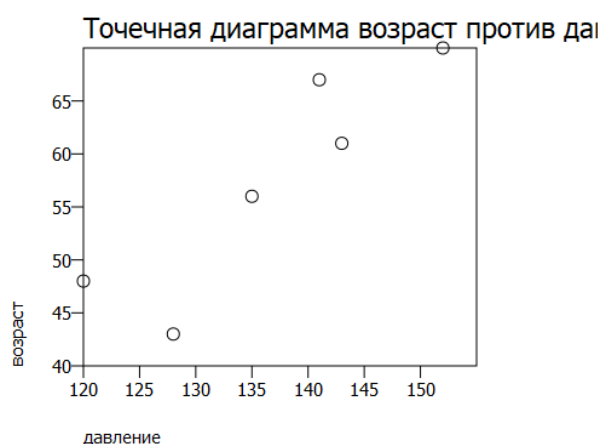


Рисунок 1.2 – Диаграмма рассеяния

Для исследования взаимосвязи между двумя переменными на основе корреляционного анализа необходимо выполнить команды **Анализ. Двумерная корреляция...**

В диалоговом окне нужно указать переменные, подлежащие исследованию и установить режимы (рис. 1.3).

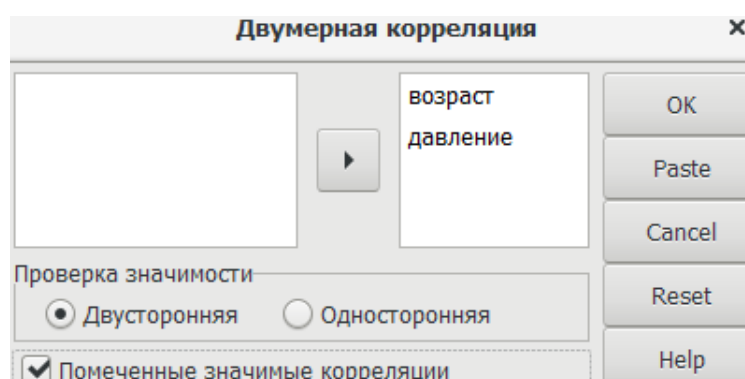


Рисунок 1.3 – Диалоговое окно корреляционного анализа

Результат представлен на рис. 1.4:

Корреляции			
		возраст	давление
возраст	Корреляция Пирсона	1,000	,897 ^a
	Знач. (двустороннее)		,015
	N	6	6
давление	Корреляция Пирсона	,897 ^a	1,000
	Знач. (двустороннее)	,015	
	N	6	6

a. Significant at .05 level

Рисунок 1.4 – Результаты корреляционного анализа

P-значение $0,015 < 0,05$, можно сделать вывод, что связь между переменными существенна. Поскольку коэффициент Пирсона $r = 0,897 > 0$, связь между переменными *Возраст пациента* и *Давление* прямая и достаточно сильная. По результатам анализа можно сделать вывод, что артериальное давление сильно зависит от возраста, при этом

с увеличением возраста увеличивается и давление.

2. Задания для выполнения практической работы

Материалы для выполнения задания находятся по ссылке:

<https://do.sevsu.ru/mod/folder/view.php?id=557496>

Задание 1. Выполнить пример 1 данных методических указаний.

Задание 2. Построить диаграмму рассеяния, вычислить коэффициент корреляции Пирсона и определить тип зависимости для двух признаков - количество пропущенных занятий и итоговая оценка.

Число пропусков	6	2	15	9	12	5	8
Итоговая оценка	82	86	43	74	58	90	78

Задание 3. Рассматриваются две переменные: «Продолжительность подготовки (в часах)» студентов перед экзаменом и «Итоговая оценка» (табл. 4.2). Построить диаграмму рассеяния, вычислить коэффициент корреляции Пирсона и определить тип зависимости для данных переменных.

Таблица 4.2 – Подготовка к экзамену.

Студент	Часы х	Оценка у
A	6	82
B	2	63
C	1	57
D	5	88
E	2	68
F	3	75

Задание 4. Определить, есть ли зависимость между интересом к моде (Q1) и тем, сколько денег человек тратит за один поход в магазин за одеждой (Q3). Исследование выполнить по данным файла fashion.sav.

Задание 5. В выборку вошли данные о 20 женщинах в возрасте от 25 до 30 лет (файл bodyfat). В наборе данных указаны следующие переменные:

у - количество жира в организме;

х1 - толщина кожной складки трицепса;

х2 - окружность бедра;

х3 - окружность предплечья.

На основе корреляционного анализа определить, есть ли зависимость между переменными:

а) толщина кожной складки трицепса и количество жира в организме;

б) окружность бедра и количество жира в организме;

в) окружность предплечья и количество жира в организме;

- г) толщина кожной складки трицепса и окружность бедра;
- д) толщина кожной складки трицепса и окружность предплечья;
- е) окружность бедра и окружность предплечья.

4. Содержание отчета

1. Цель работы.
2. Постановка задачи
3. Ход выполнения работы.
4. Анализ результатов (для каждого задания).
5. Общие выводы по работе.

Практическая работа № 2
ИССЛЕДОВАНИЕ ВЗАИМОСВЯЗЕЙ МЕЖДУ ДАННЫМИ
НА ОСНОВЕ ОДНОФАКТОРНОГО ДИСПЕРСИОННОГО АНАЛИЗА
В СТАТИСТИЧЕСКОМ ПАКЕТЕ PSPP

Цель работы: научиться выявлять наличие взаимосвязи между данными на основе однофакторного дисперсионного анализа.

1 Дисперсионный анализ. Теоретические сведения.

Дисперсионный анализ используют для изучения взаимосвязей, исследования влияния одной или нескольких качественных переменных на одну зависимую количественную переменную. Основной целью подобного исследования является установление значимости различия между средними значениями зависимой количественной переменной по группам фактора. Достигается это с помощью разложения общей дисперсии зависимой переменной на составляющие: общая дисперсия; межгрупповая дисперсия; средняя внутригрупповая дисперсия.

Величина **общей** дисперсии σ_o^2 характеризует вариацию признака под влиянием всех факторов, формирующих уровень признака у единиц данной совокупности, и определяется по формуле:

$$\sigma_o^2 = \frac{1}{k \cdot n} \sum_{j=1}^k \sum_{i=1}^n (y_{ij} - \bar{y})^2$$

где $\bar{y}$ - общая средняя арифметическая для всей изучаемой совокупности.

Межгрупповая дисперсия (дисперсия групповых средних - δ^2) отражает систематическую вариацию, т.е. различия в величине изучаемого признака, которые возникают под влиянием фактора, положенного в основу группировки.

$$\delta^2 = \frac{1}{k} \sum_{j=1}^k (\bar{y}_j - \bar{y})^2,$$

где $\bar{y}_j$ - средняя арифметическая по каждой группе, $j=1, k$.

Средняя внутригрупповая дисперсия $\bar{\sigma}^2$ характеризует случайную вариацию, возникающую под влиянием других, неучтенных факторов, и не зависит от признака-фактора, положенного в основу группировки:

$$\bar{\sigma}^2 = \frac{1}{k} \sum_{j=1}^k \sigma_j^2,$$

где дисперсия в каждой группе рассчитывается по формуле:

$$\sigma_j^2 = \frac{1}{n_j} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2,$$

Указанные дисперсии взаимосвязаны. Их связь между собой выражена равенством, которое называется правилом сложения дисперсий: **величина общей дисперсии равна сумме межгрупповой дисперсии и средней внутригрупповой**. Математическое выражение для правила сложения дисперсий представлено ниже:

$$\sigma_o^2 = \delta^2 + \bar{\sigma}^2.$$

Опираясь на это правило, можно определить, какая доля общей дисперсии складывается под влиянием признака-фактора, положенного в основу группировки δ^2 , то есть дисперсию за счет разбиения на группы и дисперсию за счет остальных факторов. Анализируя эти компоненты дисперсии, можно оценить долю воздействия каждого фактора на исследуемую величину.

В зависимости от количества факторов, определяющих вариацию признака-результата, дисперсионный анализ подразделяют на однофакторный и многофакторный.

Простейшим случаем дисперсионного анализа является однофакторный анализ, в котором n единиц совокупности распределены на m групп по значениям одного фактора. Основная идея однофакторного дисперсионного анализа заключается в сравнении дисперсии исследуемого признака, вызванной действием фактора, с дисперсией ошибок измерения этого признака. Если различие между ними значимо, то фактор оказывает существенное влияние на исследуемый признак.

В ходе расчетов вариации значений используются суммы квадратов отклонений от средних:

- **общая** вариация ***SSE*** (*the Sum of Squared estimate of Errors*);

вариация, объясняющая различия **между группами** ***SSB*** (*between groups*) - **межгрупповую сумму квадратов отклонений**

- вариация, объясняющую различия между единицами совокупности внутри группы ***SSW*** (*within groups*) - **внутригрупповую сумму квадратов отклонений**.

Поскольку суммы квадратов отклонений также описывают вариацию результирующего признака, можно представить выражением:

$$SSE = SSW + SSB.$$

Далее рассчитывается межгрупповой средний квадрат отклонений (на одну степень свободы) ***MSB***, а также внутригрупповой средний квадрат отклонений (на одну степень свободы) ***MSW***:

$$S_1^2 = \frac{SSB}{(m-1)} = MSB, \quad S_2^2 = \frac{SSW}{n-m} = MSW,$$

где n - число единиц анализируемой совокупности;

m - число групп, по которым распределены единицы совокупности.

Если различия внутригрупповых средних носят случайный характер, т.е. не

связаны с изменением факторного признака, то $S_1^2 = S_2^2$.

Это положение соответствует нулевой гипотезе **$H_0: S_1^2 = S_2^2$** .

Для выявления степени влияния изучаемого фактора X и сопоставления ее с межгрупповым разбросом значений вычисляют наблюдаемое значение критерия Фишера:

$$F_{расч} = \frac{S_1^2}{S_2^2}.$$

Наблюдаемое значение критерия сравнивают с критическим $F_{табл} = F(\alpha, df_1, df_2)$, которое находят по таблицам F-распределения для уровня значимости α , числа степеней свободы $df_1 = m - 1$ и $df_2 = n - m$. Таблица критических значений критерия Фишера приведена в Приложении А.

Если $F_{расч} \leq F_{табл}$, то влияние фактора X незначительно, и следует принять нулевую гипотезу. Следовательно, все полученные результаты измерений принадлежат одной генеральной совокупности с общим средним и дисперсией.

При $F_{расч} > F_{табл}$ влияние фактора принимают существенным и нулевую гипотезу отвергают. В этом случае полагают, что есть m нормально распределенных совокупностей, каждая из которых имеет свои среднее и групповую дисперсию.

Если установлено влияние, то связь между фактором, положенным в основу группировки можно измерить на основании эмпирического корреляционного отношения:

$$\eta_{yx} = \sqrt{\frac{SSB}{SSE}}$$

Этот показатель принимает значения в интервале от 0 до 1. Связь тем теснее, чем ближе значение η к 1, и наоборот (о направлении связи здесь речь не идет).

Для вывода (заключения) о тесноте связи можно воспользоваться следующими рекомендациями, сведенными в таблицу 1:

Таблица 2.1 – Шкала изменения тесноты связи результата с фактором

Значение эмпирического корреляционного	Теснота связи
0,3 и меньше	слабая
0,3 - 0,5	средняя
0,5 - 0,7	умеренно тесная
0,7 и больше	тесная

2 Примеры выполнения задания

Задание 1. Требуется проверить нулевую гипотезу о влиянии профессии рабочих на продолжительность заболеваний хроническим бронхитом, повлекшим временную нетрудоспособность. В ходе эксперимента были выделены следующие

профессиональные группы: станочники, слесари, прочие. Выборки сделаны на предприятии из 4 цехов с близкими условиями и характером труда работников. Предполагается, что распределение подчиняется нормальному закону.

Таким образом, в роли фактора (имеющего три уровня) выступает профессия рабочих, а в роли зависимого признака (результатирующего) – длительность заболевания в днях. Исходные данные файл *проф_заболевание.sav* (рис. 2.1):

Длительность заболевания			
Цех	станочники	Слесари	Прочие
1	10	5	4
2	14	11	6
3	9	7	11
4	12	8	4

а)

	Группа	Длит_забол
1	1	10,0
2	1	14,0
3	1	9,0
4	1	12,0
5	2	5,0
6	2	11,0
7	2	7,0
8	2	8,0
9	3	4,0
10	3	6,0
11	3	11,0
12	3	4,0

б)

Рисунок 2.1 – Исходные данные к заданию 1

До применения метода дисперсионного анализа необходимо выполнить проверку выборок на соответствие нормальному закону распределения.

Формулируем гипотезу **H₀**: выборки соответствуют нормальному распределению; **H₁**: выборки не соответствуют нормальному распределению.

Для проверки можно использовать таблицу на рис. 2.1б. В этом случае нужно последовательно выбирать наблюдения по группам и каждую группу проверять, используя критерий Колмогорова-Смирнова, на соответствие нормальному закону распределения.

Чтобы проверить одновременно все выборки, нужно создать в PSPP таблицу, изображенную на рис. 2.1а. В результате проверки имеем:

Тест Колмогорова-Смирнова для одной выборки

		станочники	слесари	прочие
N		4	4	4
Параметры нормального распределения	Среднее	11,25	7,75	6,25
	Станд. отклонение	2,22	2,50	3,30
Наибольшие разницы	Абсолютная	,21	,21	,28
	Положительная	,21	,21	,28
	Отрицательная	-,16	-,15	-,25
Колмогорова-Смирнова (Z)		,43	,42	,56
Асимпт. знач. (двусторонние)		,993	,994	,912

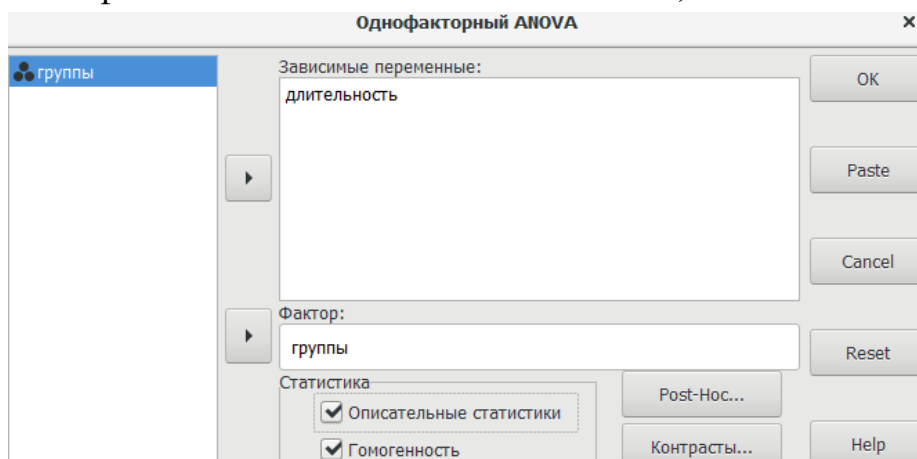
P-значения для всех выборок больше, чем 0,05. На основании этого можно сделать

вывод, все выборки подчинены нормальному закону распределения и для выяснения зависимости заболевания от профессии можно проводить дисперсионный анализ.

Формулируем гипотезу **H₀**: фактор, положенный в основу группировки влияния не оказывает, то есть длительность заболевания не зависит от профессии; **H₁**: Влияние фактора, положенного в основу группировки, существенно.

Для выполнения задания необходимо:

- выполнить команду **Анализ. Сравнение средних. Однофакторный ANOVA...**
- заполнить поля ввода диалогового окна:
- переменную *Длительность* – в **Зависимые переменные**;
- переменную *Группы* – в **Фактор**;
- установить флажок **Гомогенность** для проверки однородности дисперсий;
- установить флажок **Описательные статистики**, если это необходимо.



В результате имеем:

1. Описательные статистики по группам.

Описательные статистики						
группы		N	Среднее	Станд. отклонение	Станд. погрешность	95%-ы Ниж
длительность	станочники	4	11,25	2,22	1,11	
	слесари	4	7,75	2,50	1,25	
	прочие	4	6,25	3,30	1,65	
	Итого	12	8,42	3,29	,95	

2. Проверка однородности дисперсий.

Проверка однородности дисперсий

	Статистика Ливиня	df1	df2	Знач.
длительность	,24	2	9	<u>,788</u>

Р-значение больше 0,05, дисперсии считаем однородными.

3. Результаты дисперсионного анализа.

ANOVA

		Сумма квадратов	df	Среднее по квадратам	F	Знач.
длительность	Между группами	52,67	2	26,33	3,58	<u>,072</u>
	Внутри групп	66,25	9	7,36		
	Итого	118,92	11			

Расчетное значение критерия Фишера равно $F_{расч}=3,58$, а критическое

$F_{табл}=F(\alpha, df_1, df_2)=F(0,05, 2, 9)=4,26$ (Приложение А). Р-значение равно $0,072>0,05$.

4. Интерпретация результатов.

Поскольку $F_{табл} > F_{расч}$, то принимается нулевая гипотеза об отсутствии влияния фактора на значение зависимого признака. Кроме того, р-значение $p>0,05$, что тоже свидетельствует о том, что нулевую гипотезу надо принять.

Таким образом выяснено, что профессия рабочих не влияет на длительность заболевания.

Задание 2. Требуется определить, зависит ли время работы мобильных телефонов после полной зарядки от производителя. Рассматривается по 5 телефонов от каждого из 4-х различных производителей. Исходные данные приведены в таблице:

Порядковый номер телефона	Производитель 1	Производитель2	Производитель3	Производитель4
1	21,2	25,4	22,4	26
2	19,5	24,3	24	21,7
3	23,1	25,1	21,6	19
4	19,9	24,1	21,5	22,1
5	20,5	24,7	22,6	24,2

Для проверки выборок на соответствие нормальному распределению удобнее использовать файл **моб_телефоны_2**:

Наблюдения	proizv_1	proizv_2	proizv_3	proizv_4
1	21,2	25,4	22,4	26,0
2	19,5	24,3	24,0	21,7
3	23,1	25,1	21,6	19,0
4	19,9	24,1	21,5	22,1
5	20,5	24,7	22,6	24,2

1. Проверка нормальности распределения.

H_0 – выборки соответствуют нормальному распределению; H_1 – выборки не соответствуют нормальному распределению.

В результате проверки получим:

Тест Колмогорова-Смирнова для одной выборки

		proizv_1	proizv_2	proizv_3	proizv_4
N		5	5	5	5
Параметры нормального распределения	Среднее	20,84	24,72	22,42	22,60
	Станд. отклонение	1,42	,54	1,01	2,65
Наибольшие разницы	Абсолютная	,20	,18	,23	,17
	Положительная	,20	,18	,23	,17
	Отрицательная	-,17	-,16	-,18	-,17
Колмогорова-Смирнова (Z)		,45	,41	,51	,39
Асимпт. знач. (двусторонние)		,988	,997	,956	,998

Все 4 выборки соответствуют нормальному закону распределения, поскольку р-значение для каждой из них превышает уровень значимости 0,05.

2. Дисперсионный анализ проводится по таблице **моб_телефоны**:

Наблюдения	proizv	time
1	1	21,2
2	1	19,5
3	1	23,1
4	1	19,9
5	1	20,5
6	2	25,4
7	2	24,3

Переменная *proizv* соответствует номеру производителя, а переменная *time* – длительности работы телефона без подзарядки.

В качестве фактора выступает переменная *proizv*, а переменная *time* является зависимой.

В результате имеем:

Проверка однородности дисперсий				
	Статистика Ливиня	df1	df2	Знач.
time	3,01	3	16	,061

ANOVA						
		Сумма квадратов	df	Среднее по квадратам	F	Знач.
time	Между группами	38,08	3	12,69	4,91	,013
	Внутри групп	41,39	16	2,59		
	Итого	79,47	19			

Дисперсии выборок однородны, р-значение превышает уровень значимости, равный 0,05.

Расчетное значение критерия Фишера равно $F_{расч}=4,91$, а критическое $F_{табл}=F(\alpha, df_1, df_2)=F(0,05, 3, 16)=3,24$ (Приложение А). Р-значение равно $0,013 < 0,05$.

Поскольку $F_{табл} < F_{расч}$, то нулевая гипотеза об отсутствии влияния фактора на значение зависимого признака отвергается и принимается альтернативная гипотеза о наличии влияния. Кроме того, р-значение $p < 0,05$, что тоже свидетельствует о том, что нулевую гипотезу необходимо отклонить. Для определения тесноты связи можно воспользоваться эмпирическим корреляционным отношением η , равным отношению межгрупповой дисперсии к общей $\eta = 38,08/79,47 = 0,48$ – связь средняя.

Таким образом выяснено, что время работы телефона без подзарядки зависит от производителя.

3. Задания для выполнения практической работы

Материалы для выполнения задания находятся в приложении Б и по ссылке:

<https://do.sevsu.ru/mod/folder/view.php?id=558194>

1. Выполнить примеры 1 и 2 пункта 2.
2. Задание выполняется по вариантам.

I вариант выполняют студенты с нечетными вариантами в списке группы, используется файл **exam.sav**.

Данные состоят из экзаменационных баллов 42 студентов, посещающих занятия в некотором университете. Переменными являются:

ID: Идентификационный номер студента

Пол (gender): Пол студента

Экзамен (exam): Номер экзамена

Оценка (scores): Балл за студенческие экзамены

Выяснить, отличаются ли оценки студентов по разным экзаменам.

II вариант выполняют студенты с нечетными вариантами в списке группы, используется файл **mathassess.sav**. Выяснить, отличаются ли средние оценки студентов по вопросам о знаниях в различных колледжах.

3. Выяснить влияние фактора, положенного в основу группировки на значения средних в группах. Вариант задания выбрать из Приложения Б в соответствии с номером студента в списке группы.

4. Содержание отчета

1. Цель работы.
2. Постановка задачи
3. Ход выполнения работы.
4. Анализ результатов (для каждого задания).
5. Общие выводы по работе.

Практическая работа № 3

НЕПАРАМЕТРИЧЕСКИЕ КРИТЕРИИ В СТАТИСТИЧЕСКОМ ПАКЕТЕ PSPP

Цель работы – научиться проводить проверку гипотезы о равенстве выборок на основе непараметрических критериев.

1. Основные теоретические сведения

Достаточно часто у исследователя нет информации о параметрах исследуемых совокупностей и виде их распределения: близки ли они к нормальному типу или какому-либо другому. Анализируемые выборки могут быть небольшого объема, дифференцированными на мелкие группы, или иметь переменные с 80% и более наблюдениями в одной из групп. Кроме того, могут быть выборки, у членов которых признак измеряется в рангах, а не в натуральных единицах. В таких случаях применение параметрических критериев может быть некорректным. Строго говоря, *нельзя применять критерии, основанные на предположении нормальности, к данным, не являющимся нормальными.*

Непараметрические критерии обладают рядом преимуществ:

- не требуют предварительных предположений относительно вида исходного распределения,
- для их вычисления не требуется большого объема данных,
- являются более применимыми в широком диапазоне условий, чем их параметрические аналоги.

Недостатки непараметрических критериев:

- низкая статистическая мощность (менее чувствительные);
- меньшая гибкость,
- большая вероятность совершить ошибку II рода - отклонить нулевую гипотезу, когда она верна.

Если данные не являются нормально распределенными, а измерения, в лучшем случае, содержат ранжированную информацию, то вычисление обычных описательных статистик (среднего, стандартного отклонения) не слишком информативно. В непараметрической статистике применяют разнообразный набор *мер положения* (например, медиана, мода, квартили и т.д.) и *рассеяния* (например, ранг, квартильный размах и т.д.), позволяющий представить более «полную картину» данных.

По существу, для каждого параметрического критерия имеется, по крайней мере, один, непараметрический аналог. Эти критерии можно отнести к одной из следующих групп:

- критерии различия между независимыми выборками;
- критерии различия между зависимыми выборками.

1.1 Критерий Вилкоксона для двух зависимых выборок

Знаковый ранговый критерий Вилкоксона — это непараметрическая версия t-критерия для парных выборок. Он используется для проверки наличия существенной разницы между двумя средними значениями генеральной совокупности, когда распределение различий между двумя выборками нельзя считать нормальным. Знаковый ранговый критерий Вилкоксона — это непараметрическая версия t-критерия для парных выборок. Он используется для проверки наличия существенной разницы между двумя медианными значениями генеральной совокупности, когда распределение различий между двумя выборками нельзя считать нормальным.

Знаковый ранговый критерий Вилкоксона является непараметрической версией парного t-критерия. Он используется для проверки того, существует ли значительная разница между двумя средними значениями совокупности.

Самый простой способ определить, являются ли различия нормально распределенными, — это создать гистограмму различий и посмотреть, насколько соответствуют они нормальному, «колоколообразному» распределению.

Нужно иметь в виду, что парный t-критерий довольно устойчив к отклонениям от нормальности, поэтому отклонение от нормального распределения должно быть довольно серьезным, чтобы оправдать использование критерия знакового ранга Вилкоксона.

Суть проверки равенства для двух зависимых выборок на основе критерия Вилкоксона заключается в следующем:

1. Формулируются нулевая и альтернативная гипотезы, а также устанавливается критический уровень значимости α :

H_0 — отсутствует статистически значимое различие двух выборок.

H_1 — выборки статистически различимы.

2. Отбираются необходимые данные для двух зависимых выборок.

3. Две выборки объединяются в одну общую выборку.

4. Ранжируются все значения общей выборки по возрастанию.

5. Суммируются ранги отдельно для каждой исходной выборки.

6. Сравниваются суммы рангов с помощью статистического критерия.

7. Интерпретируется значение расчетного уровня значимости p и z -статистики. Если величина p превышает значение 0,05, то следует принять нулевую гипотезу об отсутствии статистических различий двух выборок. В противном случае, если p меньше, чем 0,05 нулевая гипотеза отклоняется и принимается альтернативная гипотеза о существующих различиях двух выборок.

1.2 Пример применения критерия Вилкоксона для двух зависимых выборок

Задание 1. Исследование аккомодации в баллах для заик студенческого возраста.

Восемнадцать заик студенческого возраста читали вслух три различных фрагмента текста последовательно по пять раз. В первой ситуации они подвергались воздействию электрошока в момент заикания, во второй – сразу же после каждого слова, на котором они заикнулись. Последняя ситуация была контрольной и никакого воздействия не предусматривала. Во время чтения фиксировался процент заиканий. В таблице 3.1 приведены оценки аккомодации в баллах для всех ситуаций.

Исследователи хотят знать, влияют ли методы лечения электрошоком на оценку аккомодации

Таблица 3.1- Оценки аккомодации в баллах для заик студенческого возраста

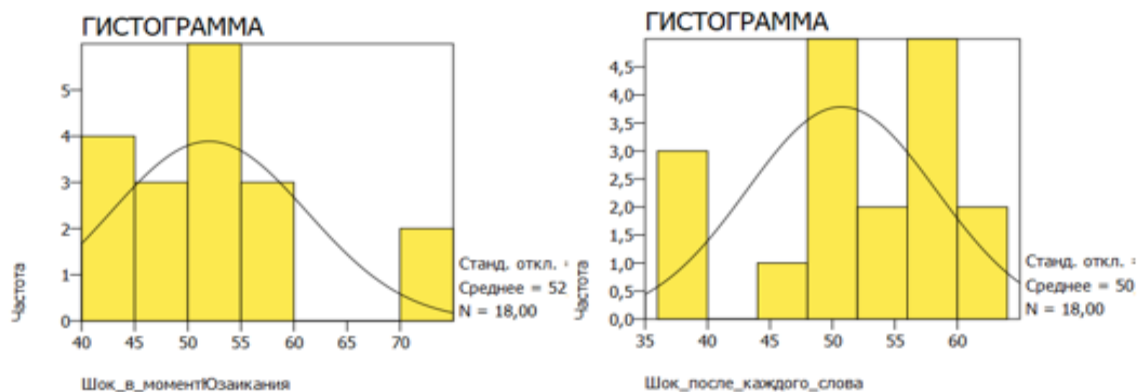
Испытуемые	Без шока	Шок после каждого слова	Шок в момент заикания
1	57	38	51
2	59	48	56
3	44	50	44
4	51	53	44
5	43	53	50
6	49	56	54
7	48	37	50
8	56	58	40
9	44	44	50
10	50	50	50
11	44	58	56
12	50	48	46
13	70	60	74
14	42	58	57
15	58	60	74
16	54	38	48
17	38	48	48
18	48	56	44

Сформулируем статистические гипотезы:

– нулевая гипотеза H_0 – разница средних в двух зависимых выборках равна нулю, то есть отсутствуют различия в лечении электрошоком заикания.

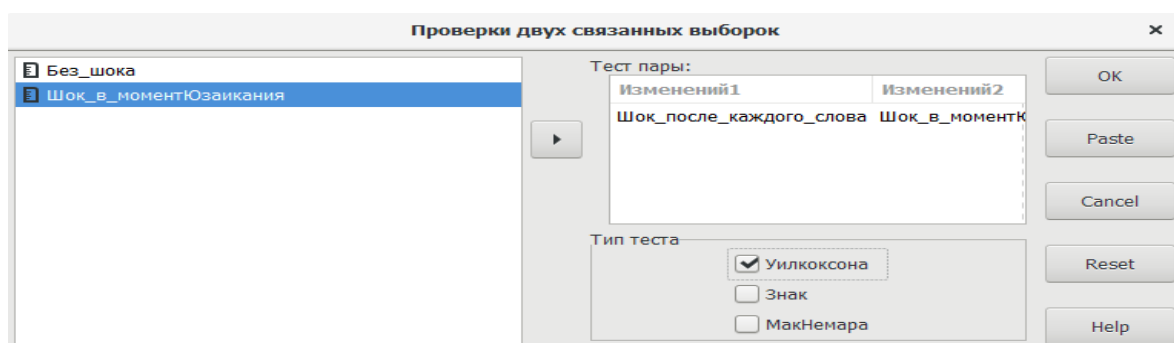
– альтернативная гипотеза H_1 – разница средних в двух зависимых выборках не равна нулю, то есть оценки аккомодации в баллах зависит от выбора метода лечения электрошоком. **Алгоритм решения:**

1. Загрузить файл *Вилкоксон_зависимые.sav*.
2. Выполнить визуальное сравнение гистограммы переменных *Шок после каждого слова* и *Шок в момент заикания* с нормальной кривой. Для каждой переменной выполнить команду **Графика (Garphs). Гистограмма (Histogram)** и установить флажок **Нормальная кривая**. В результате имеем:



Полученные гистограммы отличаются по внешнему виду от гистограмм для выборок с нормальным распределением. Выбираем непараметрический тест Вилкоксона.

3. Выполнить команду **Анализ. Непараметрическая статистика 2 связанные выборки... (2 Related Sample...)**. В диалоговом окне **Проверки двух связанных выборок** выбрать переменные для анализа – *Шок после каждого слова* и *Шок в момент заикания*, в группе **Тип теста** установим флажок в опции **Уилкоксона**. **Ок.** и нажмём кнопку **ОК**.



В результате имеем таблицы:

Ранги				
		N	Средний ранг	Сумма рангов
Шок_в_момент_заикания - Шок_после_каждого_слова	Отрицательные ранги	9	6,39	57,50
	Положительные ранги	7	11,21	78,50
	Совпадения	2		
	Итого	18		
Критерии проверки				
		Шок_в_момент_заикания - Шок_после_каждого_слова		
Z		-,54		
Асимпт. знач. (двусторонние)		,587		

Вторая таблица отображает статистику теста и соответствующее двустороннее р-значение:

Z-статистика теста: -0,54 – значение не велико.

Двустороннее р-значение: 0,587

Поскольку р-значение больше 0,05, нулевая гипотеза H_0 о равенстве выборок

принимается. Есть достаточно доказательств, чтобы заключить, что метод лечения не оказал статистически значимого влияния на аккомодацию заик.

Пример вывода:

Знаковый ранговый критерий Уилкоксона был проведен для определения наличия статистически значимой разницы в уровне аккомодации для заик при различных методах лечения. Всего в исследовании было задействовано 18 заик.

Тест показал, что не существует статистически значимой разницы в методах лечения между двумя группами ($z = -0,54$, $p = 0,587$).

1.3 Критерий Фридмана для нескольких зависимых выборок

Критерий Фридмана является непараметрическим аналогом двухфакторного дисперсионного анализа и используется для проверки нулевой гипотезы H_0 о том, что m выборок (повторных измерений) были отобраны из одной и той же генеральной совокупности или генеральных совокупностей с одной средней. Поскольку m выборок являются взаимно сопоставляемыми, количество наблюдений n является одинаковым во всех выборках.

1.4 Пример применения критерия Фридмана для нескольких зависимых выборок.

Задание 2. Пассивное курение и переносимость физической нагрузки.

В эксперименте участвовали 10 респондентов, больных ишемической болезнью. У каждого респондента переносимость физической нагрузки определяли как время до наступления приступа стенокардии при работе на велотренажере.

Затем в течение двух часов каждый отдыхал в отдельной комнате, где находились окуриватели. Они либо не курили, либо выкуривали по пять сигарет. В последнем случае комнату либо проветривали, либо не проветривали. После этого физическую нагрузку повторяли. Исследование продолжалось три дня, и каждый больной в случайном порядке испытал все три вида отдыха по одному в день.

Необходимо выяснить, влияет ли пассивное курение на переносимость физической нагрузки. Исходные данные приведены в таблице 3.2 и в файле *ранг_ДА.sav*.

Выдвигаем статистические гипотезы:

- нулевая гипотеза H_0 – между наблюдениями существуют лишь случайные различия по уровню переносимости физической нагрузки;
- альтернативная гипотеза H_1 – между наблюдениями существуют неслучайные различия по уровню исследуемого признака и, таким образом, пассивное курение влияет на переносимость физической нагрузки.

Таблица 3.2 – Данные для исследования

Больной	Свежий воздух		Пассивное курение, проветривание		Пассивное курение без проветривания	
	до	после	до	после	до	после
1	193	217	191	149	202	127
2	206	214	203	169	189	130
3	188	197	181	145	192	128
4	375	412	400	306	387	230
5	204	199	211	170	196	132
6	287	310	304	243	312	198
7	221	215	213	158	232	135
8	216	223	207	155	209	124
9	185	208	186	144	200	129
10	231	224	227	172	218	125

Алгоритм решения:

1. Загрузить файл *rang_ДА.sav*
2. Проверить выборки на соответствие нормальному закону распределения, используя тест Колмогорова-Смирнова (Sample K-S).

H₀: выборки соответствуют нормальному закону распределения при уровне значимости 0,05.

H₁: выборки не соответствуют нормальному закону распределения.

В результате:

Тест Колмогорова-Смирнова для одной выборки

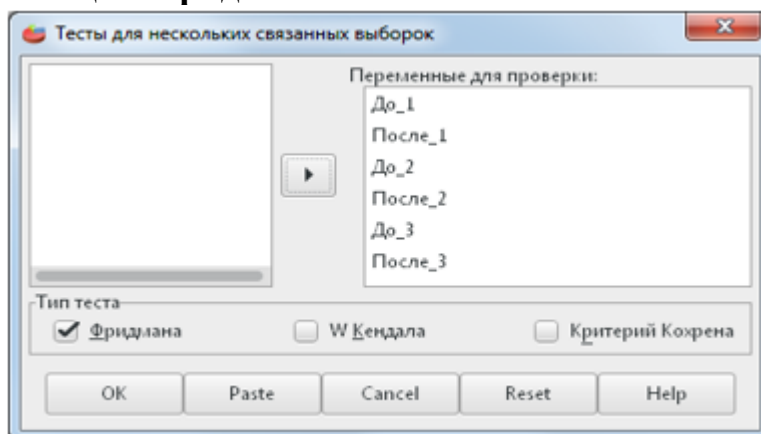
		До_1	После_1	До_2	После_2	До_3	После_3
N		10	10	10	10	10	10
Параметры нормального распределения	Среднее	230,60	241,90	232,30	181,10	233,70	145,80
	Станд. отклонение	58,65	67,80	68,39	52,40	64,80	36,87
Наибольшие разницы	Абсолютная	,30	,40	,33	,37	,31	,42
	Положительная	,30	,40	,33	,37	,31	,42
	Отрицательная	-,22	-,25	-,23	-,24	-,25	-,28
Колмогорова-Смирнова (Z)		,94	1,28	1,05	1,17	,98	1,31
Асимпт. знач. (двусторонние)		,340	,060	,214	,113	,290	,048

P-значение для одной выборки меньше уровня значимости 0,05, некоторые имеют незначительное превышения этого уровня. Нулевая гипотеза о соответствии всех выборок нормальному закону распределения отвергается, и принимается альтернативная гипотеза о несоответствии нормальному закону. Параметрический дисперсионный анализ провести невозможно, поэтому воспользуемся ранговым критерием Фридмана.

3. Выбрать команду **Анализ. Непараметрическая статистика. К Связанных выборок....**

Затем в открывшемся окне **Тесты для нескольких связанных выборок необходимо** выбрать все переменные для проверки, в группе **Тип теста** необходимо

установить флажок в опции **Фридмана**. ОК.



В результате получим:

Ранги		Критерии проверки	
	Средний ранг		
До_1	4,30	N	10
После_1	5,30	χ-квадрат	38,40
До_2	4,00	df	5
После_2	2,00	Асимпт. знач.	,000
До_3	4,40		
После_3	1,00		

Из этой таблицы следует, что критерий $\chi^2_{\text{расч}} = 38,40$ (*χ-квадрат*), а асимптотическая значимость $p=0$ для числа степеней свободы $df = 5$. Расчетное значение критерия $\chi^2_{\text{расч}} = 11,07$.

Поскольку расчётное значение критерия $\chi^2_{\text{расч}} = 38,40$ больше критического значения $\chi^2_{\text{табл}} = 11,07$ и расчетное значение уровня значимости $p < 0,05$, нулевая гипотеза об отсутствии статистических различий между выборками отклоняется, и принимается альтернативная гипотеза H_1 о том, что различия статистически значимы.

Общий вывод: пассивное курение снижает переносимость физической нагрузки.

1.5 Критерий Крускала-Уоллиса для нескольких независимых выборок

Тест Крускала-Уоллиса используется для определения наличия или отсутствия статистически значимой разницы между медианами трех или более независимых групп.

Этот тест является непараметрическим эквивалентом однофакторного дисперсионного анализа и обычно используется, когда предположение о нормальности не выполняется.

Критерий Крускала-Уоллиса не предполагает нормальности данных и гораздо менее чувствителен к выбросам, чем однофакторный дисперсионный анализ.

Критерий Крускала-Уоллиса применим при следующих допущениях:

- порядковая или непрерывная переменная отклика – переменная отклика должна быть порядковой или непрерывной переменной.
- независимость – наблюдения каждой группы должны быть независимы друг от друга.

- распределения имеют схожую форму – распределения в каждой группе должны иметь одинаковую форму.

1.6 Пример применения критерия Крускала-Уоллиса для нескольких независимых выборок.

Задание 3. Данные представляют собой результаты 60-ти учащихся трех спортивных школ одинакового направления. Определить, одинаков ли уровень подготовки спортсменов в школах. Используется файл спорт_школа_2:

Наблюдения	школа_1	школа_2	год_3
1	22,00	37,00	21,00
2	23,00	27,00	22,00
3	31,00	38,00	22,00
4	25,00	41,00	3,00
5	24,00	43,00	4,00
6	26,00	44,00	6,00
7	25,00	43,00	11,00
8	28,00	45,00	7,00
9	29,00	43,00	5,00
10	31,00	46,00	8,00
11	33,00	47,00	9,00

Для выбора метода исследования необходимо провести проверку соответствия выборок нормальному закону распределения на основе критерия Колмогорова-Смирнова. Тест показал следующий результат:

Тест Колмогорова-Смирнова для одной выборки

		школа_1	школа_2	школа_3
N		60	60	60
Параметры нормального распределения	Среднее	32,32	36,58	37,97
	Станд. отклонение	5,94	13,51	18,70
Наибольшие разницы	Абсолютная	,09	,29	,23
	Положительная	,09	,17	,17
	Отрицательная	-,06	-,29	-,23
Колмогорова-Смирнова (Z)		,69	2,28	1,78
Асимпт. знач. (двусторонние)		,720	,000	,002

Поскольку для 2-х выборок наблюдается несоответствие нормальному закону, метод дисперсионного анализа ANOVA не применим. В таком случае для исследования применяется ранговый непараметрический метод Крускала-Уоллиса.

Для применения теста Крускала-Уоллиса надо переформировать таблицу (Файл спорт_школа.sav).

Алгоритм проверки:

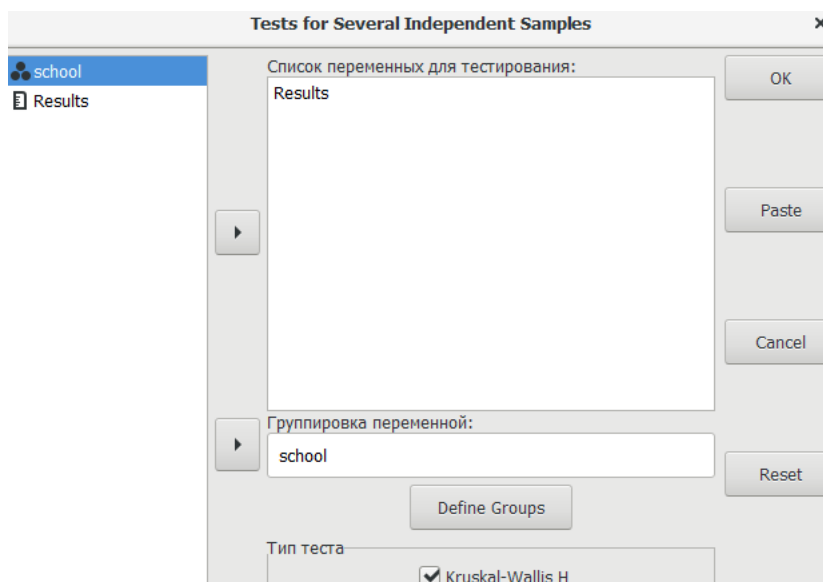
1. Сформулировать гипотезы:

Нулевая гипотеза – H_0 – выборки статистически неразличимы, то есть, все школы дают спортивную подготовку одного уровня.

Альтернативная гипотеза – H_1 – выборки статистически различаются, то есть, уровень спортивной подготовки в школах различен.

2. Выполнить команду Анализ. Непараметрическая статистика. К независимых выборок (K Independent Samples)...

3. В диалоговом окне указать зависимую переменную – *Results* и переменную группировки – *school*; В поле Тип теста установить флажок **Kruskal-Wallis**:



В результате имеем:

Ранги				Критерии проверки	
		N	Средний ранг		Results
Results	1	60	66,74	χ-квадрат	20,41
	2	60	96,23	df	2
	3	60	108,53	Асимпт. знач.	,000
Итого		180			

Расчетное значение критерия χ^2 равно **20,41**, что больше критического значения **5,99** при уровне значимости 0,05 и количестве степеней свободы 2, р-значение $0,000 < 0,05$.

Таким образом, отклоняем нулевую гипотезу H_0 и принимаем альтернативную H_1 о статистическом различии выборок.

В результате исследования установлено, что уровень спортивной подготовки в школах отличается.

2 Задания для выполнения практической работы.

Материалы для выполнения задания находятся по ссылке:

<https://do.sevsu.ru/mod/folder/view.php?id=558886>

1. Выполнить **задания 1-3** раздела 1 данных методических указаний. Студенты, присутствующие на лекции, в отчет эти задания не включают.

2. **Задание 4.** Исследователь хочет знать, оказывают ли три лекарства различное воздействие на боль в колене. Поэтому он набирает 30 человек, которые испытывают

одинаковую боль в коленях, и случайным образом делит их на три группы для приема препарата 1, препарата 2 или препарата 3. После месяца приема препарата исследователь просит каждого человека оценить боль в колене по шкале от 1 до 100, где 100 указывает на самую сильную боль. Использовать файл *к ПРЗ_3.xlsx*, лист *колени*.

3. **Задание 5.** Исследуется уровень ситуативной тревожности у пациентов до и после курса комплексной аудиовизуальной стимуляции. Определить эффективность воздействия курса комплексной аудиовизуальной стимуляции на пациента. Используется файл *к ПРЗ_3.xlsx*, лист *ур_тревожности*.

4. **Задание 6.** В файле *optimism_training.sav* содержатся результаты оценки оптимизма участников тренинга до (переменная *First*) и после (переменная *Second*) тренинга. Оценить эффективность тренинга.

5. **Задание 7.** Проверить нулевую гипотезу H_0 о том, что различные программы физической реабилитации приводят к практически одинаковому результату. Использовать файл *к ПРЗ_3.xlsx*, лист *реабилитация*.

6. **Задание 8.** Исследовать, одинаково ли относятся к инфляции вообще респонденты разных возрастных групп. Использовать файл *inflation.sav*.

7. **Задание 9.** Исследовались искажения мышления у родственников шизофреников. Переменная *group* содержит номер группы (1 — экспериментальная группа (родственники больных), 2 — контрольная группа (норма)), вторая переменная *distort* — количество ошибок. Использовать файл *Distort.sav*.

8. **Задание 10.** Рассматриваются группы респондентов по отношению к моде (*q1*). Определить есть ли разница в затратах на одежду за один поход в магазин (*q3*) в группах респондентов? Использовать файл *fashion.sav*.

3 Содержание отчета

1. Титульный лист.
2. Цель практической работы.
3. Ход выполнения работы:
 - формулировка задания должна быть непосредственно перед его выполнением;
 - порядок действий;
 - результат выполнения в виде листинга PSPP;
 - интерпретация результатов выполнения задания.
4. Общие выводы по практической работе.

Практическая работа № 4

ЛИНЕЙНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ В ПАКЕТЕ PSPP

Цель работы: получение практических навыков в проведении линейного регрессионного анализа с использованием пакета PSPP.

1 Краткие теоретические сведения

Регрессионный анализ применяется для выявления влияния одной или нескольких независимых переменных на одну зависимую переменную и позволяет дать математическое описание зависимости между переменными, построить модель (тренд), отражающий данную зависимость, и позволяющий строить прогнозы изменения зависимой переменной.

В практике социологических исследований с помощью регрессионного анализа можно найти ответы на следующие вопросы:

1. В какой зависимости находятся сумма расходов на отдых в целом и сумма расходов на питание и развлечения?
2. Чему равен вклад расходов на персонал в изменение объема продаж при фиксированных ценах и расходах на продвижение?
3. Какие объемы продаж можно ожидать при заданных ценах, расходах на продвижение и расходах на персонал?
4. Имеется ли зависимость между успеваемостью студента и количеством пропущенных занятий? Построить вид этой зависимости.

В регрессионный анализ могут быть включены только количественные переменные.

Если необходимо использовать номинальные переменные, их необходимо представить как дихотомические (бинарные). Такие переменные могут принимать значения только **0** и **1**.

Ответ на вопрос «Какую марку одежды Вы предпочитаете?» может быть закодирован следующим образом: «1» - предпочитаю, «0» - не предпочитаю.

Ответ на вопрос «Какую марку одежды Вы предпочитаете?» может быть закодирован следующим образом: «1» - предпочитаю, «0» - не предпочитаю.

Уровень качества обслуживания на предприятии общественного питания может измеряться рейтингом от 1 до 10.

1.2 Классификация методов регрессионного анализа

Линейный регрессионный анализ – переменные должны быть количественными (шкала - Масштаб):

$\hat{y} = a_0 + a_1x$ - простая линейная регрессия,

$\hat{y} = a_0 + a_1x_1 + a_2x_2 + \dots + a_nx_n$ - множественная линейная регрессия.

Бинарная логистическая регрессия - выявляет зависимость бинарной (дихотомической) переменной от одной или нескольких количественных переменных.

Мультиномиальная логистическая регрессия - Если зависимая переменная является категориальной и имеет более двух категорий.

Нелинейная регрессия – анализирует нелинейные связи между количественными переменными:

$$\hat{y} = a_0 + a_1x^2, \hat{y} = a_0 + a_1\frac{1}{x}, \hat{y} = a_0 + a_1\ln x, \hat{y} = a_0 + a_1e^x \dots$$

Здесь: y — зависимая переменная; $x_1, x_2, \dots, x_n$ — независимые переменные;

a_0 — свободный член (константа); $a_1, a_2, \dots, a_n$ — коэффициенты при независимой переменной или коэффициенты регрессии.

В результате регрессионного анализа необходимо определить свободный член (a_0) и коэффициенты регрессии (a_i).

1.3 Показатели регрессионного анализа

Обязательным условием является расчет показателей, характеризующих практическую применимость и статистическую значимость результатов анализа и построенной регрессионной модели.

К данным показателям относятся:

Коэффициент корреляции (R) — отражает связь между значениями зависимой и независимой переменных. Значения коэффициента корреляции могут изменяться в интервале от нуля до единицы. При приближении коэффициента к единице, взаимосвязь усиливается. Если R превышает 0,5, то говорят о наличии связи между переменными, и наоборот.

Коэффициент R -квадрат — коэффициент детерминации равен той доле результатов наблюдений, которая объясняется уравнением регрессии. Значения R -квадрат изменяются в интервале от нуля до единицы. Если R -квадрат равен 0,5, то только для 50% всех наблюдений справедлива построенная регрессионная модель. Для значимых результатов исследования значение коэффициента R -квадрат должно превышать 0,5.

Показатель значимости p - характеризует значимость регрессионной модели, построенной на основе данных опроса респондентов, попавших в выборку, для всей генеральной совокупности. Если $p < 0,05$, то можно говорить о справедливости регрессионной модели для всей генеральной совокупности. Результат $p > 0,05$ говорит о том, что связь между переменными слабая или не обнаружена.

1.4 Простая линейная регрессия

Простая линейная регрессия используется для оценки связи между одной независимой переменной и одной переменной отклика (зависимой переменной), линейно меняющейся в некотором диапазоне значений.

Регрессионный анализ – процедура построения уравнения регрессии или регрессионной модели методом наименьших квадратов (МНК) и анализ полученного уравнения с помощью аппарата математической статистики.

Уравнением регрессии называют функциональную зависимость математического ожидания случайной величины Y (зависимой переменной) от одной x или нескольких $X=(x_1, x_2, \dots, x_n)$ детерминированных величин (независимых переменных), то есть $M\{Y\} = \hat{Y} = \varphi(X)$.

Считаем, что функция φ линейная и однопараметрическая, тогда уравнение регрессии имеет вид:

$$y = \alpha + \beta x + \varepsilon, \quad (4.1)$$

где α и β – неизвестные параметры модели, подлежащие оценке; ε – случайная переменная, характеризующая отклонение от теоретически предполагаемой регрессии (ошибка). Относительно ε делаются следующие предположения:

- ε – является нормально распределенной случайной переменной;
- математическое ожидание ε равна нулю;
- дисперсия ε постоянна;
- последовательные значения ε не зависят друг от друга.

Предполагается, что каждое наблюдение y_i представимо в виде

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

Пусть в результате эксперимента получено n пар наблюдений (параметров) y_i и x_i , $i=1, n$. Задача регрессионного анализа состоит в том, чтобы, располагая выборочными значениями $x_1, x_2, \dots, x_n$ и $y_1, y_2, \dots, y_n$, определить значения параметров α и β . Поскольку объем выборок ограничен, то полученные значения параметров α и β являются их оценками, которые будем обозначать a_0 и a_1 соответственно. При этом уравнение регрессии (4.1), записанное в виде

$$\hat{y} = a_0 + a_1 x$$

есть оценка модели (1), где $\hat{y}$ обозначает предсказанное значение y для данного x , когда a_0 и a_1 определены.

Согласно МНК, параметры уравнения регрессии подбираются так, чтобы сумма квадратов отклонений от регрессии была минимальной:

$$\sum_{i=1}^n [y_i - \hat{y}_i]^2 = \sum_{i=1}^n [y_i - (a_0 + a_1 x_i)]^2 \rightarrow \min. \quad (4.2)$$

Дифференцируя выражение (4.2) по a_0 и a_1 и приравнявая к нулю производные,

получаем два уравнения для определения a_0 и a_1 :

$$\begin{cases} \sum_{i=1}^n [y_i - (a_0 + a_1 x_i)] x_i = 0 \\ \sum_{i=1}^n [y_i - (a_0 + a_1 x_i)] = 0 \end{cases} \quad (4.3)$$

Решая систему (4.3) относительно a_0 и a_1 , получаем следующие выражения:

$$a_0 = \frac{\sum_{i=1}^n x_i^2 \cdot \sum_{i=1}^n y_i - \sum_{i=1}^n y_i x_i \cdot \sum_{i=1}^n x_i}{n \cdot \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}$$

$$a_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n y_i \cdot \sum_{i=1}^n x_i}{n \cdot \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}.$$

Полученные оценки параметров необходимо проверить на значимость, то есть статистически обосновать вывод о том, что параметры регрессии отличны от нуля. Проверка осуществляется на основе вычисления критерия Стьюдента (t -критерия).

Вычисляется статистика $t = \frac{|a_1|}{s_{a1}}$, $s_{a1} = \frac{s_{ocm}}{s_x \sqrt{n-2}}$ – стандартная ошибка. Здесь s_{ocm}^2 –

остаточная дисперсия, равная $s_{ocm}^2 = \frac{1}{n} \cdot \sum_{i=1}^n (\hat{y}_i - y_i)^2$, s_x^2 – дисперсия переменной x .

Значимость свободного коэффициента не определяется. Статистика t имеет распределение Стьюдента с $\gamma=n-2$ степенями свободы, где n – количество наблюдений эксперимента. По уровню значимости α и числу степеней свободы γ определяется критическое значение $t_{\alpha,\gamma}$. Если $t \geq t_{\alpha,\gamma}$, то параметр a_1 значимо отличается от нуля.

Для определения критического значения t -статистики можно воспользоваться функцией ЭТ Excel: СТЬЮДЕНТ.ОБР.2Х(α , γ).

Проверка адекватности уравнения линейной регрессии (3) экспериментальным данным выполняется с использованием критерия Фишера (F-

критерия): $F = \frac{s_{ocm}^2}{s_y^2}$, s_y^2 – дисперсия наблюдаемых значений y , s_{ocm}^2 – остаточная

дисперсия.

По уровню значимости α и числам степеней свободы $\gamma_1=k-1$ и $\gamma_2=n-k$, n – количество наблюдений эксперимента, k – количество коэффициентов в модели.

определяется критическое значение $F(\alpha, \gamma_1, \gamma_2)$. Для определения критического значения F-критерия можно воспользоваться функцией ЭТ Excel: F.ОБР($1 - \alpha, \gamma_1, \gamma_2$).

Если $F \leq F(\alpha, \gamma_1, \gamma_2)$, то уравнение регрессии считается адекватным, и его можно использовать для анализа и прогнозирования значений переменной Y .

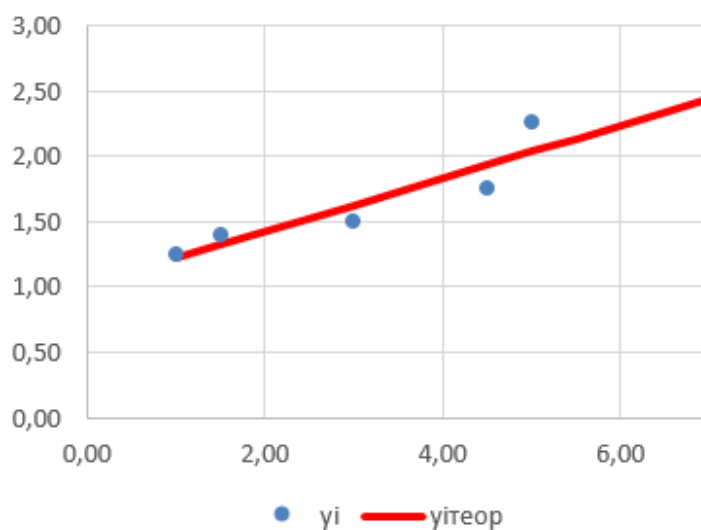
Исходными данными для простой линейной регрессии является таблица значений независимой переменной x и наблюдаемой (зависимой) переменной y :

МНК		
	1	2
	x_i	y_i
1	1,00	1,25
2	1,50	1,40
3	3,00	1,50
4	4,50	1,75
5	5,00	2,25
Σ	15,00	8,15

Применяя метод наименьших квадратов, можно построить регрессионную модель

$$y = 1,024 + 0,202x$$

и использовать ее для прогнозирования значений зависимой переменной:



Красная линия на графике называется линией тренда.

1.5 Простая линейная регрессия в PSPP.

Рассмотрим этапы выполнения регрессионного анализа на примере.

Задание 1. Используя результаты оценки потребителями 15 магазинов с точки зрения уровня обслуживания и приемлемости цен, проведите регрессионный анализ и выявите зависимость в предпочтениях от приемлемости цен и уровня обслуживания. При оценке магазинов использовалась 10-балльная шкала. 10 - соответствует максимальной оценке, 1 - минимальной (табл. 4.1). Использовать файл *предпочтение_mag.sav*

Таблица 4.1 - Предпочтения магазинов

№	Предпочтение	Качество обслуживания	Цена
1	6	5	3
2	9	6	10
3	8	6	4
4	3	2	1
5	10	6	10
6	4	3	1
7	5	4	7
8	2	1	4

№	Предпочтение	Качество обслуживания	Цена
9	10	9	8
10	9	5	10
11	10	8	8
12	2	1	5
13	9	8	5
14	5	3	2
15	3	8	3

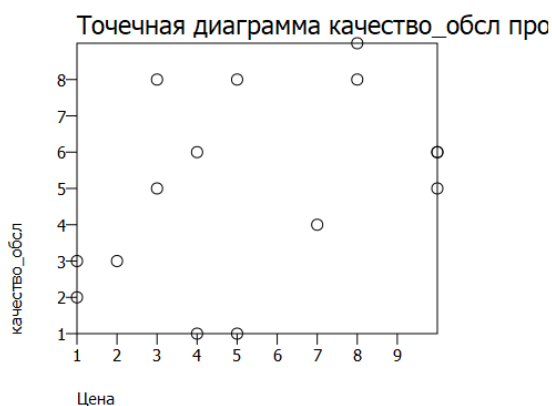
Порядок действий.

Шаг 1. Создать таблицу в PSPP. В качестве шкалы для всех переменных выбрать **Масштаб**. Данные скопировать из файла **РА_задания.xlsx**, Лист **Таблица 1**.

Шаг 2. Диаграмму рассеяния независимых признаков можно построить, используя меню

Graphs. Scatterplot.

В диалоговом окне ввести в поля Оси X и Y независимые переменные. В результате получим:



Шаг 3. Построить регрессионную модель (тренд), отражающую зависимость между переменной «Предпочтение» и переменной «Качество обслуживания», составить уравнение регрессии.

1. Анализ - Регрессия - Линейная

2. Переменная «**Предпочтение**» - в поле **Зависимая** переменная. Переменная «**Качество обслуживания**» - в поле **Независимая** переменная.

3. Кнопка **Статистика...** – установить флажки:

- ✓ **Коэффициенты**,
- ✓ **R**,
- ✓ **ANOVA**.

4. Кнопка **Сохранить...** - установить флажок **Прогнозные значения** для вывода

значений зависимой переменной, вычисленных по уравнению регрессии (столбец PRED1).

блюдения	Предпочтение	качество_обсл	Цена	PRED1
1	6	5	3	6,33
2	9	6	10	7,22
3	8	6	4	7,22
4	3	2	1	3,68
5	10	6	10	7,22
6	4	3	1	4,56
7	5	4	7	5,45

Шаг 4. Интерпретация результатов.

Резюме модели (Предпочтение)

R	R квадрат	Скорректированный R квадрат	Ст. погрешность оценки
,75	,56	,53	2,11

1. Значение коэффициента детерминации $R = 0,75$ ($>0,5$) свидетельствует о тесной взаимосвязи между предпочтением магазина и качеством обслуживания.

2. Значение коэффициента детерминации (R квадрат) равно 0,56 означает, что построенная модель верна только для 56% случаев, в результате которых рост качества обслуживания влечет за собой рост предпочтения данного магазина.

ANOVA (Предпочтение)

	Сумма квадратов	df	Среднее по квадратам	F	Знач.
Регрессия	75,26	1	75,26	16,85	,001
Остатки	58,07	13	4,47		
Итого	133,33	14			

Таблица «ANOVA»: показатель Значимость равен $0,001 \ll 0,05$.

Можно утверждать, что построенная регрессионная модель верна для всех покупателей, посещающих выбранные магазины, то есть результат можно распространить на генеральную совокупность.

Коэффициенты (Предпочтение)

	Нестандартизованы коэффициенты		Стандартизованные коэффициенты	t	Знач.
	B	Станд. погрешность	Бетта		
(Константа)	1,91	1,21	,00	1,58	,137
качество obsl	,89	,22	,75	4,10	,001

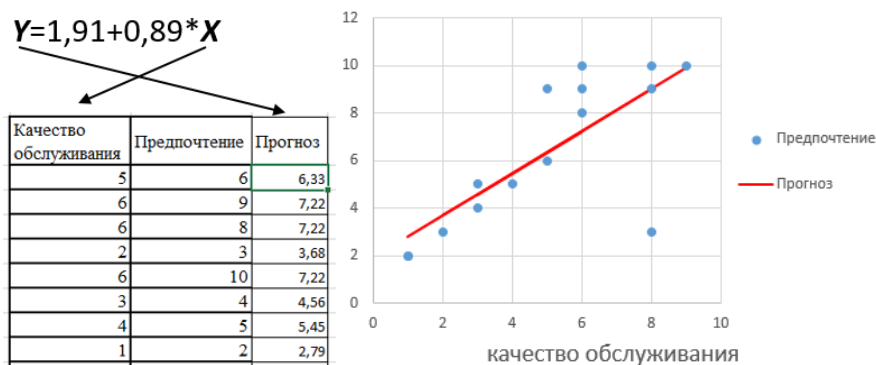
В: коэффициенты регрессии, уравнение регрессии примет вид:

$$Y=1,91+0,89*X$$

При этом, с учетом стандартной ошибки, равной 1,21 при доверительном интервале 95% уровень предпочтения магазина без учета качества обслуживания составит $1,91 \pm 1,21$.

Значение коэффициента регрессии в построенной модели составляет 0,89. Это значит, что рост качества обслуживания почти на 1 п. влечет за собой рост предпочтения почти на 2 пункта.

Шаг 5. Графическая интерпретация результатов. Построение графического отображения наблюдаемых данных и теоретического (прогнозируемого) значения зависимой переменной производится в ЭТ Excel. Для этого в ЭТ необходимо скопировать из PSPP значения столбца PRED1. Независимая переменная Качество обслуживания должна располагаться в самом левом столбце. Ее значения откладываются на оси абсцисс. Все остальные ряды находятся правее:



Выполнить команды меню: **Вставка. Диаграммы. Тип «Точечная»**. В результате будет построено в области диаграммы два ряда: наблюдаемое значение в столбце **Предпочтение** (на графике голубые точки) и линия тренда или теоретическое значение по столбцу **Прогноз** (на графике красная линия).

Задание 2. Построить регрессионную модель (тренд), отражающую зависимость между переменной «Предпочтение» и переменной «Цена», составить уравнение регрессии.

Выполнив команды Анализ. Регрессия. Линейная..., в результате получим:

Резюме модели (Предпочтение)			
R	R квадрат	Скорректированный R квадрат	Ст. погрешность оценки
,72	,53	,49	2,21

ANOVA (Предпочтение)					
	Сумма квадратов	df	Среднее по квадратам	F	Знач.
Регрессия	70,06	1	70,06	14,40	,002
Остатки	63,27	13	4,87		
Итого	133,33	14			

Коэффициенты (Предпочтение)					
	Нестандартизованы коэффициенты		Стандартизованные коэффициенты		Знач.
	B	Станд. погрешность	Бетта	t	
(Константа)	2,59	1,14	,00	2,27	,040
Цена	,69	,18	,72	3,79	,002

1. $R = 0,75 (>0,5)$ свидетельствует о тесной взаимосвязи между предпочтением магазина и уровнем цен.

2. R квадрат = 0,53 означает, что построенная модель верна только для 53% случаев, в результате которых рост рейтинга цены влечет за собой рост предпочтения данного магазина.

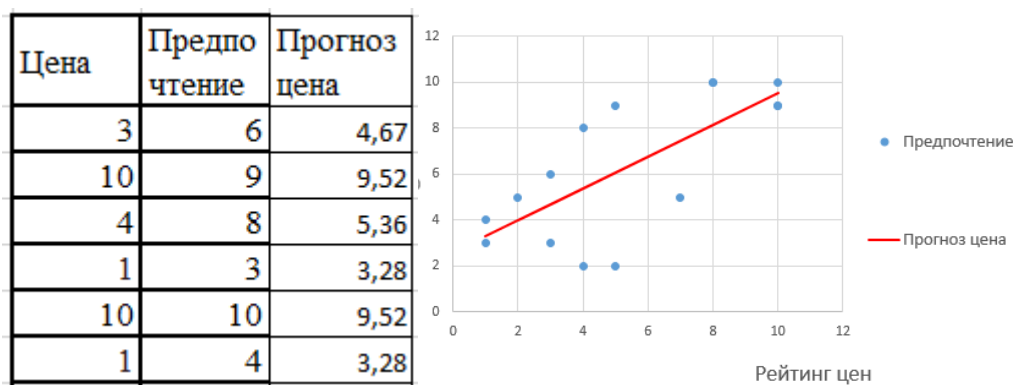
3. «ANOVA»: показатель Значимость равен $0,002 < 0,05$ выводы можно распространить на всю генеральную совокупность.

4. Уравнение регрессии имеет вид: $Y = 2,59 + 0,69 * X$

5. Стандартная ошибка равна 1,14 при доверительном интервале 95% уровень предпочтения магазина без учета уровня цены составит $2,59 \pm 1,14$.

6. Значение коэффициента регрессии составляет 0,89. Это значит, что рост рейтинга цены на 0,69 п. влечет за собой рост предпочтения магазина на 2,6 пункта.

7. Графическая интерпретация результатов:



1. 6 Множественная линейная регрессия

Множественная регрессия исследует влияние двух и более факторов на зависимую переменную.

Условия получения приемлемых результатов анализа:

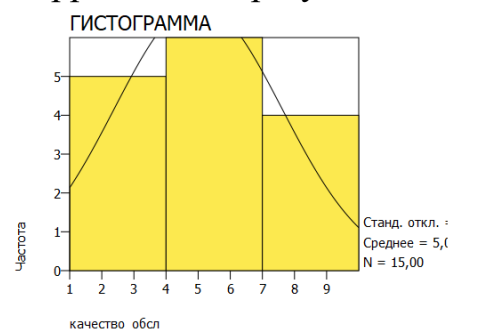
- Для того, чтобы существующие зависимости между переменными были признаны значимыми, необходимо иметь достаточно большие размеры выборки (N не менее 50).
- Данные необходимо проверить на корректность и наличие ошибок.
- Необходимо проверить распределение на нормальность (показатели асимметрии и эксцесса не превосходят 1).
- Запрет на использование зависимых переменных, корреляции между которыми близки к 1 (-1).

Задание 3. Построить регрессионную модель (тренд), отражающую зависимость между переменной «Предпочтение» и независимыми переменными «Цена» и «Качество обслуживания», составить уравнение регрессии.

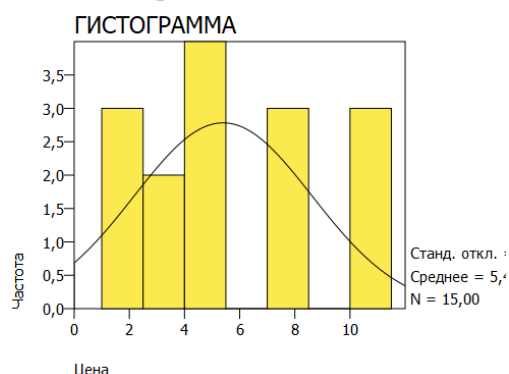
1. Проверка зависимых переменных на соответствие нормальному закону распределения.

Проверку можно проводить на основании теста Колмогорова-Смирнова, по значениям характеристик распределения или построением гистограммы с наложенной на нее нормальной кривой. Независимость переменных можно проверить по коэффициенту корреляции, выполнив команду Анализ. Корреляция. В результате:

Описательные статистики			Статистика
качество_обсл	Среднее		5,00
	Медиана		5,00
	Асимметрия		-,11
	Эксцесс		-1,13
Цена	Среднее		5,40
	Медиана		5,00
	Асимметрия		,22
	Эксцесс		-1,33



Корреляции			
		качество_обсл	Цена
качество_обсл	Корреляция Пирсона	1,000	,440
	Знач. (двустороннее)		,101
	N	15	15
Цена	Корреляция Пирсона	,440	1,000
	Знач. (двустороннее)	,101	
	N	15	15



Распределения переменных *Цена* и *Уровень обслуживания* можно считать нормальными. Зависимость между переменными слабая (0,44).

Для построения регрессионной модели, выполняя команды **Анализ. Регрессия. Линейная...**, в диалоговом окне в поле Независимая переменная нужно внести две переменных – **Цена** и **Качество обслуживания**. В результате:

Резюме модели (Предпочтение)			
R	R квадрат	Скорректированный R квадрат	Ст. погрешность оценки
,87	,76	,72	1,64

ANOVA (Предпочтение)					
	Сумма квадратов	df	Среднее по квадратам	F	Знач.
Регрессия	100,98	2	50,49	18,73	,000
Остатки	32,35	12	2,70		
Итого	133,33	14			

Коэффициенты (Предпочтение)					
	Нестандартизованы коэффициенты		Стандартизованные коэффициенты		Знач.
	B	Станд. погрешность	Бетта	t	
(Константа)	,65	1,02	,00	,63	,539
качество_обсл	,63	,19	,54	3,39	,005
Цена	,47	,15	,49	3,09	,009

1. $R = 0,87$ ($>0,5$) свидетельствует о тесной взаимосвязи между зависимой и независимыми переменными.

2. R квадрат = 0,76 означает, что построенная модель верна в 76% случаев, в результате которых рост зависимых переменных влечет за собой рост предпочтения данного магазина.

3. «ANOVA»: Значимость равна $0,000 \ll 0,05$, результаты можно распространить на всю генеральную совокупность.

4. уравнение регрессии имеет вид: $Y=0,65+0,63 \cdot X_1+0,47 \cdot X_2$.

Качество обслуживания существенно влияет на предпочтения покупателей, чем уровень цен.

Отличие от нуля константы говорит о том, что есть еще какие-то факторы, влияющие на предпочтения покупателей.

2 задания для выполнения практической работы

Материалы для выполнения задания находятся по ссылке:

<https://do.sevsu.ru/mod/folder/view.php?id=559007>

1. Выполнить задания 1-3 пункта 1 данных методических указаний.

2. **Задание 2.** Фирма провела рекламную кампанию. Через 10 недель руководство фирмы решило проанализировать эффективность рекламной компании, сопоставив недельные объемы продаж в тыс. руб. с расходами на рекламу в тыс. руб. Использовать файл *реклама_продажи.sav*.

3. **Задание 3.** Для оценки удовлетворенности отдыхающих базы отдыха было проведено исследование и опрошено 50 человек. В качестве факторов, влияющих на удовлетворенность отдыхом, выделили месторасположение базы отдыха (1), качество обслуживания персоналом (2) и уровень затрат на отдых (3). Для оценки удовлетворенности отдыхающих и каждого фактора была предложена 5-балльная шкала. Чем выше балл, тем выше оценка. Проведите регрессионный анализ и выявите зависимость в удовлетворенности от вышеперечисленных факторов в соответствии с индивидуальным вариантом. Использовать файл *база_отдыха.sav*.

№ вар-та	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
простая	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2
множеств	1,2	1,3	2,3	1,2	1,3	2,3	1,2	1,3	2,3	1,2	1,3	2,3	1,2	1,3	2,3	1,2	1,3	2,3	1,3	2,3
множеств	1,2,3																			

4. **Задание 4.** На основе регрессионного анализа определить зависимость количества жира в организме – у от переменных x_1 - толщины кожной складки трицепса; x_2 - окружности бедра и x_3 - окружности предплечья. Использовать файл *bodyfat.sav*.

Для каждого задания придерживаться следующего порядка действий:

1. Проверьте условия получения приемлемых результатов анализа.
2. Проведите простой/множественный линейный регрессионный анализ.
3. Постройте регрессионную модель, отражающую зависимость между переменными, составьте уравнение регрессии.
4. Сделайте вывод о том, как будет изменяться зависимая переменная при изменении независимых.
5. Сделайте вывод о вероятности ошибки, при использовании данной регрессионной модели.

6. Выполните графическую интерпретацию результатов простой линейной регрессии в случае простой регрессии.

3 Содержание отчета

1. Цель работы.
 2. Постановка задачи
 3. Ход выполнения работы.
 4. Анализ результатов.
 5. Общие выводы по работе.
- } для каждого задания

Практическая работа № 5 **БИНАРНАЯ ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ В СТАТИСТИЧЕСКОМ ПАКЕТЕ PSPP**

Цель работы: получение практических навыков в проведении бинарного логистического регрессионного анализа (БЛРА) с использованием пакета PSPP.

1 Краткие теоретические сведения

С помощью метода бинарной логистической регрессии можно исследовать зависимость дихотомических переменных от независимых переменных, имеющих любой вид шкалы.

Как правило, в случае с дихотомическими (бинарными) переменными речь идет о некотором событии, которое может произойти или не произойти; бинарная логистическая регрессия в таком случае рассчитывает вероятность наступления события в зависимости от значений независимых переменных.

Вероятность наступления события для некоторого случая рассчитывается по формуле:

$$P = \frac{1}{1 + e^{-z}}$$

где $z = b_1x_1 + b_2x_2 + \dots + b_nx_n + a$, x_i – значения независимых переменных, b_i – коэффициенты, расчет которых является задачей бинарной логистической регрессии, a – некоторая константа.

Если для P получится значение меньше 0.5, то можно предположить, что событие не наступит; в противном случае предполагается наступление события.

Практическую применимость и статистическую значимость результатов БЛРА определяют следующие показатели построенной регрессионной модели:

Р-квадрат Нагелькерке, показывает процентную долю вариации (дисперсии) переменной отклика, которая может быть объяснена независимыми переменными.

Общий процент показывает процент наблюдений, которые модель смогла правильно классифицировать.

Кроме этих показателей используется статистика Вальда для каждой независимой переменной, которая определяет статистическую значимость коэффициента регрессии и соответствующее ей p -значение. В идеале эта величина должна быть меньше 0,05, но это может быть лишь при 100%-й классификации.

Коэффициенты регрессионной модели представлены в столбце «В» таблицы Переменные в уравнении. Если коэффициент положительный, то это увеличивает шансы наблюдаемой переменной (столбец $\text{Exp}(B)$) при увеличении соответствующей независимой переменной, если отрицательный, то при увеличении независимой переменной шансы уменьшаются.

2 Пример выполнения задания

Рассмотрим порядок построения уравнения БЛРА и применение его для определения вероятности события.

Задание 1. Оценить выход баскетболистов из университетской команды в НБА на основании их среднего количества очков за игру и уровня дивизиона команды. Использовать файл *НБА.sav*:

Наблюдения	выход_НБА	очки	дивизион
1	0	5	2
2	0	8	2
3	1	28	2
4	1	12	1
5	0	14	2
6	1	36	1
7	0	9	1
8	0	9	2
9	1	21	1
10	0	21	1
11	1	24	1
12	1	24	2
13	1	34	2
14	1	25	1

Порядок действий.

Шаг 1. Выполнить команду: **Анализ. Регрессия. Бинарная логистическая...**

Шаг 2. Указать **зависимую** переменную **выход_НБА**.

Шаг 3. Указать **независимые** переменные - **очки** и **дивизион**.

Шаг 4. Кнопка **Параметры** – указать доверительный интервал.

Шаг 5. ОК.

В результате имеем:

Кодирование зависимой переменной

Начальное значение	Внутреннее значение
0	,00
1	1,00

Резюме обработки наблюдений

Невзвешенные наблюдения	N	Процент
Включено в анализ	14	100,0%
Пропущенные наблюдения	0	,0%
Итого	14	100,0%

Суммарная модель

Step	-2 log-правдоподобия	R квадрат Кокса и Снела	R квадрат Нагелкере
1	8,37	,54	,72

R-квадрат Нагелькерке, показывает **72%**. Переменные **очки** и **дивизион** могут объяснить изменчивость отклика на 72%. Остальная вариация независимой переменной объясняется влиянием какими-либо другими независимыми переменными.

Таблица классификации

		Предусмотрено		Percentage Correct
		выход_НБА		
Просмотрено		0	1	
Шаг 1	выход_НБА 0	5	1	83,3%
	1	1	7	87,5%
Общий процент				85,7%

Модель может правильно классифицировать 85,7% (Общий процент) всех наблюдений.

Переменные в уравнении

	В	Стандартная ошибка	Вальда	df	Знач.	Exp(B)	95% коэф. спряж. для Exp(B)	
							Нижняя	Верхняя
Шаг 1 очки	,28	,14	4,43	1	,035	1,33	1,02	1,73
дивизион	-1,17	1,85	,40	1	,525	,31	,01	11,57
Константа	-3,11	3,44	,82	1	,365	,04		

1. Значимость коэффициентов:

- для переменной **Очки** статистика Вальда составляет 4,43, а Знач.(p=значение) равно 0,035, что меньше установленного уровня значимости 0,05; эта переменная оказывает достаточно сильное на переменную **Выход_НБА**.

- для переменной **Дивизион** статистика Вальда составляет всего 0,40, а Знач.(p=значение) равна 0,525, что больше установленного уровня значимости 0,05. Этот коэффициент регрессии статистически незначим, влияние его на зависимую переменную неопределенное. Здесь можно обратить внимание на 95% -й интервал для шансов.

2. Отношение шансов **Exp(B)** для каждой переменной-предиктора указывает на изменение шансов игроков быть выбранным в НБА, связанном с увеличением на одну единицу данной предикторной переменной:

Если среднее число очков увеличится на 1, то шанс попасть в НБА увеличится в 1,33 раза. Переход из первого дивизиона во второй (+1) увеличивает шанс попасть в НБА в 0,31 раз.

3. По вычисленным коэффициентам регрессии (В) строится регрессионная модель:

$$Z=0,28*x_1-1,17*x_2-3,11,$$

где $b_1=0,28$ – среднее число набранных очков; $b_2=-1,17$ – дивизион, $a=-3,11$ – константа.

4. Теперь можно определить вероятность того, что игрок, имеющий определенное среднее число очков за игру и играющий в том или ином дивизионе будет выбран в

НБА, используя следующую формулу:

$$P = \frac{1}{1 + e^{-z}}, \text{ где } z = 0,28 \cdot x_1 - 1,17 \cdot x_2 - 3,11.$$

Расчеты удобно производить в ЭТ Excel. Примеры расчётов представлены ниже.

	A	B	C	D
29	a	b1	b2	
30	-3,11	0,28	-1,17	
31				
32	x1	x2	z	p
33	20		=B30*A33+C30*B33+A30	

Вычисление переменной z:

$$z = 0,28 \cdot x_1 - 1,17 \cdot x_2 - 3,11.$$

	A	B	C	D
29	a	b1	b2	
30	-3,11	0,28	-1,17	
31				
32	x1	x2	z	p
33	20	1	=1/(1+EXP(-C33))	

Вычисление вероятности P:

$$P = \frac{1}{1 + e^{-z}}$$

	A	B	C	D
29	a	b1	b2	
30	-3,11	0,28	-1,17	
31				
32	x1	x2	z	p
33	20	1	1,32	0,78918

Результаты вычислений

Выводы по результатам анализа могут быть следующими.

- Был проведен бинарный регрессионный анализ, цель которого определить, как количество набранных очков и уровень дивизиона влияют на вероятность того, что баскетболист будет выбран в НБА. Всего в анализе участвовало 14 игроков.
- В результате модель объяснила 72,5% вариаций признака и правильно классифицировала 85,7% случаев.
- Если среднее число очков увеличится на 1, то шанс попасть в НБА увеличится в 1,33 раза.
- Переход из 2 дивизиона в первый (1) увеличивает шанс попасть в НБА в 0,31 раз.
- Построенная модель может быть использована для определения вероятности выхода игроков в НБА и их классификации в соответствии с вычисленной вероятностью. Например, если игрок играет в 1-м дивизионе и имеет 20 очков, то вероятность выбора его в НБА равна 0,79%.

4 Задание для выполнения практической работы

Материалы для выполнения задания находятся по ссылке:

<https://do.sevsu.ru/mod/folder/view.php?id=559114>

1. Выполнить задание 1 пункта 2 данных методических указаний.

2. **Задание 2.** По построенной бинарной логистической модели определить вероятность заболевания у пациентов по результатам анализа и теста. Примеры значений анализа и теста (3 примера) придумать самостоятельно. Использовать файл *медицина_бин регр.sav*.

3. **Задание 3.** На основе построенной модели определите вероятность покупки фургона американским гражданином с доходом 40 тыс.дол., 5 членами семьи, образованием 9 лет, проживающим в северном регионе (1), придерживающегося консервативных взглядов (2) и имеющего в распоряжении 2 автомобиля. Придумать еще 2 примера самостоятельно. Использовать файл *фургон_бин регр.sav*.

4 Содержание отчета.

1. Титульный лист.
2. Постановка задачи.
3. Результаты решения задач.
4. Анализ и выводы по результатам решения.
5. Общий вывод по практическому заданию.

Практическая работа № 6

ФАКТОРНЫЙ АНАЛИЗ В СТАТИСТИЧЕСКОМ ПАКЕТЕ PSPP

Цель работы: получение практических навыков в проведении факторного анализа с использованием пакета PSPP.

1 Краткие теоретические сведения

1.1 Постановка задачи

Социально-экономические объекты, как правило, являются очень сложными, многообразными, так как их формирование обычно обусловлено действием множества разнообразных факторов (причин). Поэтому для более полной характеристики таких объектов их обычно характеризуют набором разнообразных признаков (параметров, переменных), а результаты измерений совокупности этих признаков представляют в виде многомерных случайных величин. При исследовании таких многопараметрических объектов всегда встает вопрос: нельзя ли отбросить часть параметров или заменить их меньшим числом каких-либо функций от них, сохранив при этом всю информацию? Методы факторного анализа позволяют это сделать. Факторный анализ является одним из разделов современной многомерной статистики и широко используется в различных областях исследовательской деятельности.

1.2 Общие представления о факторном анализе

Факторный анализ – это общее название для класса методов, используемых, главным образом, для сокращения числа переменных и их обобщения.

Можно выделить две цели факторного анализа:

- определение взаимосвязей между признаками (классификация переменных);
- сокращение числа переменных.

Фактор – латентная (скрытая) переменная, конструируемая таким образом, чтобы можно было объяснить корреляцию между набором наблюдаемых переменных. В факторном анализе нет разграничения переменных на зависимые и независимые.

В факторном анализе решаются следующие задачи:

1. Определить количество действующих латентных факторов и указать их относительную интенсивность.
2. Выявить признаковую структуру факторов, т.е. показать, какими признаками объекта (наблюдаемыми переменными) обусловлено действие того или иного фактора и в какой относительной мере.
3. Выявить факторную структуру изучаемых признаков объекта, т.е. показать долю влияния каждого из факторов на значение каждой наблюдаемой переменной объекта.

4. Воссоздать в факторном координатном пространстве облик изучаемого объекта, используя вычисляемые значения факторов для каждого наблюдения исходной выборочной совокупности.

1.3 Алгоритм осуществления факторного анализа

Факторный анализ начинается с определения объекта исследования и описания поставленной проблемы. Формулировка проблемы включает в себя четкое определение назначения и целей факторного анализа.

1. Подготовка данных — это очистка данных от выбросов и пропущенных значений, а также приведение данных к нужному формату.

2. Выбор метода факторного анализа. Метод выбирается в зависимости от цели исследования: метод главных компонент, метод максимального правдоподобия и метод обобщенных наименьших квадратов.

3. Определение числа факторов. На этом этапе необходимо определить, сколько факторов следует использовать для анализа данных.

4. Оценка параметров модели. Это включает в себя оценку факторных нагрузок, факторных весов и дисперсии ошибок.

5. Интерпретация результатов. Это включает в себя анализ факторных нагрузок, которые показывают, насколько каждая переменная связана с каждым фактором, а также анализ факторных весов, которые показывают, какие переменные имеют наибольшее влияние на каждый фактор.

6. Проверка адекватности модели. Осуществляется с помощью различных статистических тестов. Если модель не является адекватной, то необходимо внести корректировки или выбрать другой метод факторного анализа.

Для получения надежных результатов факторного анализа исходные данные должны отвечать определенным требованиям:

1. Модель факторного анализа разрабатывалась для метрических данных. Включение в анализ порядковых и, тем более, бинарных данных, возможно. Но в общем случае желательно перейти к единой шкале для всех признаков, что чревато потерей информации и потерей достоверности и ценности конечности результата.

2. Факторный анализ является одним из разделов многомерного статистического анализа. Он основан на многомерном нормальном распределении, то есть каждый из используемых признаков изучаемого объекта должен быть распределен по нормальному закону. Отсутствие нормальности ухудшает качество решения. Если переменные имеют существенные асимметрию и эксцесс, то ее можно однозначно преобразовать, введя новую, нормально распределенную переменную.

3. Для получения обоснованных результатов объем выборки должен превышать число включенных в исследование переменных в 4-5 раз.

1.4 Модели факторного анализа

При построении модели факторного анализа предполагается, что признаки линейно зависят от факторов.

Существует два подхода:

1) **метод главных компонент (МГК)**, в котором наблюдаемые значения каждого из признаков представляются в виде линейных комбинаций факторных нагрузок и факторов. Факторные нагрузки показывают тесноту взаимосвязи фактора и признака. **Метод главных компонент** основан на попытке объяснить общую дисперсию в заданном наборе переменных при помощи латентных факторов. Основной объект анализа в МГК – это дисперсии. Метод главных компонент ориентирован на выделение малого набора ортогональных компонент таким образом, чтобы они объясняли максимум дисперсии для анализируемого набора данных.

2) **модель собственно факторного анализа (ФА)**, когда наблюдаемые значения определяются не только факторами, но и действием локальных случайных причин.

Общности представляют несколько исходных переменных, а каждая характеристика относится только к одной исходной переменной. Предполагается, что общие факторы и характерные факторы не коррелированы.

Метод ФА пытается объяснить корреляции между переменными. Основной объект анализа в факторном анализе – ковариации (корреляции).

В ФА для каждой переменной анализируется только общая дисперсия, присущая сразу нескольким наблюдаемым переменным и не включаются дисперсия ошибки измерения и специфическая дисперсия какой-то отдельной переменной. Общая дисперсия оценивается **общностями**, стоящими на главной диагонали матрицы взаимосвязей и принимающими значения от 0 до 1. Факторное решение выбирается на основе переменных с высокими общностями. Сумма общностей - это дисперсия взаимодействия факторов, которая не может превысить общую дисперсию наблюдаемых переменных, поскольку специфичная и ошибочная дисперсия опускаются.

Таким образом, можно сказать, что основной объект анализа в методе главных компонент – это дисперсии, а в факторном анализе – ковариации (общности).

МГК ориентирован на выделение малого набора ортогональных компонент таким образом, чтобы они объясняли максимум дисперсии для анализируемого набора данных. Решение является единственным и если сохранить все компоненты, то можно точно воспроизвести наблюдаемую матрицу взаимосвязей. Если необходимо в первую очередь сократить число переменных, то есть уменьшить размерность, то МГК даст возможность определиться как с количеством, так и с природой факторов. Обычно именно МГК используют на этапе разведочного анализа.

Цель **факторного анализа** – при помощи малого набора факторов воспроизвести матрицу взаимосвязей. С математической точки зрения метод главных компонент дает

единственное решение. Это объясняется тем, что теоретически в МГК количество выделенных компонент равно количеству наблюдаемых переменных и общая дисперсия учитывается полностью. В то же время разные виды факторного анализа дают разные решения для одного и того же набора данных.

Оба метода направлены на аппроксимацию корреляционной матрицы определенным числом факторов — меньшим, чем количество переменных, но способами аппроксимации они отличаются. Разумеется, методы дают схожие результаты: если главные компоненты объясняют большую часть дисперсии наблюдаемых переменных, они объясняют и почти всю корреляцию; в том случае, если факторы объясняют корреляцию между наблюдаемыми переменными, они также должны объяснять их дисперсию (пусть и не полностью).

1.5 Ограничения применимости факторного анализа

В основе корреляционного анализа обычно лежит корреляционная матрица. Целесообразность выполнения анализа определяется корреляцией между переменными. Если корреляции между переменными небольшие, то факторный анализ проводить бессмысленно.

Для проверки целесообразности использования факторной модели анализа зависимости переменных существует несколько статистик.

С помощью критерия сферичности **Бартлетта** проверяется нулевая гипотеза об отсутствии корреляции между переменными в генеральной совокупности. Проверка основана на преобразовании детерминанта корреляционной матрицы в статистику χ^2 . При большом значении статистики и малом значении наблюдаемого уровня значимости нулевую гипотезу отвергают. Значение уровня значимости меньше 0,05 указывает статистическую значимость отличия коэффициента корреляции от 0 и соответственно, свидетельствуют о возможности проведения факторного анализа.

Мера выборочной адекватности Кайзера-Мейера-Олкина (**КМО**) основан на сравнении наблюдаемых парных коэффициентов корреляции со значениями частных коэффициентов корреляции. Небольшие значения статистики указывают на то, что корреляции между парами переменных нельзя объяснить другими переменными и что использование факторного анализа нецелесообразно.

1.6 Проблема числа факторов

Обычно заранее неизвестно, сколько факторов необходимо и достаточно для представления данного набора переменных. Процедура факторного анализа предполагает предварительное задание числа факторов. Поэтому на первом, поисковом, этапе имеет смысл использовать анализ главных компонент. Для выбора числа факторов наиболее часто применяют следующие подходы:

1. Критерий **Кайзера** или критерий **собственных чисел**. Этот критерий предложен Кайзером, и является, вероятно, наиболее широко используемым. Отбираются только факторы с собственными значениями равными или большими 1. Это означает, что если фактор не выделяет дисперсию, эквивалентную, по крайней мере, дисперсии одной переменной, то он опускается. **Критерий Кайзера: число факторов равно числу компонент, собственные числа которых больше 1.**

2. Критерий каменистой осыпи или критерий отсеивания (график главных компонент). Он является графическим методом, впервые предложенным психологом Кэттелом. Собственные значения можно изобразить в виде простого графика (рис. 6.1).

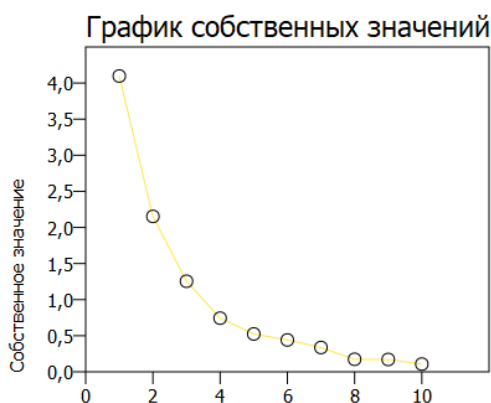


Рисунок 6.1 – График каменистой осыпи

Кэттел предложил найти такое место на графике, где убывание собственных значений слева направо максимально замедляется. Предполагается, что справа от этой точки находится только «факториальная осыпь» — «осыпь» является геологическим термином, обозначающим обломки горных пород, скапливающиеся в нижней части скалистого склона. Однако этот критерий отличается высокой субъективностью и, в отличие от предыдущего критерия, статистически не обоснован. Поэтому обычно проверяют три значения возможного количества факторов — соответствующее точке замедления, меньшее на 1 и большее на 1.

Кроме того, число существенных факторов можно оценить из содержательных соображений, а окончательное решение о числе факторов принимается после их интерпретации.

1.7 Проблема общности

Единичная дисперсия каждой переменной представлена в факторном анализе как сумма общности и характерности.

Общность — это часть дисперсии, обусловленная действием общих факторов. Общность равна сумме квадратов ее нагрузок по всем факторам. Характерность — часть ее дисперсии, обусловленная спецификой данной переменной и ошибками измерения. Характерность переменной дополняет ее же общность до единицы.

Качество факторного анализа тем выше, чем больше полнота факторизации. Низкие ее значения, меньше 0,7, свидетельствуют о желательности сокращения количества признаков или увеличения количества факторов. В компонентном анализе общность каждой переменной равна единице при условии выделения всех главных компонент.

1.8 Проблема вращения и интерпретации

Полученные факторы можно заменить их линейными комбинациями, которые также взаимно не коррелированы. Таким образом, имеется бесконечное множество наборов факторов, удовлетворяющих данной модели. Процедура получения нового набора факторов называется ортогональным вращением факторов. Чаще всего результаты первоначальной факторизации не подлежат интерпретации. Для решения вопроса о распределении переменных по факторам нужно произвести вращение факторов относительно признаков. Соотношение признаков в осях факторов при этом никак не изменятся.

Процедура вращения факторов — это перераспределения дисперсии по определённому методу. Цель вращений — получение простой структуры факторных нагрузок, когда большинство факторных нагрузок близко к нулю. И только некоторые имеют существенные значения.

Вращение бывает ортогональным и косоугольным. При ортогональном вращении сохраняется прямоугольная система координат, то есть факторы не коррелированы. Косоугольное вращение, при котором не сохраняется ортогональность факторов, применяется, если факторы могут быть взаимосвязаны между собой.

Существует около 13 методов вращения, однако из всех наиболее употребителен ортогональный метод «варимакс».

Чаще других используются ортогональные методы вращения:

- **Варимакс** минимизирует количество переменных с высокими значениями нагрузок, связанных с этим фактором. Метод «варимакс» максимизирует дисперсию квадратов нагрузок для каждого фактора, что приводит к увеличению больших и уменьшению малых значений факторных нагрузок. В результате простая структура получается для каждого фактора в отдельности;
- **Квартимакс** минимизирует число факторов, необходимых для объяснения каждой переменной. Этот метод упрощает интерпретацию наблюдаемых переменных;
- **Эквимакс** объединяет методы варимакс, упрощающий факторы, и квартимакс, упрощающий переменные. Минимизируется число переменных с большими факторными нагрузками и число факторов, требуемых для объяснения переменной.

Неортогональные методы вращения – например, облимин, квартимин, квартимакс. Например, аналитические методы, в которых ищется простая вторичная структура, называются непрямыми методами «облимин».

Интерпретация факторов производится по таблице факторных нагрузок после вращения:

1. По каждой переменной выделяется наибольшая **по модулю** нагрузка – как доминирующая и следующая – второстепенная, если она отличается от доминирующей не более, чем на 0,2.

2. По каждому фактору выделяются переменные, имеющие большие нагрузки по этому фактору. При этом учитывается знак нагрузки: если он отрицательный, это означает противоположный полюс переменной.

3. Каждому фактору подбирается наименование исходя из смысла связанных с ним переменных.

1.9 Принятие решения о качестве факторной структуры

Формальное требование к факторной структуре называется принципом простой структуры. Это выражается в том, что каждая переменная имеет близкие к нулю нагрузки по всем факторам, кроме одного.

В настоящее время не существует формальных критериев простоты факторной структуры. Основным критерием является возможность содержательной интерпретации фактора по двум и более исходным переменным.

Разработаны подходы приближения к простой структуре путем пошагового сокращения факторов и переменных.

1. Если выявлен фактор, по которому ни одна из переменных не получила существенно большей (по сравнению с другими факторами) нагрузки, то стоит уменьшить число факторов.

2. Если фактор идентифицируется только по одной переменной, а остальные не вошли в него даже с второстепенными нагрузками, то стоит уменьшить число факторов.

3. Если имеются неоднозначные переменные, то есть переменные, которые имеют примерно одинаковые по модулю факторные нагрузки по двум и более факторам, то эти переменные следует поочередно удалять из исследования.

1.10 Вычисление оценок факторов

Оценки факторных нагрузок являются коэффициентами линейного уравнения, связывающего значения факторов и значения исходных переменных. Они показывают, с каким весом входят исходные значения каждой переменной в оценку факторов. Факторные коэффициенты можно использовать для вычисления факторных оценок для новых объектов, не включенных ранее в факторный анализ.

Факторные оценки – значения факторов для конкретного объекта. Факторные оценки отражают структуру взаимосвязей исходных признаков

1.11 Задачи, решаемые с использованием факторного анализа

С появлением специализированных пакетов статистической обработки информации появилась возможность широкого применения факторного анализа при решении исследовательских и практических задач:

1. Факторный анализ позволяет выделять группы взаимосвязанных переменных.
2. Факторный анализ позволяет сократить исходное множество признаков до нескольких факторов, которые отражают разные стороны исследуемого объекта.
3. Факторный анализ позволяет оценивать количественно комплексные характеристики объектов, учитывая реальную структуру и взаимосвязь исходных переменных, избегая потерь информации и оценивания путем простого суммирования.
4. Результаты факторного анализа могут быть использованы для последующего регрессионного, дискриминантного и кластерного анализа.

2 Факторный анализ в PSPP

1. Вызов процедуры ФА

Вызов процедуры осуществляется командами *Анализ. Factor Analysis...*

Диалоговое окно имеет структуру, представленную на рис. 6.2.

В поле *Переменные* необходимо перенести переменные, включенные в ФА. В зависимости от цели исследования в ФА может участвовать только некоторая часть переменных. Кнопка *Извлечение...* задает метод проведения ФА и устанавливает некоторые параметры (рис. 6.2).

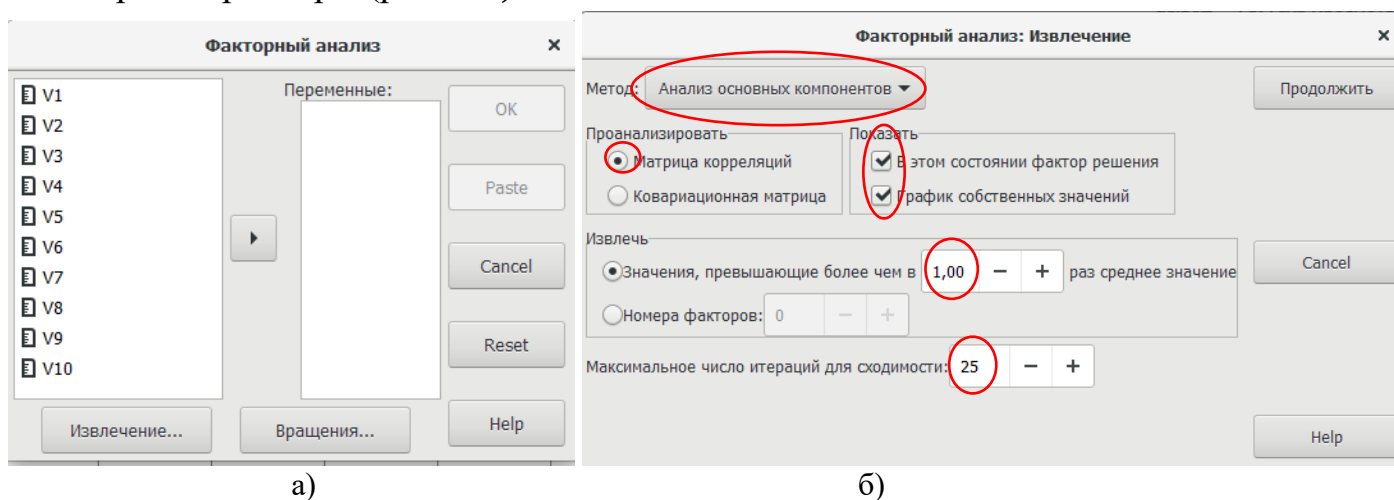


Рисунок 2 – ДО Факторный анализ: а) выбор переменных; б) извлечение

В окне «Извлечение»:

- выбрать метод «Анализ главных компонент»;
- установить переключатель «Матрица корреляций»;

- В поле **Показать** установить флажки «В этом состоянии фактор решения» для вывода таблицы Общностей и «График собственных значений» для вывода графика типа «каменистой осыпи», график можно использовать для выбора числа факторов или порогового собственного значения, поскольку он наглядно отражает различие между большими и маленькими собственными значениями;

- в поле «Значения, превышающие...» установить 1;

- в поле «Максимальное число...» установить 25.

В окне «Вращения...» (рис. 6.3) выбирается метод вращения, рекомендуется метод «Варимакс».

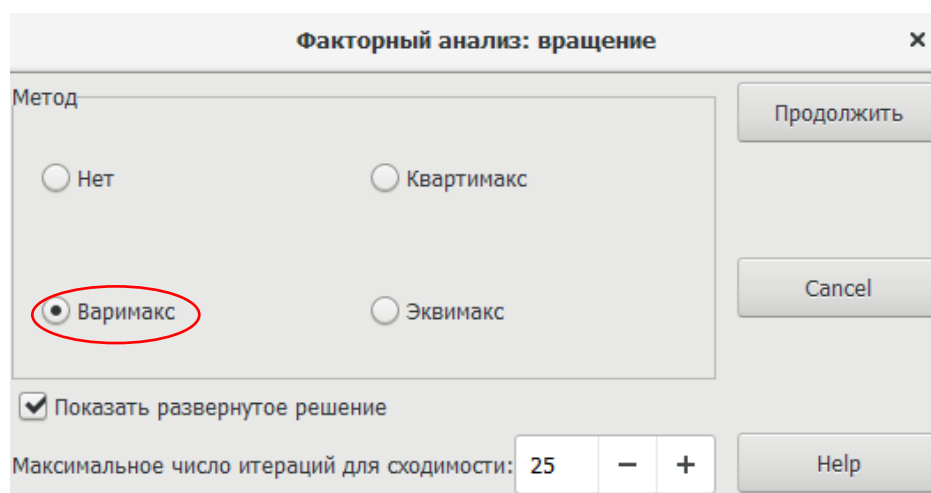


Рисунок 6.3 – ДО «Факторный анализ. вращение»

В результате имеем:

1. Таблица «**Общности**».

Общности		
	Начальный	Извлечение
V1	1,00	,63
V2	1,00	,75
V3	1,00	,78
V4	1,00	,68
V5	1,00	,83
V6	1,00	,81
V7	1,00	,79
V8	1,00	,79
V9	1,00	,86
V10	1,00	,59

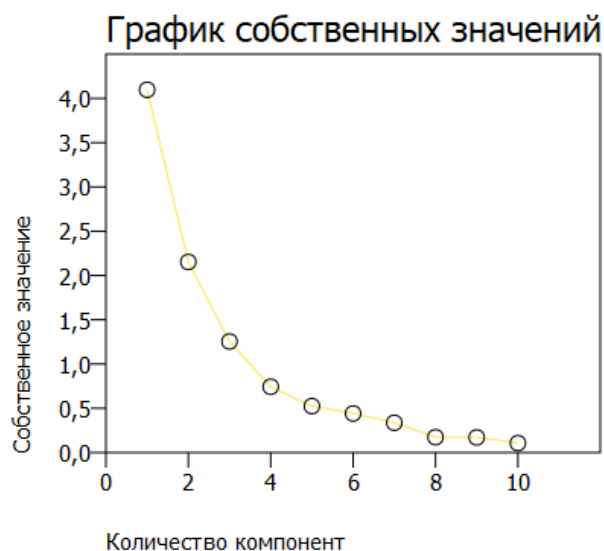
Здесь представлен множественный коэффициент корреляции каждой переменной со всеми остальными. В идеальном случае этот коэффициент равен 1. Можно отметить, что связь между переменными достаточная для проведения ФА.

2. Таблица «Распределение совокупной дисперсии».

Распределение совокупной дисперсии										
	Исходные собственные значения			Извлечения суммы квадратов нагрузок			Суммы квадратов нагрузок вращения			
	Итого	% дисперсии	Накопительный процент	Итого	% дисперсии	Накопительный процент	Итого	% дисперсии	Накопительный процент	
1	4,10	41,0%	41,0%	4,10	41,0%	41,0%	4,01	40,1%	40,1%	
2	2,15	21,5%	62,5%	2,15	21,5%	62,5%	1,94	19,4%	59,6%	
3	1,25	12,5%	75,0%	1,25	12,5%	75,0%	1,55	15,5%	75,0%	
4	,74	7,4%	82,5%							
5	,52	5,2%	87,7%							
6	,44	4,4%	92,1%							
7	,34	3,4%	95,5%							
8	,17	1,7%	97,2%							
9	,17	1,7%	98,9%							
10	,11	1,1%	100,0%							

В этой таблице приводятся проценты общей вариации, объясняемой каждым фактором. Так 1- фактор объясняет 41,0% вариации; 2-й из оставшихся объясняет 21,5%, третий – 12,5% (итого вместе три первых фактора объясняют 75,0% общей вариации). Третья часть таблицы показывает факторные нагрузки после выполнения операции вращения. Как видно, общий процент нагрузки не изменился, но произошло перераспределение нагрузки между факторами. Итак, выделено три фактора.

3. График собственных значений.



Визуально можно определить, что излом графика происходит на 3-4-м значении, остальные переменные имеют небольшие собственные числа и будут распределены между факторами.

4. Таблицы «Матрица компонент» и «Повернутая матрица компонент».

Матрица компонент Повернутая матрица компонент

	Компонент				Компонент		
	1	2	3		1	2	3
V1	,11	-,75	,24	V1	,04	,77	,16
V2	-,34	,28	,75	V2	-,19	,11	-,84
V3	,05	-,56	,68	V3	,07	,83	-,31
V4	,72	-,37	-,13	V4	,64	,29	,42
V5	,84	,30	,17	V5	,89	-,13	-,14
V6	,90	-,04	,05	V6	,89	,10	,14
V7	-,55	-,69	-,10	V7	-,64	,53	,32
V8	,88	-,04	-,11	V8	,84	,02	,28
V9	,90	,07	,21	V9	,92	,08	-,05
V10	,21	-,70	-,24	V10	,08	,50	,58

Обе таблицы представляют информацию о факторной нагрузке, но одна до использования опции вращения, а другая после. Повернутая матрица является оптимальной, поэтому для распределения переменных по факторам используют именно ее.

Для каждой переменной (в строке) выбирают максимальное по абсолютной величине значение и в этот фактор включают рассматриваемую переменную. Для переменной V1 максимальным является значение 0,77, поэтому первая переменная включается во второй фактор. Для второй переменной максимальное по абсолютной величине значение -0,84 относит ее к третьему фактору и т.д.

В результате имеем:

Фактор 1 объединяет следующие переменные:

4. уровень комфорта
5. тишина, уединение
6. хорошее питание
7. наличие развлечений
8. комфорт отдыха с детьми
9. возможность лечения, ухода за здоровьем.

Фактор 2 включает в себя переменные:

1. близость к городу
3. близость водоема, леса.

Фактор 3 в своем составе имеет переменные:

2. приемлемые цены
10. организация детского отдыха.

Итак, в первый фактор попали любители комфортного отдыха, хотя есть вопросы по переменным 5 и 9. Второй фактор объединил любителей природы, которая недалеко от дома. Третий фактор включил предпочитающих экономный отдых ради детей.

3 Задания для выполнения практической работы.

Материалы для выполнения задания находятся по ссылке:

<https://do.sevsu.ru/mod/folder/view.php?id=559989>

Таблицы для выполнения задания представлены в файле **к ПР 3_6.xlsx**.

Задание 1. Проведите факторный анализ с целью сокращения массива данных, содержащих информацию о мотивах туристов, при выборе места отдыха за городом (лист ЭТ Excel **отдых**). Оптимизируйте структуру данных, сократив число переменных. Для этого:

- выявите и извлеките необходимое количество факторов для создания упрощенной структуры:
- разбейте базу данных на группы факторов на основе значений совместной

корреляции;

- подберите названия созданным факторам.

На вопрос «Что для Вас важно при выборе места отдыха за городом?» респонденты выбирали следующие варианты ответа (1 - очень важно, 2 - важно, 3 - и да, и нет, 4 - не важно, 5 - совсем не важно):

1. близость к городу
2. приемлемые цены
3. близость водоема, леса
4. уровень комфорта
5. тишина, уединение
6. хорошее питание
7. наличие развлечений
8. комфорт отдыха с детьми
9. возможность лечения, ухода за здоровьем
10. организация детского отдыха

Задание 2. В исследовании взаимосвязи между семейным и покупательским поведением проводилось измерение следующих утверждений о стиле жизни респондента (1 - не согласен, 7 - согласен) (Таблица в ЭТ Excel **стиль жизни**),

V1: я больше люблю проводить время дома, чем ходить на вечеринки

V2: Покупая товар, я всегда смотрю на цены

V3: мне больше нравится ходить в магазины, чем в кино

V4: я не доверяю рекламе товаров

V5: я - любитель сидеть дома.

V6: я часто пользуюсь купонами скидок.

V7: на рекламу зря тратятся такие большие деньги

Проведите факторный анализ. Оптимизируйте структуру данных, сократив число переменных.

- выявите и извлеките необходимое количество факторов для создания упрощенной структуры;
- разбейте базу данных на группы факторов на основе значений совместной корреляции;
- подберите названия созданным переменным.

Задание 3. Проведите факторный анализ с целью сокращения массива данных, содержащих информацию о потребностях туристов организации отдыха за городом (Таблица в ЭТ Excel **досуг**)

На вопрос: «Хотели бы Вы, чтобы Ваш досуг за городом был организован?» отдыхающие отвечали следующим образом (1 - выбирает, 2 - не выбирает):

1. проведение спортивных состязаний, конкурсов
2. проведение вечерних дискотек
3. пешие прогулки по лесу, тропа здоровья
4. туристические походы
5. катание на лошадях
6. баня, сауна, массаж
7. посещение бассейна
8. катание на катерах, скутерах, катамаранах
9. конкурсы, игры, концерты
10. дневные и вечерние мероприятия для детей, детская комната

Оптимизируйте структуру данных, сократив число переменных. Для этого:

- выявите и извлеките необходимое количество факторов для создания упрощенной структуры;
- разбейте базу данных на группы факторов на основе значений совместной корреляции;
- подберите названия созданным переменным.

4 Содержание отчета.

1. Титульный лист.
2. Постановка задачи.
3. Результаты решения задач.
4. Анализ и выводы по результатам решения.
5. Общий вывод по практическому заданию.

Библиографический список

1. Толстова Ю. Н. Математическая статистика для социологов : учебник и практикум для вузов / Ю. Н. Толстова. — 2-е изд., испр. и доп. — Москва : Юрайт, 2025. — 196 с. — (Высшее образование). — URL: <https://urait.ru/bcode/558884> (дата обращения: 03.03.2025) — ISBN 978-5-534-20860-3. — Текст : электронный.
2. Наследов А. Д. SPSS 15: профессиональный статистический анализ данных / А. Д. Наследов. — СПб. : Питер, 2008. — 416 с. — ISBN 978-5-388-00193-1. — Текст : непосредственный.

ПРИЛОЖЕНИЕ А

Таблица значений F-критерия Фишера при уровне значимости $\alpha=0,05$

Уровень значимости $\alpha = 0,05$											
v_1 v_2	1	2	3	4	5	6	7	8	9	10	12
1	161	200	216	225	230	234	237	239	241	242	244
2	18,51	19,00	19,16	19,25	19,30	19,33	19,36	19,37	19,38	19,39	19,41
3	10,13	9,55	9,28	9,12	9,01	8,94	8,88	8,84	8,81	8,78	8,74
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,90
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74	4,68
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63	3,57
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,34	3,28
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13	3,07
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97	2,91
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,86	2,79
12	4,75	3,88	3,49	3,26	3,11	3,00	2,92	2,85	2,80	2,76	2,69
13	4,67	3,80	3,41	3,18	3,02	2,92	2,84	2,77	2,72	2,67	2,60
14	4,60	3,74	3,34	3,11	2,96	2,85	2,77	2,70	2,65	2,60	2,53
15	4,54	3,68	3,29	3,06	2,90	2,79	2,70	2,64	2,59	2,55	2,48
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42
17	4,45	3,59	3,20	2,96	2,81	2,70	2,62	2,55	2,50	2,50	2,38

Степени свободы: $v_1=df_1=m-1$; $v_2=df_2=n-m$.

Приложение Б

Задания для самостоятельной работы

Задача 1. Четыре бригады выполняли работы по посадке декоративных кустарников. Объем работы, выполненный каждой бригадой за каждый из четырех рабочих дней, приведен в таблице 27. При уровне значимости $\alpha = 0,05$ требуется выяснить, зависит ли объем выполненных работ от работающей бригады?

День	Бригада (фактор)			
	Бригада 1	Бригада 2	Бригада 3	Бригада 4
1-й	140	150	148	150
2-й	144	149	149	155
3-й	142	152	146	154
4-й	145	150	147	152

Задача 2.

Группой из пяти экспертов проведена оценка качества объекта при трех значениях уровня фактора, влияние которого на качество объекта изучается. Значения уровня фактора во время оценки экспертам не известны. При уровне значимости $\alpha = 0,05$ проверить нулевую гипотезу о влиянии фактора на качество объекта. Экспертные оценки уровня качества объекта приведены в таблице:

№	Уровень фактора (Ф)		
	Ф1	Ф2	Ф3
1	18	24	36
2	28	36	12
3	12	28	22
4	14	40	45
5	32	16	40

Задача 3. Используя анализ однофакторной модели, проверить гипотезу о влиянии методик на оценку качества признака при уровне значимости $\alpha = 0,05$.

Исследуется эффективность обучения тремя различными методами. Студентам дается задание изучить тему «Денежная система в Англии». Для этого 10 студентов конспектируют первоисточник, 10 изучают ее по электронному учебнику, 10 - с помощью обучающих компьютерных программ. По окончании их уровень знаний проверяется с помощью теста, состоящего из 100 вопросов. Результаты представлены в таблице:

№	Баллы по тесту для трех методик		
	Первоисточник	Учебник	Компьютер
1	28	39	41
2	33	52	49
3	42	53	56
4	47	54	62
5	48	56	63
6	50	58	64
7	50	59	65
8	51	63	72
9	60	64	77
10	71	77	87

Как в результате проведенного однофакторного дисперсионного анализа можно ответить на следующие вопросы:

1) Влияет ли методика изучения темы на результат? Есть ли значимые различия между тремя выборками по уровню усвоения материала?

2) Есть ли статистически значимая тенденция возрастания показателей в порядке «первоисточник» - «учебник» - «компьютер»?

Задача 4.

В течение шести лет в хозяйстве использовались четыре различных технологии по

выращиванию сельскохозяйственной культуры. Необходимо установить при уровне значимости $\alpha = 0,05$ влияние различных технологий на урожайность культуры. Данные наблюдений представлены в таблице:

Год	Технология (фактор А)			
	A1	A2	A3	A4
1	0,9	1	1,3	1,8
2	0,8	0,7	1,5	1,9
3	1	1,3	1,6	1,1
4	1,3	1	1,4	1,4
5	1,4	1,3	1,2	1,3
6	0,8	1,3	1,1	1,9

Задача 5. Требуется оценить влияние квалификации наладчиков (фактор К) на рассеяние диаметров шариков, которые они изготавливают, при уровне значимости $\alpha = 0,05$. Замеры отклонения диаметра от номинала для каждого из пяти наладчиков проводились по 6 раз. Результаты представлены в таблице:

№	Квалификация наладчика				
	K1	K2	K3	K4	K5
1	1,2	0,6	0,9	1,7	1,0
2	1,1	1,1	0,6	1,4	1,4
3	1,0	0,8	0,8	1,3	1,1
4	1,3	0,7	1,0	1,6	0,9
5	1,1	0,7	1,0	1,2	1,2

Проверяется нулевая гипотеза о равенстве математических ожиданий отклонения для всех пяти наладчиков, то есть предполагается, что квалификация наладчика не влияет на точность изготовления шариков.

Задача 6. Компания, работающая на рынке, хочет оценить влияние обучения персонала на уровень продаж. Кадровая служба компании осуществляет обучение сотрудников отдела продаж в группах (групповое обучение) и организует индивидуальное обучение. Результаты продаж восьми сотрудников после группового и индивидуального обучения приведены в таблице А.8. Используйте однофакторный дисперсионный анализ, задав уровень значимости $\alpha = 0,05$, чтобы выяснить, влияет ли обучение персонала на уровень продаж.

Объем продаж за различные периоды			
Статистический период	Объем продаж до обучения (млн. руб.)	Объем продаж после группового обучения	Объем продаж после индивидуального обучения
1	203	217	213
2	223	230	231
3	206	221	221
4	205	227	222
5	234	218	229
6	204	231	235
7	219	224	213
8	210	225	228
9	233	211	214
10	223	229	225

Задача 7. Используя анализ однофакторной модели, проверить гипотезу о влиянии наглядности задания на оценку качества признака при уровне значимости $\alpha = 0,05$.

Исследуется эффективность методики выдачи задания тремя различными способами. Студентам дается задание разного уровня наглядности. Для этого 10 студентов получают задание от преподавателя устно, 10 - письменно, 10 - в виде текста с иллюстрациями. По окончании их уровень

знаний проверяется с помощью теста, состоящего из 100 вопросов. Результаты представлены в таблице:

№	Баллы по тесту для трех методик		
	Устно	Письменно	Текст с иллюстрациями
1	28	39	41
2	33	52	49
3	42	53	56
4	47	54	62
5	48	56	63
6	50	58	64
7	50	59	65
8	51	63	72
9	60	64	77
10	71	77	87

Как в результате проведенного однофакторного дисперсионного анализа можно ответить на вопрос: влияет ли методика выдачи задания на результат? Есть ли значимые различия между тремя выборками по уровню усвоения материала?

Задача 8. Используя анализ однофакторной модели, проверить гипотезу о влиянии технологии чистовой обработки детали на точность ее изготовления при уровне значимости $\alpha = 0,05$.

Возможно применение трех видов технологии чистовой обработки детали. Исследуется влияние технологии чистовой обработки детали на точность ее изготовления. Проводятся замеры точности обработки детали, по 4 замера отклонения размера детали от номинала в мкм (для каждого вида технологии). Результаты представлены в таблице:

Номер замера	Вид технологии		
	1	2	3
1	1	2	3
2	2	1	2
3	2	3	2
4	1	2	3

Задача 9. Четыре группы студентов на уборке картофеля за четыре дня выполнили объем работы, приведенный в таблице. При уровне значимости $\alpha = 0,05$ требуется выяснить, зависит ли объем выполненных работ от работающей группы.

Дата	Количество убранных рядков картофеля за рабочий день			
	Группа 1	Группа 2	Группа 3	Группа 4
20 августа	140	150	148	150
21 августа	144	149	149	155
22 августа	142	152	146	154
23 августа	145	150	147	152

Задача 10. Достоверны ли при уровне значимости $\alpha = 0,05$ различия приростов бычков в четырех экспериментальных группах по пять животных, данные о которых приведены в таблице?

Группа	№ животного				
	1	2	3	4	5
1	40	24	46	20	35
2	29	27	20	39	45
3	11	31	17	37	39
4	17	21	38	33	21

Задача 11. На одной из опытных станций испытывались шесть местных сортов пшеницы.

Результаты учета урожайности приведены в таблице. Используя однофакторный дисперсионный анализ, выясните, достоверны ли различия в урожайности между сортами при уровне значимости $\alpha = 0,05$.

№ сорта	Повторение			
	1	2	3	4
1	26,1	29,2	30,0	27,3
2	25,0	24,3	28,5	29,0
3	27,2	26,4	31,0	26,4
4	23,6	27,2	25,2	24,8
5	30,0	33,0	36,0	29,8
6	23,0	26,0	26,0	24,8

Задача 12. Результаты оценки влияния различных норм минеральных удобрений на урожайность яровой пшеницы в полевом опыте приведены в таблице. Используя однофакторный дисперсионный анализ, выясните достоверность и степень влияние минеральных удобрений на урожайность при уровне значимости $\alpha = 0,05$.

Норма удобрений, кг д. в./га	Повторение			
	1	2	3	4
0 (Без удобрений)	16,2	15,8	14,0	15,1
Удобрение 1	17,5	18,9	16,7	18,2
Удобрение 2	18,9	18,7	19,3	20,6
Удобрение 3	26,3	24,5	24,3	25,5
Удобрение 4	23,0	23,9	24,2	23,9

Задача 13. Изучается, как влияет тип магазина на товарооборот. Данные о зафиксированном товарообороте (в млн руб.) в магазинах трех типов за каждые 8 месяцев работы приведены в таблице. Используя однофакторный дисперсионный анализ, выясните достоверность влияния тип магазина на товарооборот при уровне значимости $\alpha = 0,05$.

Магазин 1	Магазин 2	Магазин 3
19	20	16
23	20	15
26	32	18
18	37	26
20	40	19
20	24	17
18	22	19
35	18	18

Задача 14. Крупной организации для принятия решения о закупках лицензионного программного обеспечения (ПО) необходимо выяснить значимость выбора поставщика с точки зрения возможности сокращения при этом издержек. Требуется выяснить, насколько выбор поставщика комплекта лицензионного ПО влияет на издержки предприятия. В шести классах градации фактора П (1-й поставщик, 2-й поставщик и т.д.) собраны данные о потенциальных затратах на покупку 100 единиц продукции в тысячах рублей в течение 4 лет. Данные о затратах за год (в тыс. руб.) приведены в таблице. Используйте однофакторный дисперсионный анализ, задав уровень значимости $\alpha = 0,05$, чтобы выяснить, влияет ли выбор поставщика на размер потенциальных затрат.

Поставщик	2018	2019	2020	2021
1-й	72,1	75,5	72,9	76,0
2-й	78,9	82,7	73,9	81,8
3-й	72,5	70,1	73,7	75,3
4-й	77,5	74,1	72,7	74,2
5-й	77,0	82,8	85,5	86,0
6-й	75,6	80,5	80,7	78,4

Задача 15. Менеджер по продажам в сети супермаркетов хочет знать, влияет ли расположение

рекламных щитов на объем продаж товара. Для каждого из трех видов щитов отобрано случайным образом по 6 магазинов, расположенных в соответствующем районе. Объемы продаж за месяц (в млн руб.) приведены в таблице. Используйте однофакторный дисперсионный анализ, задав уровень значимости $\alpha = 0,05$, чтобы выяснить, влияет ли расположение рекламных щитов на объем продаж товара.

Расположение щитов		
1	2	3
8,1	3,9	4,8
7,5	4,4	6,0
5,8	2,8	4,4
6,6	3,4	5,8
5,9	4,8	6,2
4,8	3,6	4,9

Задача 16. Деталь может изготавливаться из трех видов материалов. Производитель проверяет влияние на износостойкость детали материала, из которого она изготовлена. Получены данные по износостойкости пяти деталей для каждого материала: время работы детали до износа, тыс. час. Результаты испытаний приведены в таблице. Используйте однофакторный дисперсионный анализ, задав уровень значимости $\alpha = 0,05$, чтобы проверить влияние на износостойкость детали материала (три вида), из которого она изготовлена.

Материал 1	1,25	1,32	1,28	1,26	1,29
Материал 2	1,12	1,15	1,26	1,19	1,21
Материал 3	1,32	1,33	1,34	1,29	1,30

Задача 17. Требуется оценить эффективность нескольких различных способов действий, направленных на достижение одной цели - получение максимально дешевого кредита на покупку автомобиля. Для этого рассмотрим данные о годовой процентной ставке на кредиты для новых автомобилей в девяти самых крупных банках шести городов США. Данные представлены в таблице. Проверяется нулевая гипотеза о независимости в среднем ставки от города при уровне значимости $\alpha = 0,05$. Выполнить однофакторный дисперсионный анализ.

Данные					
Atlanta	Chicago	Houston	Memphis	New York	Philadelphia
13,75	14,25	14	15	14,5	13,5
13,75	13	14	14	14	12,25
13,5	12,75	13,51	13,75	14	12,25
13,5	12,5	13,5	13,59	13,9	12
13	12,5	13,5	13,25	13,75	12
13	12,4	13,25	12,97	13,25	12
13	12,3	13	12,5	13	12
12,75	11,9	12,5	12,25	12,5	11,9
12,5	11,9	12,5	11,89	12,45	11,9

Проверяется нулевая гипотеза о независимости в среднем ставки от города.

Задача 18. Используя анализ однофакторной модели, проверить гипотезу о влиянии методик на оценку качества признака при уровне значимости $\alpha = 0,05$.

Исследуется эффективность обучения тремя различными методами. Студентам дается задание изучить тему «Денежная система в Англии». Для этого 10 студентов конспектируют первоисточник, 10 изучают ее по электронному учебнику, 10 - с помощью обучающих компьютерных программ. По окончании их уровень знаний проверяется с помощью теста, состоящего из 100 вопросов. Результаты представлены в таблице:

№	Баллы по тесту для трех методик		
	Первоисточник	Учебник	Компьютер
1	28	39	41
2	33	52	49
3	42	53	56

4	47	54	62
5	48	56	63
6	50	58	64
7	50	59	65
8	51	63	72
9	60	64	77
10	71	77	87

Задача 18. Требуется оценить влияние квалификации наладчиков (фактор К) на рассеяние диаметров шариков при уровне значимости $\alpha = 0,05$. Замеры отклонения диаметра от номинала для каждого из шести наладчиков проводились по 6 раз. Результаты представлены в таблице:

№	Квалификация наладчика					
	K1	K2	K3	K4	K5	K6
1	1,2	0,6	0,9	1,7	1,0	1,4
2	1,1	1,1	0,6	1,4	1,4	1,3
3	1,0	0,8	0,8	1,3	1,1	1,1
4	1,3	0,7	1,0	1,6	0,9	1,6
5	1,1	0,7	1,0	1,2	1,2	1,3
6	0,8	0,9	1,1	1,3	1,5	1,5

Проверяется нулевая гипотеза о равенстве математических ожиданий отклонения для всех шести наладчиков, то есть предполагается, что квалификация наладчика не влияет на точность изготовления шариков.

Задача 19. Изучается, как влияет тип магазина на товарооборот. Данные о зафиксированном товарообороте (в млн руб.) в магазинах трех типов за каждые 8 месяцев работы приведены в таблице. Используя однофакторный дисперсионный анализ, нужно выяснить достоверность влияния тип магазина на товарооборот при уровне значимости $\alpha = 0,05$.

Магазин 1	Магазин 2	Магазин 3
19	20	19
23	20	23
26	32	18
18	37	26
20	40	20
20	24	20
18	22	18
35	18	35

Задача 20. Используя анализ однофакторной модели, проверить гипотезу о влиянии технологии чистовой обработки детали на точность ее изготовления при уровне значимости $\alpha = 0,05$.

Возможно применение трех видов технологии чистовой обработки детали. Исследуется влияние технологии чистовой обработки детали на точность ее изготовления. Проводятся замеры точности обработки детали, по 4 замера отклонения размера детали от номинала в мкм (для каждого вида технологии). Результаты представлены в таблице:

Номер замера	Вид технологии		
	1	2	3
1	1	2	3
2	2	1	2
3	2	3	2
4	1	2	3

СТАТИСТИЧЕСКИЙ ПАКЕТ SPSS ДЛЯ СОЦИАЛЬНЫХ НАУК

Методические указания

Часть III

**Анализ взаимосвязей между данными
в статистическом пакете PSPP**

Составитель: **Балакирева** Ирина Аркадьевна

В авторской редакции

Изд. № 107/2025. Объем 4 п. л. Усл. печ. л. 3,72. Уч.-изд. л. 3,92.
Издательский центр ФГАОУ ВО «Севастопольский государственный университет»