Министерство образования и науки РФ
Федеральное агентство по образованию
ГОУ ВПО «Амурский государственный университет»

УТВЕРЖДАЮ
Зав. кафедрой ОМиИ
_____Г.В.Литовка
«__» _____2007 г.

УЧЕБНО-МЕТОДИЧЕСКИЙ КОМПЛЕКС
ПО ДИСЦИПЛИНЕ «Математические методы в социологии»

для специальности 040201 – «Социология»

Составитель: Н.Н. Двоерядкина , к.п.н.

Благовещенск, 2007

Печатается по решению

редакционно-издательского совета

факультета математики и информатики

Амурского государственного университета

Двоерядкина Н.Н.

Учебно-методический комплекс дисциплины «Математические методы в социологии» для специальности 040201 – Благовещенск: АмГУ, 2007. – 137 с.

Содержание

1. РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ.

1.1. Цели и задачи дисциплины, ее место в учебном процессе.

Математическое образование следует рассматривать как важную составляющую подготовки специалиста-социолога. Обусловлено это тем, что математические методы являются не только мощным средством решения прикладных задач и универсальным языком науки, но также и элементом общей культуры.

Современная математика является междисциплинарным инструментарием, который выполняет две основные функции: первую – обучающую специалиста-профессионала умению правильно задавать цель тому или иному процессу, определить условия и ограничения в достижении цели; вторую – аналитическую, т.е. «проигрывание» на моделях возможных ситуаций и получение оптимальных решений.

Основной целью курса «Математические методы в социологии» является развитие мышления, прежде всего формирование абстрактного мышления, способности к абстрагированию, и умения «работать» с абстрактными, «неосязаемыми» объектами.

Конкретные фундаментальные математические знания не являются «предметом первой необходимости» для большинства людей и не могут составлять целевую основу обучения математике социологов.

Именно поэтому в качестве основополагающего принципа математического образования для социологов на первый план выдвигается принцип приоритета развивающей функции в обучении математике. Иными словами, обучение математическим методам ориентировано не столько на математическое образование, сколько на образование с помощью математики.

В соответствии с этим принципом главной задачей математических методов в социологии становится не изучение фундаментальных основ математической науки, а формирование у студентов-социологов в процессе изучения дисциплины базы для организации полноценной профессиональной исследовательской деятельности.

1.2.Содержание дисциплины.

Дисциплина «Математические методы в социологии» изучается в 5 семестре. Учебным планом специальности «Социология» на ее изучение отводится 140 часов, в том числе 36 ч – лекций, 36 ч – практические занятия, 68 ч – самостоятельная ра-

бота.

Лекционные занятия, наименование тем, содержание, объем в часах.

1. Основные понятия математической статистики: выборка ее репрезентативность, нормальное распределение, правило «трех сигм», теорема Ляпунова и ее следствия, понятие статистической гипотезы, уровень значимости, доверительный интервал, зона неопределенности, алгоритм проверки статистических гипотез (2 ч).

2. Параметрические критерии проверки статистических гипотез: критерии Стьюдента, Фишера для зависимых и независимых выборок, однофакторный и многофакторный дисперсионный анализ. Ограничения параметрических критериев (6 ч).

3. Непараметрические критерии проверки гипотез: ранг; ранжирование переменных; критерии различий (Q-Розенбаума, U-Манна-Уитни, H-Крускала-Уоллиса, S-тенденций Джонкира); критерии изменения (G-знаков, T-Вилкоксона, χ^2-Фридмана, L-тенденций Пейджа); алгоритмы, сходства, различия и ограничения критериев; виды задач, решаемых с помощью данных критериев(4 ч).

4. Критерии согласия: таблицы сопряженности и их анализ; эмпирические и теоретические частоты; критерии согласия χ2-Пирсона, λ-Колмогорова-Смирнова, φ-Фишера, их алгоритмы, сходства и различия; примеры задач (4 ч).

5. Корреляционный анализ: типы шкал; измерение связи между переменными, измеренными в разных шкалах; коэффициенты корреляции Пирсона (r_{xy}), Спирмена (ρ), Кендалла (τ), ассоциации (φ), рангово-бисериальный ($R_{гв}$), бисериальный ($R_{бис}$), корреляционное отношение (η) (4 ч).

6. Многомерные статистические методы исследования: многомерный анализ как один из наиболее действенных количественных инструментов исследования социальных процессов, описываемых большим числом характеристик, классификация многомерных методов. Кластерный анализ: постановка задачи, построение дендрограммы, иерархические и неиерархические структуры, агломеративные и дивизитивные иерархические методы. Факторный анализ, и его использование в исследовании связи, выделение латентных переменных (факторов), интерпретация факторных нагрузок и факторных весов, моделирование значений наблюдаемых переменных на

основе выделенных латентных факторов (16 ч).

Практические занятия, их содержание и объем в часах

1. Основные понятия математической статистики: создание выборки, критерии репрезентативная выборка, определение оптимального объема выборки, построение доверительных интервалов(2 ч).

2. Параметрические критерии проверки статистических гипотез: критерии Стьюдента, Фишера для зависимых и независимых выборок; использование критериев для сравнения средних значений и сравнения «разбросов» значений около среднего. Решение задач с комбинированным применением критериев. Однофакторный и многофакторный дисперсионный анализ (8 ч).

3. Непараметрические критерии проверки гипотез: ранжирование переменных; использование критериев различий и изменения для решения задач (4 ч).

4. Критерии согласия: таблицы сопряженности и их анализ; вычисление эмпирических и теоретических частот; анализ данных с помощью критериев согласия $\chi 2$-Пирсона и λ-Колмогорова-Смирнова (4 ч).

5. Корреляционный анализ: определение типа шкалы; измерение связи между переменными, в разных шкалах; нахождение коэффициентов корреляции Пирсона, Спирмена, Кендалла, ассоциации, рангово-бисериальный, бисериальный, корреляционное отношение (4 ч).

6. Многомерные статистические методы исследования: кластерный анализ: определение мер сходства, вычисление расстояний, меры объединения или связи, построение дендрограммы, агломеративные и дивизитивные иерархические методы (дендрограммы), последовательный кластерный анализ, метод к – средних.

Факторный анализ: определение оптимального количества собственных чисел корреляционной матрицы (количества факторов), критерий Кайзера, критерий факторной осыпи, выделение латентных переменных (факторов), нахождение и интерпретация матрицы факторных нагрузок и факторных весов, моделирование значений наблюдаемых переменных на основе выделенных латентных факторов (14 ч).

1.3. График самостоятельной учебной работы студентов.

В качестве самостоятельной работы студентам предлагается:

1. Творческое задание по теме «Многомерные статистические методы исследования», которое заключается в том, что студенты составляют задачу, связанную с их будущей профессиональной деятельностью и находят ее решение с помощью кластерного и факторного анализов. Задача составляется на основании статистических данных, которые собирают студенты во время выполнения курсовых проектов по профессиональным дисциплинам. Таблица данных должна содержать 10-15 различных переменных и не менее 50 наблюдений.

2. Расчетно-графическая работа: по теме «Применение непараметрических критериев к решению задач». Работа рассчитана на 6-7 недель, выполняется по индивидуальным вариантам. Выбор варианта осуществляется согласно порядковому номеру студента в списке группы. По выполнению студент защищает свою работу в индивидуальной беседе с преподавателем.

Кроме того, время, выделенное на самостоятельную работу, распределяется также на выполнение домашних заданий и подготовку к контрольным работам. Домашнее задание задается после каждого занятия и проверяется в начале следующего практического занятия.

№ недели	**Вопросы, изучаемые на лекции**	**Вопросы, изучаемые на практическом занятии**	**Самостоятельная работа**	**Формы контроля**
1	Основные вопросы математической статистики	Репрезентативная выборка, определение оптимального объема выборки, построение доверительных интервалов.		к/р №1
2	Критерий Стьюдента	Критерии Стьюдента для сравнения средних значений двух и более выборок.		
3	Критерий Фишера	Критерии Фишера для сравнения «разбросов» значений. Решение задач с комбинированным применением параметрических критериев.		
4	Дисперсионный анализ.	Однофакторный дисперсионный анализ.		
5	Критерии различий	Многофакторный дисперсионный анализ.		
6	Критерии изменений	Непараметрические критерии различий	Расчетно-графическая работа	Сдача РГР
7	Таблицы сопряженности	Непараметрические критерии изменения.		
8	Критерии согласия	Таблицы сопряженности и их анализ; вычисление эмпирических и теоретических частот.		
9	Корреляционный анализ, типы шкал; измерение связи между переменными, измеренными в разных шкалах.	Анализ данных с помощью критериев согласия $\chi 2$-Пирсона и λ-Колмогорова-Смирнова.		

10	Коэффициенты корреляции Пирсона, Спирмена, Кендалла, ассоциации, ранговобисериальный, бисериальный, корреляционное отношение.	Определение типа шкалы; измерение связи между переменными, в разных шкалах.		
11	Многомерный анализ для исследования социальных процессов	Использование коэффициентов корреляции для решения задач.		
12	Постановка задачи кластерного анализа.	Меры сходства и связи, вычисление расстояний.	Выполнение творческого задания	Защита задания
13	Иерархические кластерные структуры	Построение агломеративных и дивизитивных иерархических дендрограмм.		
14	Метод к-средних	Последовательный кластерный анализ, метод к – средних		
15	Факторный анализ, и его использование в исследовании связи.	Определение оптимального количества факторов, критерий Кайзера, критерий факторной осыпи,		
16	Выделение латентных переменных, факторные нагрузки и факторные весы.	Выделение латентных переменных, нахождение и интерпретация матрицы факторных нагрузок		
17	моделирование значений наблюдаемых переменных на основе выделенных латентных факторов.	Нахождение и интерпретация матрицы факторных весов.		
18	Методы главных компонент и главных факторов, варимаксное вращение.	Вращение факторов.		

1.4. Вопросы к экзамену

1. Репрезентативность выборки.
2. Определение оптимального объема выборки.
3. Нормально распределенные величины, их свойства.
4. Правило «трех сигм».
5. Теорема Ляпунова и ее следствие.
6. Понятие доверительного интервала.
7. Построение доверительных интервалов.
8. Проверка статистических гипотез, алгоритм, цель.
9. Понятие статистического критерия, мощности, уровня значимости.
10. Примеры статистических критериев.
11. Область допустимых значений критерия, критическая область, зона неопределенности.
12. Параметрические критерии проверки статистических гипотез (критерии Стьюдента, Фишера для зависимых и независимых выборок).

13. Использование критериев для сравнения средних значений и сравнения «разбросов» значений около среднего
14. Однофакторный дисперсионный анализ.
15. Многофакторный дисперсионный анализ.
16. Непараметрические критерии проверки гипотез.
17. Критерии различий: Q-Розенбаума, U-Манна-Уитни, H-Крускала-Уоллиса, S-тенденций Джонкира; алгоритмы, сходства, различия и ограничения критериев; виды задач, решаемых с помощью данных критериев.
18. Критерии изменения: G-знаков, T-Вилкоксона, χ^2-Фридмана, L-тенденций Пейджа; алгоритмы, сходства, различия и ограничения критериев; виды задач, решаемых с помощью данных критериев.
19. Критерии согласия χ2-Пирсона и λ-Колмогорова-Смирнова, их алгоритмы, сходства и различия; примеры задач
20. Таблицы сопряженности и их анализ
21. Эмпирические и теоретические частоты, методы их вычисления
22. Корреляционный анализ.
23. Типы переменных и шкал.
24. Измерение связи между переменными в разных шкалах.
25. Коэффициенты корреляции Пирсона (rxy), Спирмена (ρ), Кендалла (τ), ассоциации (φ), рангово-бисериальный (Rгв), бисериальный (Rбис), корреляционное отношение (η).
26. Понятие о многомерных статистических методах исследования.
27. Границы применимости многомерных статистических методов.
28. Классификация многомерных статистических методов.
29. Примеры задач, решаемых с помощью многомерных статистических методов исследования.
30. Постановка задачи кластерного анализа.
31. Меры сходства в кластерном анализе, способы их вычисления.
32. Меры объединения или связи.
33. Построение дендрограммы, агломеративные и дивизитивные иерархические денограммы.

34. Последовательный кластерный анализ, метод к - средних.
35. Постановка задачи факторного анализа.
36. Стандартизация данных.
37. Корреляционные матрицы для исходных и стандартизированных данных, связь между ними.
38. Определение оптимального количества факторов на основании собственных чисел матрицы, критерий Кайзера, критерий факторной осыпи.
39. Выделение латентных переменных.
40. Понятие, методы нахождения и интерпретация матрицы факторных нагрузок и факторных весов.
41. Моделирование значений наблюдаемых переменных на основе выделенных латентных факторов.

1.5. Рекомендуемая литература.

Основная литература

1. Толстова Ю.Н. Измерение в социологии. – М.: ИНФРА-М, 1998. – 223 с.
2. Двоерядкина Н.Н., Киселева А.Н., Юрьева Т.А. Математические методы в психологии и социологии. – АмГУ, ФМиИ: изд-во Амурского гос.ун-та, 2005. – 125 с.
3. Гладышев И. Анализ и обработка данных: специальный справочник. – СПб.: Питер, 2001. – 752 с.

Дополнительная литература

1. Наследов А.Д. SPSS: компьютерный анализ данных в психологии и социальных науках. – СПб.: Питер, 2005. – 415 с..
2. Елисеева И.И. Общая теория статистики. – М.: Финансы и статистика, 2003 г.
3.Сидоренко Е.В. Методы математической обработки в психологии. – Спб.: Соцпсих. центр, 2004. –278 с.
4. Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы. – М., Финансы и статистика, 2003. – 352 с.

2. МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ И УКАЗАНИЯ.

2.1. Методические рекомендации по проведению лекций

Современное исследование общественных процессов, которым занимаются социологи, опирается в значительной степени на использование математических методов анализа. В связи с этим в настоящее время все более интенсивно происходит проникновение математики в социологию, в особенности в конкретные социальные исследования. Это проникновение связано с большими трудностями, т.к. точность и абстрактность математики, с одной стороны обеспечивают ее силу, универсализм и общность, а с другой определенную трудность усвоения этой науки.

Для математики важна не природа рассматриваемых объектов, а лишь существующие между ними соотношения, в то время как социолог работает в основном с реальными данными, результатами проведенного им эксперимента.

Курс «Математические методы в социологии» призван размыть грань между формализмом математики и реальными происходящими в обществе процессами, интересующими социолога.

Значение лекционных занятий по данному курсу обусловлено следующими причинами:

- отсутствием единого учебника, в котором изложены всевозможные математические методы, используемые в социологии;
- необходимостью адаптировать лектором некоторые математические методы для нужд социологии;
- невозможностью студента-социолога самостоятельно представить социологический смысл полученных им знаний по математике.

Каждая лекция сопровождается высоким научным стилем изложения и достаточным количеством примеров профессионального характера, которые разрешают противоречие между желанием поскорее приобщиться к профессии и необходимостью терпеливого изучения фундаментальных дисциплин.

Лекция 1 носит вводный и обзорно-повторительный характер. На ней происходит знакомство студентов с целью, назначением и местом курса в системе учебных дисциплин. А также повторение материала из курса математики, необходимого для осознан-

ного восприятия понятийного аппарата курса «Математические методы в социологии».

Основная цель лекций 2, 3, 4, на которых излагаются параметрические критерии, заключается в демонстрации особенностей данных критериев, систематизации задач для которых их можно использовать. При изложении этих лекций особое внимание следует уделить дисперсионному анализу и его использованию в социологических исследованиях.

Лекции 5 и 6 посвящены непараметрическим критериям различий и изменений. При изложении лекций необходимо четко указать диапазон применения каждого критерия, найти аналог для него среди параметрических критериев. По завершению систематизировать все изученные критерии в виде схемы.

Таблицы сопряженности – очень часто встречаются в социологических исследованиях. Поэтому их анализу уделяется лекция 7. При изложении необходимо обратить внимание на внутреннее содержание этих таблиц, их интерпретацию, понятие частот и возможность применения ранее изученных критериев для анализа таблиц.

На лекции 8 излагаются критерии согласия, которые используются для анализа таблиц сопряженности, а также для решения других исследовательских задач.

9 и 10 лекции посвящены корреляционному анализу, причем, на первой из них дается весь общий понятийный аппарат корреляционного анализа, а на второй конкретные коэффициенты корреляции для различных типов данных, встречающихся в социологии.

Общий анализ многомерных методов исследования приводится в лекции 11. Особое внимание следует уделить классификации данных методов, т.к. в разных источниках можно обнаружить различную классификацию.

Один из самых простых и многомерных методов является кластерный анализ. Его изложению отводится 12, 13, 14 лекции, на которых целесообразно рассмотреть: постановку задачи кластерного анализа, меры сходства и связи, построение агломеративных и дивизитивных иерархических дендрограмм, последовательный кластерный анализ, метод к-средних.

Последние лекции (15, 16, 17, 18) посвящены изложению методов факторного анализа. Необходимо классифицировать данные методы, указать их математические

модели, четко определить матрицы факторных нагрузок и факторных весов, понятие латентного фактора. Указать способы определения оптимального количества факторов и необходимость факторного вращения.

Краткий конспект лекций по каждой теме приводится в п.3.1.

2.2. Методические рекомендации к практическим занятиям.

Лекционный курс дисциплины «Математические методы в социологии» сопровождается практическими занятиями. Теоретические знания, представления, образы должны быть прожиты. Афоризм одного из известных физиков М. Лауэ: знание есть то, что остается, когда все выученное уже забыто, характеризует важную роль практики.

Практические занятия должны проводиться в логичном единстве с теоретическим курсом, подкрепляя и уточняя понятийный аппарат, путем решения задач профессиональной направленности.

Каждый практическое занятие начинается с теоретического опроса необходимого материала и проверки домашнего задания. Далее на конкретных примерах из разных областей социологии рассматриваются пути и способы применения тех математических методов, которые не требуют использования электронных вычислительных машин. При этом выявляются особенности каждой из сформулированных задач, выясняется возможность применения для их решения других известных методов. При этом необходимо активизировать самостоятельную работу студентов. Задания и методические указания к ним выдаются студентам, каждый из которых выбирает оптимальный для себя темп работы. Преподавателю отводится роль консультанта и помощника. Задания, вызвавшие трудности у большинства студентов, разбираются на доске.

При работе студенты должны опираться на систему базовых математических знаний, приобретенных при изучении высшей математики, теории вероятностей и математической статистики, и понимать качественный смысл тех количественных преобразований в области социологии, которые они осуществляют с помощью математических методов.

В конце занятия выдается домашнее задание, состоящее из теоретических вопро-

сов, уяснение которых необходимо для следующего занятия и практических заданий по пройденному материалу.

2.3. Методические указания по выполнению домашних заданий.

При выполнении домашнего задания решать задачи удобнее поэтапно, в той последовательности, в какой эти задания сформулированы. В этом случае при возникновении трудностей будет легче обратиться к анализу тех тем, которые изложены в лекции и задач, разобранных на практическом занятии.

Следует иметь в виду, что решение задач направлено на выработку навыков распознания возможности применения тех или иных математических методов. Поэтому при выполнении заданий контрольной работы требуется абстрагироваться от содержательного анализа предлагаемых задач и формально применить необходимый метод.

При выполнении заданий ответы должны быть аргументированными, то есть недостаточно просто привести ответ, необходимо указать путь, каким Вы пришли к данному ответу, и те основания, которыми Вы руководствовались. При этом следует обратить внимание на то, что ряд заданий предусматривает несколько последовательных шагов или операций для ответа на вопрос. При получении ответа в задаче необходимо правильно интерпретировать его, согласно условию, даже если на Ваш взгляд, данный результат не соответствует действительности.

В случае затруднения с определением алгоритма, необходимого для решения конкретных задач, а также типового оформления ответа на задание, рекомендуется обратиться к образцам выполнения типичных задач, которые представлены на практическом занятии.

После выполнения практической части задания следует найти ответы на теоретические вопросы, заданные преподавателем и таким образом подготовится к осознанному восприятию следующего материала.

Активная, регулярная самостоятельная работа над домашним заданием – путь к успешному усвоению дисциплины.

2.4. Методические указания по выполнению контрольных работ.

По курсу «Математические методы в социологии» предусмотрена одна

контрольная работа по теме: «Параметрические метод проверки статистических гипотез».

Целью написания контрольной работы является выявление уровня знаний студентов по данной теме и умений определять виды задач, к которым применимы параметрические методы.

Написание контрольной работы формирует у студентов способность абстрагироваться от фабулы задачи, строить формализованную математическую модель предложенных явлений, выделять общие закономерности и особенности параметрических методов исследования.

При подготовке к контрольной работе студенту необходимо изучить и систематизировать теоретический материал по теме. Разобрать конкретные примеры проверки статистических гипотез с помощью параметрических методов. Решить достаточное количество задач и упражнений, во время аудиторной и самостоятельной домашней работы.

2.5. Методические указания по организации контроля знаний студентов.

Основной целью учебного процесса в вузе является подготовка высококвалифицированных специалистов, способных творчески решать профессиональные задачи. Контроль и оценка знаний умений и навыков является одним из важных аспектов обучения, который существенно влияет на его качество.

Контролю знаний присущи определенные дидактические правила: объективность, действенность, систематичность, индивидуальность, единство требований.

Отчет по материалу курса только на экзамене не может обеспечить полноту его усвоения студентами. Поэтому в течение семестра предусмотрены и другие виды контроля. При преподавании дисциплины «Математические методы в социологии» используются три основных вида контроля знаний студентов – текущий, тематический и итоговый.

При текущем контроле оценивается уровень участия студентов в аудиторной работе, степень усвоения ими учебного материала и выявляются недочеты по подготовке студентов в целях дальнейшего совершенствования методики преподавания данной дисциплины, активизации работы студентов в ходе занятия и оказания им индивидуальной помощи.

Текущий контроль проводится непосредственно на лекциях и практических занятиях. В процессе чтения лекций преподаватель работает с аудиторией и по ее реакции оценивает степень усвоения материала. В ходе или в конце лекции студентам задается несколько вопросов по изложенной теме, что способствует закреплению полученных знаний. На практических занятиях текущий контроль проводится индивидуально. Полученные знания и степень усвоения материала проверяются в устной или письменной форме.

Тематический контроль проводится после прохождения крупных тем или разделов и осуществляется в следующих формах: контрольная работа, защита расчетно-графической работы, защита творческих заданий.

Итоговым контролем является экзамен. Оценка на экзамене складывается из оценок, полученных за знание теории, умения решать практические задачи и работы в семестре.

3. КОМПЛЕКТЫ ЗАДАНИЙ К ЗАНЯТИЯМ.

3.1.Краткий конспект лекций.

Лекция 1.Основные вопросы математической статистики.

Современное исследование общественных и социальных процессов опирается в значительной степени на использование математических методов анализа. Это требует от специалиста, способного к проведению исследований, достаточно свободного владения математическими методами изучения статистических данных.

Работа по применению математических методов в социологии развивается в нескольких направлениях:

1. решение проблем методологии исследования (выборка, анализа данных, измерение, моделирование);
2. расширение сферы и отраслей социологических знаний, в которых возможно использование математических методов;
3. увеличение количества средств и методов и их модификаций из всевозможных разделов математики, которые применяются при исследованиях;
4. создания собственных математических методов в социологии.

Эти четыре направления представляют единый процесс - ни одно не существует вне других. Поиск нового математического формализма происходит не абстрактно, а на базе уже имеющегося аппарата, и одновременно в ходе решения одной из методологических проблем исследования - выборки, анализа данных, измерения или моделирования.

Проникновение математики в социологию происходит все более интенсивно. Наиболее естественным путем, которым математика проникает в социологию, является математическая статистика. Современное развитие обширного комплекса наук о человеке предполагает широкое использование методов математической статистики уже потому, что именно в этих науках объекты исследования в наибольшей мере удовлетворяют понятию случайных явлений.

Основная задача математической статистики состоит в создании методов сбора и обработки статистических данных для получения научных и практических выводов.

Совокупность предметов или явлений, объединенных каким-либо общим признаком или свойством качественного или количественного характера, называется *объектом наблюдения.*

Результаты статистического наблюдения представляют собой числовую информацию - *данные.*

Статистические данные – это сведения о том, какие значения принял интересующий исследователя признак в статистической совокупности. Признаки бывают количественными и качественными.

Количественным называется признак, значения которого выражаются числами.

Качественным называется признак, характеризующийся некоторым свойством или состоянием элементов совокупности**.**

Статистическая совокупность называется *генеральной*, если исследованию подлежат все элементы совокупности.

Часть элементов генеральной совокупности, подлежащая исследованию, называется *выборочной совокупностью (выборкой*). Она извлекается из генеральной совокупности случайно, так чтобы каждый из *n* элементов выборки имел равные шансы быть отобранным.

Различают повторную и бесповторную выборку. А также механический; случайный; серийный; типический методы отбора.

Независимо от способа организации выборки она должна представлять собой уменьшенную копию генеральной совокупности, т. е. быть *представительной (репрезентативной).*

Число элементов статистической совокупности называется ее *объемом***.**

Размах выборки – это разность между максимальным и минимальным значениями элементов выборки.

Пусть из генеральной совокупности извлекается выборка объемом *n*, причем значение признака x_1 наблюдается m_1 раз, x_2 m_2 раз,..., x_k наблюдается m_k раз,

Значения признака, которые при переходе от одного элемента совокупности к другому изменяются (варьируют), называются вариантами и обычно обозначаются малыми латинскими буквами x_1, x_2 ....

Порядковый номер варианта называется рангом: x_1 – 1-й вариант (1-е значение

признака), x_2 – 2-й вариант (2-е значение признака), x_i– i-й вариант *(i-е* значение признака).

Ряд значений признака (вариантов), расположенных в порядке возрастания или убывания с соответствующими им весами, называется вариационным рядом (рядом распределения).

В качестве *весов* выступают частоты или частости.

Частота (m) показывает, сколько раз встречается тот или иной вариант (значение признака) в статистической совокупности.

Частость или относительная частота (ω_i) показывает, какая часть единиц совокупности имеет тот или иной вариант. Частость рассчитывается как отношение частоты того или иного варианта к сумме всех частот ряда. Сумма всех частостей равна 1

Мы можем сопоставить каждому значению x_i относительную частоту m_i/n.

Статистическим распределением выборки называется перечень возможных значений признака x_i и соответствующих ему частот или относительных частот (частостей) m_i *(ω_i)*.

Для любой случайной величины желательно указать числовые характеристики, важнейшими из которых являются математическое ожидание (средняя), дисперсия и среднее квадратическое отклонение.

Числовые характеристики генеральной совокупности, как правило, неизвестные (средняя, дисперсия и др.), называются параметрами генеральной совокупности (обозначают, например, $\overline{X}$ или $X_{ген}$, $\sigma^2_{ген}$).

По данным выборки рассчитывают числовые характеристики, которые называют статистиками**.** Статистики, получаемые по различным выборкам, как правило, отличаются друг от друга. Поэтому статистика, полученная из выборки, является только *оценкой* неизвестного параметра генеральной совокупности. Когда оценка определяется одним числом, ее называют *точечной оценкой,* а когда двумя числами– границами интервала – *интервальной*.

Для того чтобы любые статистики служили хорошими оценками параметров генеральной совокупности, они должны обладать свойствами: несмещенности, эффективности, состоятельности.

Оценкой генеральной средней является выборочная средняя, которая для различных выборок может быть различна, но *средняя всех возможных выборочных средних равна генеральной средней (*суть центральной предельной теоремы). И что бы мы ни взяли и качестве предмета выборочного обследования, всегда случайные выборки будут распределяться вокруг генеральной средней. Сама по себе центральная предельная теорема малопрактична, но на ее основании можно утверждать, что вероятность ошибки выборки может быть рассчитана путем деления частоты выборки на количество всех возможных выборок.

При большом количестве выборок их параметры распределяются в соответствии с нормальным законом распределения, суть которого состоит в том, что наибольшее число выборочных средних располагается в середине ряда плотности, а крайние значения маловероятны.

При идеальном случайном отборе в пределах одного среднего квадратического отклонения варьируют результаты 68,27 % всех возможных выборок; в пределах двух одного средних квадратических отклонений – 95,45 %; в пределах трех средних квадратических отклонений – 99,73 %.

В процессе статистического анализа иногда бывает необходимо сформулировать и проверить предположения относительно величины независимых параметров или закона распределения изучаемой генеральной совокупности (совокупностей). Такие предположения называются *статистическими гипотезами.*

Статистические гипотезы подразделяются на нулевые и альтернативные.

Выдвинутая гипотеза называется нулевой (основной)**.** Ее принято обозначать H_0. Обычно нулевая гипотеза – это гипотеза об отсутствии различий.

По отношению к высказанной нулевой гипотезе всегда можно сформулировать *альтернативную (конкурирующую),* противоречащую ей. Альтернативную гипотезу принято обозначать H_1.

Цель статистической проверки гипотез состоит в том, чтобы на основании выборочных данных принять решение о справедливости нулевой гипотезы H_0

Так как проверка статистических гипотез осуществляется на основании выборочных данных, то такое решение неизбежно сопровождается некоторой, хотя возможно и очень малой, ошибкой.

Ошибка, состоящая в том, что мы отклонили нулевую гипотезу, в то время как она верна, называется *ошибкой I рода*, а ее вероятность – *уровнем значимости* α.

Ошибка, состоящая в том, что мы приняли нулевую гипотезу, в то время как она неверна, называется *ошибкой II рода,* а ее вероятность обозначают β. Величину равную 1 –β называют *мощностью критерия.*

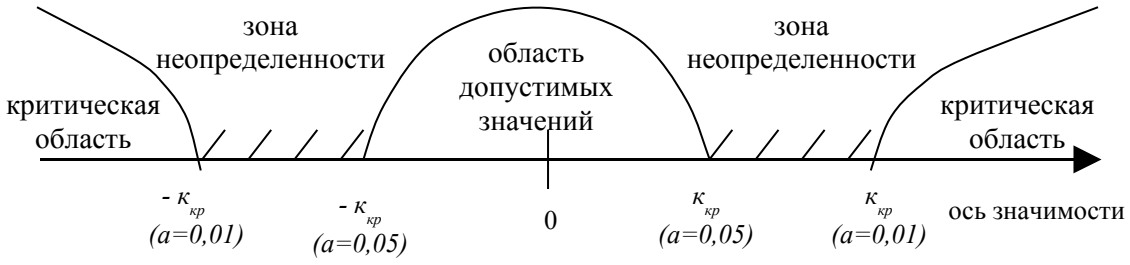
Мощность критерия определяется эмпирическим путем, а уровень значимости задается исследователем. В социологических исследованиях низшим уровнем значимости принято считать α.=0,05 а достаточным α= 0,01.

Статистический критерий – это правило (формула), по которому определяется мера расхождения результатов выборочного наблюдения с высказанной гипотезой H_0.

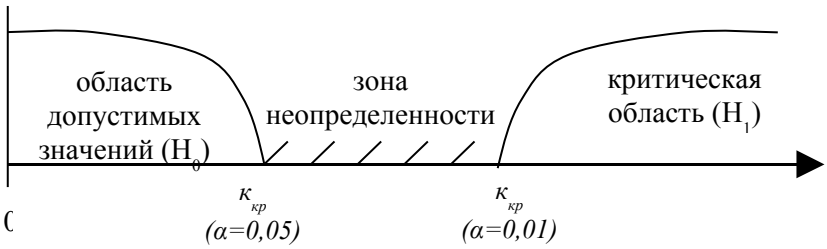
Значение критерия, рассчитываемое по специальным правилам на основании выборочных данных, называется *наблюдаемым* значением критерия.

Значения критерия, определяемые на заданном уровне значимости α по таблицам распределения случайной величины, выбранной в качестве критерия, называются *критическими точками*.

В социологических исследованиях принято определять значения критерия при α=0,01 и α=0,05. Полученные критические точки делят совокупность значений критерия на область допустимых значений или зону незначимости (область принятия нулевой гипотезы), критическую область или зону значимости (область принятия альтернативной гипотезы) и зону неопределенности.



Чаще всего критерии, используемые при социологических исследованиях, имеют положительные значения. Поэтому для простоты при решении прикладных задач изображают только неотрицательную часть оси значимости (рис.2).



Основной принцип проверки статистических гипотез состоит в следующем:

- если наблюдаемое значение критерия принадлежит критической области, то нулевая гипотеза H_0 отклоняется и принимается конкурирующая H_1;

- если наблюдаемое значение критерия принадлежит области допустимых значений, то нулевую гипотезу H_0 нельзя отклонить;

-если наблюдаемое значение критерия принадлежит зоне неопределенности, то мы уже можем отклонить нулевую гипотезу H_0, но еще не можем принять конкурирующую H_1.

Лекция 2. Параметрические критерии: критерий Стьюдента.

Критерии проверки статистических гипотез делятся на параметрические – включающие в формулу расчета параметры распределения (средние и дисперсии) и непараметрические – основанные на оперировании частотами или рангами.

К параметрическим критериям относятся критерии Стьюдента и Фишера.

t-критерий Стьюдента используется

а) для сравнения выборочной средней $\bar{x}$ с некоторым известным числовым значением a_0.

Возможны гипотезы:

H_0: $\bar{x} = a_0$ выборочная средняя генеральной совокупности равна заданному числу a_0.

H_1: $\bar{x} \neq a_0$ ($\bar{x} < a_0$, $\bar{x} > a_0$)выборочная средняя генеральной совокупности не равна (меньше, больше) заданному числу a_0.

Наблюдаемое значение t-критерия рассчитывается по формуле:

- если дисперсия генеральной совокупности неизвестна

$$t_{набл} = \frac{\bar{x} - a_0}{S}\sqrt{n}$$

- если дисперсия генеральной совокупности известна

$$t_{набл} = \frac{\bar{x} - a_0}{\sigma_{ген}}\sqrt{n}$$

где $\bar{x}$ - выборочная средняя;

a_0 – числовое значение генеральной средней;

S – исправленное среднее квадратическое отклонение;

$\sigma^2_{ген}$ - известная дисперсия генеральной совокупности;

n – объем выборки.

Критическое значение $t_{кр}$ следует находить с помощью таблиц распределения Стьюдента по уровню значимости α и числу степеней свободы $k = n - 1$.

b) для обнаружения различия между средними значениями $\bar{x}$, $\bar{y}$ двух выборок.

Возможны гипотезы:

H_0: $\bar{x} = \bar{y}$ средние значения двух выборок равны,

H_1: $\bar{x} \neq \bar{y}$ средние значения двух выборок не равны.

Наблюдаемое значение t-критерия рассчитывается по формуле:

- для независимых выборок

$$t_{набл} = \frac{\bar{x} - \bar{y}}{\sqrt{(n_1 - 1)S_x^2 + (n_2 - 1)S_y^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}$$

- для зависимых выборок

$$t_{эмп} = \frac{\sum d}{\sqrt{\frac{n\sum d^2 - (\sum d)^2}{n - 1}}}$$

где $S_x^2 = \frac{1}{n_1 - 1}\sum (x - \bar{x})^2$ - выборочная дисперсия 1 выборки;

$S_y^2 = \frac{1}{n_2 - 1}\sum (y - \bar{y})^2$ - выборочная дисперсия 2 выборки;

$\bar{x}$ - среднее значение признака для 1 выборки;

$\bar{y}$ - среднее значение признака для 2 выборки;

n_1 – объем 1 выборки;

n_2 – объем 2 выборки;

d – разность между результатами в каждой паре («*после» минус «до»*);

n –число пар данных в зависимых выборках

Критическое значение $t_{кр}$ следует находить с помощью таблиц распределения Стьюдента по уровню значимости α и числу степеней свободы $k = n_1 + n_2 - 2$.

t-критерий для независимых выборок можно использовать для сравнения средних показателей экспериментальной группы с контрольной группой.

t-критерий для зависимых выборок очень полезен в тех ситуациях, когда две сравниваемые группы основываются на одной и той же совокупности наблюдений (субъектов), которые тестировались *дважды* (например, *до* и *после* эксперимента).

Лекция 3. Параметрические критерии: критерий Фишера.

Очень часто при проведении исследований возникает необходимость не только определить наличие различий и сдвигов в значениях признаков, но и воздействие различных факторов на уровень одного и того же признака. Кроме того возникают ситуации, когда наблюдаются явные изменения в значениях признака, а критерий Стьюдента показывает достоверное отсутствие различий.

В таких ситуациях выводы об изменения можно осуществлять с помощью критерия Фишера.

F – критерий Фишера-Снедекора используют

а) для сравнения разброса значений двух выборок, т.е. для проверки гипотезы о равенстве дисперсий.

Возможны гипотезы:

H_0 : $S_x^2 = S_y^2$ - разброс значений признака относительно среднего одинаковый в обеих выборках.

H_1: $S_x^2 \neq S_y^2$ - разброс значений признака не совпадает.

Наблюдаемое значение F – критерия рассчитывается по формуле:

$$F_{набл} = \frac{S_x^2}{S_y^2},$$

где S_x^2 – большая (по величине) выборочная дисперсия;

S_y^2 – меньшая (по величине) выборочная дисперсия.

Критическое значение $F_{крит}$ следует находить с помощью таблицы распределения Фишера-Снедекора по уровню значимости α и числу степеней свободы $k_1 = n - 1$ и $k_2 = n_2 - 1$,

где k_1 – число степеней свободы большей (по величине) дисперсии;

k_2 – число степеней свободы меньшей (по величине) дисперсии;

n_1 – объем выборки большей (по величине) дисперсии;

n_2 – объем выборки меньшей (по величине) дисперсии.

b) для выявлений тенденций изменения признака в трех и более выборках при переходе от условия к условию, т.е. в однофакторном дисперсионном анализе.

Возможны гипотезы:

для независимых выборок: (влияние разных условий на разных испытуемых)

H_0 : разные условия не влияют на изменение значений признака;

H_1: условия влияют на изменение значений признака.

для зависимых выборок (одни и те же испытуемые, но в разных условиях) возможно два варианта гипотез:

а) H_0: условия не влияют на изменение признака;

H_1: условия влияют на изменение значений признака.

б) H_0: индивидуальные различия испытуемых не влияют на изменение значений признака;

H_1: индивидуальные различия между испытуемыми влияют на изменение значений признака.

Наблюдаемое значение критерия рассчитывается по формулам:

для независимых выборок $F_{набл} = \frac{S^2_{факт}}{S^2_{сл}}$,

для зависимых выборок а) $F_{набл} = \frac{S^2_{факт}}{S^2_{сл}}$,

б) $F_{набл} = \frac{S^2_{испыт}}{S^2_{сл}}$,

где $S^2_{факт} = \frac{n \cdot \sum \left(\bar{x}_{гр} - \bar{x}\right)^2}{\kappa_{факт}}$ - характеризует изменение признака, обусловленное действием фактора (условия).

$S^2_{общ} = \frac{\sum \left(x_{ij} - \bar{x}\right)^2}{\kappa_{общ}}$ - характеризует общее изменение признака по всем выборкам и наблюдениям.

$S^2_{испыт} = \frac{m \cdot \sum \left(\bar{x}_{испыт} - \bar{x}\right)^2}{\kappa_{испыт}}$ - характеризует изменение признака, обусловленное ин-

дивидуальными особенностями испытуемых.

$$S_{сл}^2 = \frac{\sum \left(x_{ij} - \bar{x}\right)^2 - n \cdot \sum \left(\bar{x}_{гр} - \bar{x}\right)^2}{к_{общ} - к_{факт}}$$ - для независимых выборок

$$S_{сл}^2 = \frac{\sum \left(x_{ij} - \bar{x}\right)^2 - n \cdot \sum \left(\bar{x}_{гр} - \bar{x}\right)^2 - m \cdot \sum \left(\bar{x}_{испыт} - \bar{x}\right)^2}{к_{общ} - к_{факт} - к_{испыт}}$$ - для зависимых выборок.

n – количество наблюдений,

m – количество выборок (групп),

$\bar{x}_{гр}$ - среднее значение признака в каждой группе,

$\bar{x}$ - среднее значение признака по всей совокупности,

$\bar{x}_{испыт}$ - среднее значение признака каждого испытуемого,

к – число степеней свободы:

$к_{общ} = n \cdot m - 1 \quad к_{факт} = m - 1 \quad к_{испыт} = n - 1$

Критические значения $F_{крит}$ следует находить с помощью таблицы распределения Фишера-Снедекора по уровню значимости α и числу степеней свободы $k_2 = k_{сл}$, и в зависимости от проверяемых гипотез $k_1 = k_{факт}$ или $k_1 = k_{испыт}$

Лекция 4. Параметрические критерии: многофакторный дисперсионный анализ.

Мир по своей природе сложен и многомерен. Ситуации, когда некоторое явление полностью описывается одной переменной, чрезвычайно редки. При проведении типичного эксперимента приходится иметь дело с большим количеством факторов и в этом случае использовать многофакторный дисперсионный анализ.

Основная идея многофакторного дисперсионного анализа состоит в сравнении дисперсий, порождаемых факторами и остаточной дисперсии, порождаемой случайными причинами. Если различие между дисперсиями значимо, то фактор оказывает существенное влияние на переменную. Кроме того, исследуется взаимодействие и воздействие нескольких факторов на несколько постоянных или случайных уровнях и выясняется влияние отдельных уровней и их комбинаций.

Общая дисперсия в случае многофакторного дисперсионного анализа имеет следующие источники: 1 – случайная ошибка (внутригрупповая дисперсия); 2 – изменчивость, связанная с принадлежностью к различным группам; 3 – изменчи-

вость, связанная с уровнем наблюдения; 4 – изменчивость, связанная с взаимодействием факторов. Последний тип изменчивости специфичен для многофакторного дисперсионного анализа.

Многофакторный дисперсионный анализ позволяет изучить каждый фактор, управляя значениями других факторов и обнаружить эффекты взаимодействия между факторами.

Рассмотрим эффекты взаимодействия на примере.

Пусть имеется две группы студентов, причем студенты первой группы настроены на выполнение поставленных задач более целеустремленно, чем студенты второй группы. Разобьем каждую группу пополам и предложим одной подгруппе сложное задание, а другой простое. После этого измерим, насколько напряженно студенты работают над этими заданиями. Средние значения вымышленного эксперимента представлены в таблице:

	целеустремленные	ленивые
Сложное задание	10	5
Простое задание	5	10

Анализируя таблицу нельзя утверждать, что над трудным заданием студенты трудятся напряженнее, а целеустремленные студенты работают упорнее, чем ленивые. Можно лишь сказать, что над сложными заданиями целеустремленные работают более упорно чем над легкими, в то время как над легкими только ленивые студенты работают более упорно.

То есть характер студентов и сложность задания, взаимодействуя между собой, влияют на затрачиваемые усилия – это пример попарного взаимодействия.

Попарные взаимодействия объяснить сравнительно легко, взаимодействия высших порядков объясняются намного сложнее, поэтому в общем случае взаимодействие между факторами описывается в виде изменения одного эффекта под воздействием другого, т.е. изменение фактора, характеризующего сложность задачи под воздействием фактора, описывающего характер студента.

Для взаимодействия трех факторов можно сказать, что два фактора изменяются под воздействием третьего, например, сложность задачи и характер студента под воздействием пола.

Суть метода многофакторного дисперсионного анализа аналогична однофакторному дисперсионному анализу. Но количество гипотез увеличивается, напри-

мер, в случае двухфакторного дисперсионного анализа:

1) H_0: различия, обусловленные действиями фактора А, не более выражены чем случайные различия.

H_1: различия, обусловленные действиями фактора А, более выражены чем случайные различия.

2) H_0: различия, обусловленные действиями фактора В, не более выражены чем случайные различия.

H_1: различия, обусловленные действиями фактора В, более выражены чем случайные различия.

3) H_0: влияние фактора А одинаково при разных градациях фактора В.

H_1: влияние фактора А при разных градациях фактора В различно.

Лекция 5. Непараметрические критерии: критерии различий.

Параметрические методы проверки статистических гипотез требуют знания закона распределения вероятностей в генеральной совокупности. При работе с этими методами предполагается, что распределение экспериментальных данных близко к нормальному, а для измерения использована интервальная шкала.

На практике требования нормальности и интервальности используемой шкалы часто не выполняются , поэтому возникает необходимость использовать непараметрические критерии, которые оперируют частотами или рангами.

Непараметрические критерии делятся на критерии различий и критерии изменений.

Рассмотрим критерии различий.

а) Q-критерий Розенбаума

Критерий используется для оценки различий между двумя выборками по уровню какого-либо признака, количественно измеренного.

Возможны гипотезы:

H_0: Уровень признака в выборке 1 не превышает уровня признака в выборке 2.

H_1: Уровень признака в выборке 1 превышает уровень признака в выборке 2.

Ограничения критерия Q

1) В каждой из сопоставляемых выборок должно быть не менее 11 наблюдений. При этом объемы выборок должны примерно совпадать. Если в обеих выборках меньше 50 наблюдений, то абсолютная величина разности между объемами выборок n_1 и n_2, соответственно, не должна быть больше 10 наблюдений. Если в каждой из выборок больше 51 наблюдения, но меньше 100, то абсолютная величина разности между объемами выборок n_1 и n_2, соответственно, не должна быть больше 20 наблюдений. Если в каждой из выборок больше 100 наблюдений, то допускается, чтобы одна из выборок была больше другой не более чем в 1,5-2 раза.
2) Диапазоны разброса значений в двух выборках должны не совпадать между собой, в противном случае применение критерия бессмысленно.
3) Измерение может быть проведено по шкале порядка, интервалов или отношений.
4) Выборки должны быть независимыми.

Эмпирическое значение критерия подсчитывается по формуле:

$$Q_{эмп} = S_1 + S_2,$$

где S_1 – количество наблюдений в выборке 1, которые выше максимального значения в выборке 2;

S_2 - количество наблюдений в выборке 2, которые ниже минимального значения выборки 1.

Значения в выборках должны быть упорядочены по возрастанию признака.

Критические значения Q-критерия определяются по таблице 2 приложения для данных n_1 и n_2 и для выбранного уровня значимости. Если $Q_{эмп}$ не меньше $Q_{кр}$, то H_o отвергается.

б) U- критерий Манна-Уитни

Критерий предназначен для оценки различий между двумя выборками по уровню какого-либо признака, количественно измеренного.

Возможны гипотезы:

H_0: Уровень признака в группе 2 не ниже уровня признака в группе 1.

H_1: Уровень признака в группе 2 ниже уровня признака в группе 1.

Ограничения критерия U

1) В каждой выборке должно быть не менее 3 наблюдений. Допускается, чтобы в одной выборке было 2 наблюдения, но тогда во второй их должно быть не менее 5.
2) В каждой выборке должно быть не более 60 наблюдений.
3) Измерение может быть проведено по шкале интервалов или отношений.
4) Выборки должны быть несвязными.

Наблюдения обеих выборок необходимо объединить и проранжировать по степени нарастания признака.

Эмпирическое значение критерия рассчитывается по формуле:

$$U_{эмп} = (n_1 \cdot n_2) + \frac{n_x \cdot (n_x + 1)}{2} - T_x,$$

где n_1, n_2 - количество испытуемых в выборках 1 и 2 соответственно,

T_x – большая из ранговых сумм,

n_x – количество испытуемых в группе с большей суммой рангов.

Критическое значение критерия определяется по таблице для данных n_1 и n_2 и выбранного уровня значимости. Если $U_{эмп}$ больше $U_{кр}$, то принимается H_o.

в) H - критерий Крускала – Уоллиса

Критерий предназначен для оценки различий между тремя и более выборками по уровню какого-либо признака. Он позволяет установить, что уровень признака изменяется при переходе от группы к группе, но не указывает направление этих изменений.

Возможны гипотезы:

H_o: Между выборками 1, 2, 3 и т. д. существуют лишь случайные различия по уровню исследуемого признака.

H_1: Между выборками 1, 2, 3 и т. д. существуют неслучайные различия по уровню исследуемого признака.

Ограничения критерия H

1) Измерение может быть проведено по шкале интервалов или отношений.
2) Выборки должны быть независимыми.
3) При сопоставлении 3-х выборок допускается, чтобы в одной из них $n=3$, а в двух других $n=2$. Но при таких численных составах установить различия можно лишь на низшем уровне значимости.

4) Критические значения критерия H и соответствующие им уровни значимости приведены в таблице. Таблица предусмотрена только для трех выборок. При этом максимальное число испытуемых во всех трех выборках может быть не больше 5. При большем количестве выборок и испытуемых в каждой выборке необходимо пользоваться таблицей критических значений критерия χ^2. Число степеней свободы при этом определяется как ν=c-1, где c – количество сопоставляемых выборок.

Наблюдения всех выборок необходимо объединить и проранжировать по степени нарастания признака.

Эмпирическое значение критерия H подсчитывается по формуле:

$$H_{эмп} = \frac{12}{N(N+1)} \cdot \sum_{i=1}^{c} \frac{T_i^2}{n_i} - 3 \cdot (N+1),$$

где N – общее количество испытуемых в объединенной выборке,

n_i – количество испытуемых в каждой выборке,

T_i^2 – квадраты сумм рангов по каждой i–ой выборке.

Если эмпирическое значение критерия меньше критического значения, то H_o принимается.

г) S – критерий тенденций Джонкира

Критерий S предназначен для выявления тенденций изменения признака при переходе от выборки к выборке при сопоставлении трех и более выборок.

Гипотезы:

H_o: Тенденция возрастания значений признака при переходе от выборки к выборке является случайной.

H_1: Тенденция возрастания значений признака при переходе от выборки к выборке не является случайной.

Ограничения критерия

1) Измерение может быть проведено по шкале интервалов или отношений.
2) Выборки должны быть независимыми.
3) Количество наблюдений в каждой выборке должно быть одинаковым.
4) Нижняя граница применимости критерия: не менее трех выборок и не менее двух элементов в каждом наблюдении. Верхняя граница определяется

таблицей приложения: не более 6 выборок и не более 10 наблюдений в каждой выборке.

Выборки необходимо располагать по возрастанию суммы значений признака слева на право.

Эмпирическое значение критерия рассчитывается по формуле:

$$S_{эмп}=2A-B,$$

где A – общая сумма инверсий,

$B = \frac{c \cdot (c-1)}{2} \cdot n^2$ – максимально возможное значение величины А.

Под числом инверсий понимается число значений признака, больших каждого конкретного значения рассматриваемой выборки и расположенных правее от нее.

Критические значения критерия определяются по таблице, в соответствии с выбранным уровнем значимости, количеством выборок (c) и числом наблюдений (n) в каждой выборке.

Если эмпирическое значение критерия меньше критического значения, то принимается гипотеза H_o.

Лекция 6. Непараметрические критерии: критерии изменений.

В социологических исследованиях часто бывает важно доказать, что в результате действия каких-либо факторов произошли достоверные изменения (сдвиги) в измеряемых показателях. В зависимости от воздействующих факторов выделяют временные, ситуационные, умозрительные, измерительные сдвиги, сдвиги под влиянием экспериментальных воздействий и структурные сдвиги. Для оценки достоверности сдвигов применяют критерии изменений.

а) G-критерий знаков

Критерий знаков предназначен для установления общего направления сдвига исследуемого признака. Он позволяет установить, в какую сторону в выборке в целом изменяются значения признака при переходе от первого измерения ко второму: изменяются ли показатели в сторону улучшения, повышения или усиления или, наоборот, в сторону ухудшения, понижения или ослабления.

Возможны гипотезы:

H_0: Преобладание типичного направление сдвига является случайным.

H_1: Преобладание типичного направление сдвига не является случайным.

Ограничения критерия G

1) Измерение может быть проведено по шкале порядка, интервалов или отношений.
2) Выборка должна быть однородной и связной.
3) Объем выборки должен быть равным от 5 до 300.
4) При равенстве типичных и нетипичных сдвигов критерий знаков неприменим.

Эмпирическое значение критерия $G_{эмп}$ принимают равным числу нетипичных сдвигов, т. е. не преобладающих сдвигов в сторону увеличения или уменьшения показателя.

Критическое значение критерия $G_{кр}$ определяют по таблице в соответствии с выбранным уровнем значимости и объемом выборки без учета нулевых сдвигов. Если $G_{эмп}$ не превосходит $G_{кр}$, то гипотеза H_0 отвергается.

б) Парный критерий Т – Вилкоксона

Критерий применяется для сопоставления показателей, измеренных в двух разных условиях на одной и той же выборке испытуемых. Он позволяет установить не только направленность изменений, но и их выраженность. С его помощью мы определяем, является ли сдвиг показателей в каком-то одном направлении более интенсивным, чем в другом.

Возможны гипотезы:

H_0: Интенсивность сдвигов в типичном направлении не превосходит интенсивности сдвигов в нетипичном направлении.

H_1: Интенсивность сдвигов в типичном направлении превышает интенсивности сдвигов в нетипичном направлении.

Ограничения критерия Т

1) Измерение может быть проведено по любой шкале, кроме номинальной.
2) Выборка должна быть связной.
3) Объем выборки должен быть равным от 5 до 50.

Эмпирическое значение критерия подсчитывают по формуле:

$$T_{эмп} = \sum R_r ,$$

где R_r – ранговые значения сдвигов с более редким знаком.

Критическое значение критерия $T_{кр}$ определяется для данного объема выборки и выбранного уровня значимости по таблице 7 приложения. Если $T_{эмп}$ не превосходит $T_{кр}$, то гипотеза H_0 отвергается.

в) Критерий χ_r^2 Фридмана

Критерий применяется для сопоставления показателей, измеренных в трех или более условиях на одной и той же выборке испытуемых. Он позволяет установить, что величины показателей от условия к условию изменяются, но при этом не указывает на направление изменений.

Возможны гипотезы:

H_0: Между показателями, полученными (измеренными) в разных условиях, существуют лишь случайные различия.

H_1: Между показателями, полученными (измеренными) в разных условиях, существуют неслучайные различия.

Ограничения критерия χ_r^2

1) Измерение может быть проведено по шкале интервалов или отношений.
2) Выборка должна быть связной.
3) В выборке должно быть не менее двух испытуемых, каждый из которых имеет не менее трех показателей. Количество измерений не может превышать 100.

Эмпирическое значение критерия вычисляется по формуле:

$$\chi^2_{r\,эмп} = \left[\frac{12}{n \cdot c \cdot (c+1)} \cdot \sum_{i=1}^{c} T_i^2 \right] - 3 \cdot n \cdot (c+1) \ ,$$

где c – количество условий,

n – количество испытуемых,

T_i – суммы рангов по каждому из условий.

Критическое значение критерия $\chi_r^2{}_{кр}$ определяем при выбранном уровне значимости и данном объеме выборки по правилам:

1) При $c=3$ и $n \leq 9$, критические значения определяются по одной таблице.

2) При *c=4* и *n≤4*, критические значения определяются по другой таблице.

3) При большем числе измерений и испытуемых критические определяются по таблице для критерия χ^2. В этом случае число степеней свободы определяется по формуле *v =c – 1*.

Если $\chi_r^2{}_{эмп}$ не меньше $\chi_r^2{}_{кр}$ то гипотеза H_0 отклоняется.

г) L – критерий тенденций Пейджа.

Критерий L Пейджа применяется для сопоставления показателей, измеренных в трех и более условиях на одной и той же выборке испытуемых. Критерий позволяет выявить тенденции в изменении величин признака при переходе от условия к условию, а также указывает на направление этих изменений.

Возможны гипотезы:

H_0: Увеличение индивидуальных показателей при переходе от первого условия ко второму, а затем к третьему и далее, случайно.

H_1: Увеличение индивидуальных показателей при переходе от первого условия ко второму, а затем к третьему и далее, неслучайно.

Ограничения критерия L

1) Измерение может быть проведено по ранговой шкале, шкале интервалов или отношений.
2) Выборка должна быть связной.
3) В выборке должно быть не менее двух и не больше 12 испытуемых, каждый из которых имеет не менее трех показателей. Максимальное число условий – 6.

Эмпирическое значение критерия определяется по формуле:

$$L_{эмп} = \sum_{i=1}^{c} (T_i \cdot i),$$

где c – количество условий,

T_i – суммы рангов по каждому из условий,

i –порядковый номер, приписанный каждому условию, после упорядочения по возрастанию сумм рангов.

Критическое значение критерия $L_{кр}$ определяем при выбранном уровне значи-

мости, данном объеме выборки и данном количестве условий по таблице. Если $L_{эмп}$ не меньше $L_{кр}$, то гипотеза H_0 отклоняется.

Лекция 7. Критерии согласия.

Во многих ситуациях необходимо выполнять проверку на соответствие нормальному закону распределения случайной величины, полученной по экспериментальным данным. От этого зависит корректность применения параметрических критериев проверки статистических гипотез и некоторых многомерных методов анализа. Для проверки соответствия экспериментально распределения нормальному или какому-либо другому теоретическому распределению используют критерии согласия.

Наиболее часто используют следующие статистические критерии: χ^2 – Пирсона, . λ – Колмогорова-Смирнова и φ – Фишера.

1. χ^2 *- критерий Пирсона*.

Основная расчетная формула эмпирического значения χ^2:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_э - f_m)^2}{f_m},$$

где k - количество разрядов признака,

f_m - теоретическая частота,

$f_э$ - эмпирическая частота.

Если количество разрядов равно двум (принимает минимально возможное значение), то в расчетную формулу вносится поправка на непрерывность:

$$\chi^2 = \sum_{i=1}^{k} \frac{(|f_э - f_m| - 0{,}5)^2}{f_m}. ,$$

Критическое значение $\chi^2_{кр}$ определяется по таблице в соответствии с определенным числом степеней свободы и уровнем значимости.

Используется в двух вариантах:

а) для сопоставления эмпирического распределения с теоретическим; в этом случае проверяется нулевая гипотеза H_0 об отсутствии различий между теоретическим и эмпирическим распределением.

В качестве теоретических распределений могут выступать, например, равномерное или нормальное распределения.

В случае равномерного распределения теоретические частоты подсчитываются по формуле:

$$f_m = \frac{n}{k},$$

где n – количество наблюдений.

Число степеней свободы $v=k-1$.

В случае нормального распределения теоретическая частота подсчитывается по формуле:

$$f_m = \frac{nh}{\sigma} \cdot \varphi(u_i),$$

где h – шаг (разность между двумя соседними значениями признака),

$$u_i = \frac{x_i - \bar{x}}{\sigma},$$

$\varphi(u)$- нормированная дифференциальная функция Лапласа.

Число степеней свободы рассчитывают как $k-3$.

б) как расчет однородности двух и более независимых экспериментальных выборок; в этом случае проверяется гипотеза H_0 об отсутствии различий между эмпирическими (экспериментальными) распределениями.

В случае, когда число переменных в двух сравниваемых выборках велико, можно использовать для вычисления $\chi^2_{эмп}$ следующую формулу:

$$\chi^2_{эмп} = 4\sum_{i=1}^{k} \frac{(f_{k1})^2}{f_{k1} + f_{k2}} - 2n,$$

где f_{k1}- частоты первого распределения,

f_{k2}- частоты второго распределения,

n- число элементов в каждой выборке.

Если число значений в выборках различно, то используют формулу:

$$\chi^2_{эмп} = \frac{(n_1 + n_2)^2}{n_1 n_2}\left(\sum_{i=1}^{k} \frac{(f_{k1})^2}{f_{k1} + f_{k2}} - \frac{n_1^2}{n_1 + n_2}\right).$$

Для применения критерия хи-квадрат необходимо соблюдать следующие

условия:

1) измерение может быть проведено по любой шкале;
2) выборки должны быть случайными и независимыми;
3) желательно, чтобы объем выборки был больше 20. С увеличением объема выборки точность критерия повышается;
4) теоретическая частота для каждого выборочного интервала не должны быть меньше 5;
5) сумма наблюдений по всем интервалам должна быть равна общему количеству наблюдений.

2. λ- критерий Колмогорова-Смирнова

Критерий λ предназначен для сопоставления двух распределений:

а) эмпирического распределения с теоретическим;

б) одного эмпирического распределения с другим эмпирическим распределением.

Для применения λ- критерия необходимо соблюдать следующие условия:

1) измерение может быть проведено по шкале интервалов или отношений;
2) выборки должны быть случайными и независимыми;
3) желательно, чтобы суммарный объем двух выборок был большим или равным 50. С увеличением объема выборки точность критерия повышается;
4) эмпирические данные должны допускать возможность упорядочения по возрастанию или убыванию какого-либо признака и обязательно отражать какое-то его однонаправленное изменение.

Возможны гипотезы:

H_0: Различия между двумя распределениями не достоверны.

H_1: Различия между двумя распределениями достоверны.

Эмпирическое значение критерия для сопоставления эмпирического распределения с теоретическим определяется по формуле:

$$d_{эмп} = \max \frac{\left| f_э^* - f_m^* \right|}{n},$$

где $f_э^*$ - накопленные эмпирические частоты,

f_m^* - накопленные теоретические частоты.

Критическое значение критерия определяется по таблице 11 приложения, при определенном уровне значимости и данном объеме выборки. Если число элементов выборки больше 100, то величина критических значений вычисляется по формуле:

$$d_{кр} = \begin{cases} 1{,}36/\sqrt{n} \\ 1{,}63/\sqrt{n} \end{cases}.$$

Эмпирическое значение критерия для сопоставления эмпирического распределения с другим эмпирическим распределением определяется по формуле:

$$\lambda_{эмп} = d_{\max} \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}},$$

где n_1-количество наблюдений в первой выборке,

n_2- количество наблюдений во второй выборке,

d_{max}- наибольшая абсолютная величина разности между накопленными частостями по каждому разряду.

Уровень значимости соответствующий полученному значению λ определяется по таблице.

3. Критерий Фишера – φ

Критерий Фишера предназначен для сопоставления двух рядов выборочных значений по частоте встречаемости какого-либо признака. Этот критерий можно применять для оценки различий в любых двух выборках зависимых или независимых. С его помощью можно сравнивать показатели одной и той же выборки, измеренные в разных условиях.

Эмпирическое значение критерия подсчитываются по формуле:

$$\varphi_{эмп} = (\varphi_1 - \varphi_2) \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}},$$

где φ_1 - величина, определяемая по таблице, соответствующая большей процентной доле,

φ_2 - величина, определяемая по таблице, соответствующая меньшей процентной доле.

Уровень значимости соответствующий полученному значению φ определяется

по таблице.

Для применения критерия Фишера φ необходимо соблюдать следующие условия:

1) измерение может быть проведено по любой шкале;
2) характеристики выборок могут быть любыми;
3) нижняя граница – в одной из выборок может быть только два наблюдения, при этом во второй должно быть не менее 30 наблюдений. Верхняя граница не определена;
4) нижние границы двух выборок должны содержать не меньше 5 элементов (наблюдений) в каждой.

Лекция 8. Анализ таблиц сопряженности.

В социологических исследованиях чаще приходится встречаться с ситуациями, когда необходимо сравнить показатели внутри одной выборки.

В этом случае проверяется гипотеза H_0: сравниваемые признаки не влияют друг на друга. В качестве критерия обычно выбирается χ^2 - *критерий Пирсона*.

Исходные данные в подобных ситуациях удобно представлять в виде таблицы сопряженности:

Разряды	Эмпирические частоты					
	первое распределение	второе распределение	…	j-ое распределение	…	с-ое распределение
1	n_{11}	n_{12}	…	n_{1j}	…	n_{1c}
2	n_{21}	n_{22}	…	n_{2j}	…	n_{2c}
…	…	…	…	…	…	…
i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{ic}
…	…	…	…	…	…	…
k	n_{k1}	n_{r2}	…	n_{kj}	…	n_{kc}

Где n_{ij} – эмпирическая частота, соответствующая *i*-ому разряду *j*-ого распределения.

Для каждой ячейки таблицы, соответствующая теоретическая частота рассчитывается по формуле:

$$f_{mij} = \frac{\left(\sum_{j=1}^{c} n_{ij}\right) \cdot \left(\sum_{i=1}^{k} n_{ij}\right)}{\sum_{i=1}^{k}\sum_{j=1}^{c} n_{ij}}$$

или $$f_{mij} = \frac{(\text{сумма частот по строке})(\text{сумма частот по столбцу})}{\text{общее количество наблюдений}}.$$

Суть критерия χ^2 в этом случае не меняется. Наблюдаемое значение критерия рассчитывается по формуле:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{c} \frac{(n_{ij} - f_{mij})^2}{f_{mij}},$$

где n_{ij} – эмпирическая частота;

f_{mij} – теоретическая частота;

κ – количество строк таблицы сопряженности;

с – количество столбцов таблицы сопряженности.

Критическое значение χ^2 определяется по таблице критических точек в соответствии с выбранным уровнем значимости, и числом степеней свободы $v=(k-1)(c-1)$.

Лекция 9. Корреляционный анализ.

Связь, когда каждому значению независимой переменной (аргумента x) ставится определенное значение зависимой переменной (функции y) называется функциональной.

Функция может быть однозначной и многозначной. В случае однозначной функции, каждой паре значений x и y соответствует некоторая точка плоскости. Множество этих точек на плоскости представляет собой графическое изображение функциональной связи, или график функции y=f(x).

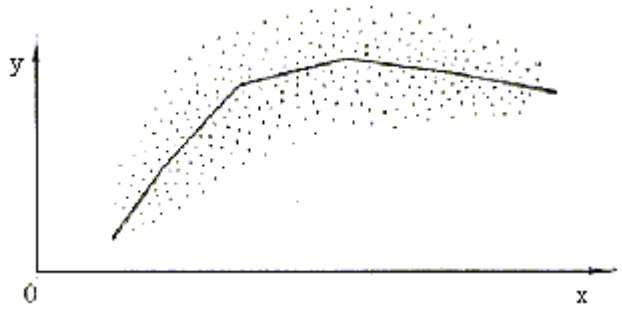
Однако в природе существуют не только функциональные связи такого рода.

Случается, что с изменением одной переменной происходит изменение распределения другой переменной. Связь этих переменных называется статистической. Если оказывается, что с изменением одной переменной изменяется среднее значение другой, то говорят, что между этими переменными существует корреляционная связь.

Изучением приемов, с помощью которых исследуются и обобщаются корреляционные связи, занимается корреляционный анализ. А количественное описание связи осуществляется на основе регрессионного анализа.

Например, требуется определить зависимость между ростом жены и мужа. Для

примера рассмотрим 100 супружеских пар. На плоскости дана прямоугольная система координат, по оси x откладывается рост мужа, по оси y – рост жены. Точкой на плоскости отмечается каждая супружеская пара. Полученное графическое изображение называется корреляционным полем.



В нашем случае должно быть 100 точек, которые как-то заполняют плоскость этого корреляционного поля. Разобьем все значения переменной x на несколько интервалов. Для каждого класс-интервала x отбираем все соответствующие ему точки. Находим их среднее значение $\bar{y}$. Эту точку наносим на график, обозначая ее крестиком, чтобы выделить среди прочих. Соединяем ломаной все отмеченные крестиком точки. Полученная линия показывает изменение среднего значения роста жены с изменением роста мужа от одного класс-интервала к другому. Эта линия называется эмпирической линией регрессии.

Если рассмотреть 100 других пар, то получится несколько иная эмпирическая линия регрессии. Если уменьшить величину класс-интервала, то линия покажет увеличение числа звеньев, сохранив в целом контур. Можно убедиться, что все эмпирические линии регрессии каких-либо двух переменных всегда лежат около некоторой плавной линии, называемой теоретической линией регрессии, или просто линией регрессии. Ее уравнение называется уравнением регрессии. Если мы рассматриваем изменение среднего y от x, то получится уравнение регрессии y на x: $\bar{y}_x = \varphi(x)$.

Если рассматриваем изменение среднего $\bar{x}$ от y, то уравнение регрессии x на y: $\bar{x}_y = \varphi(y)$.

При $\varphi(x) = ax + b$ говорят о линейной регрессии y на x, т.е. $\bar{y}_x = ax + b$. Аналогично можно ввести уравнение регрессии x на y: $\bar{x}_y = cy + d$.

Коэффициенты теоретической линии регрессии находят по методу наименьших квадратов: ищут эту линию при том условии, чтобы сумма квадратов расстояний эм-

пирической линии регрессии от теоретической была бы минимальной. Иначе говоря, теоретическая линия регрессии должна иметь наиближайшее расположение ко всем точкам эмпирической линии регрессии.

Если мы обозначили ординату теоретической линии регрессии y_m, эмпирической – $y_э$, то надо найти минимум величины:

$$f = \sum (y_э - y_m)^2 = \sum (y_э - a - bx)^2.$$

Это означает, что

$$\begin{cases} \dfrac{\partial f}{\partial a} = 2\sum (y - a - bx)(-1) = 0 \\ \dfrac{\partial f}{\partial b} = 2\sum (y - a - bx)(-x) = 0 \end{cases}$$

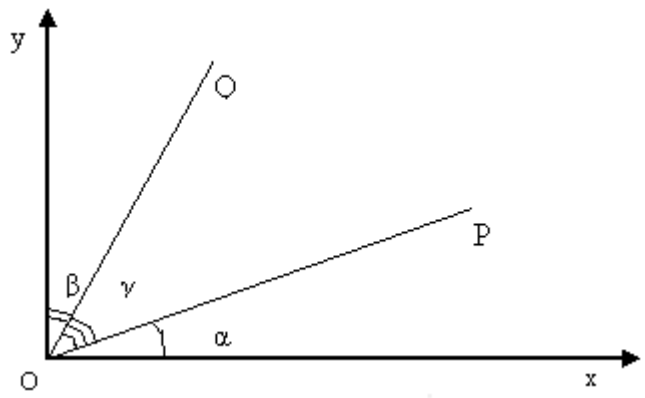
Получаем нормальные уравнения для определения коэффициентов линии регрессии:

$$\begin{cases} \sum [y - (a + bx)] = 0; \\ \sum [y - (a + bx)]x = 0. \end{cases} \text{или} \begin{cases} \sum y = na + b\sum x; \\ \sum xy = a\sum x + b\sum x^2. \end{cases} \text{или} \begin{cases} \bar{y} = a + b\bar{x} \\ \overline{xy} = a\bar{x} + b\overline{(x^2)} \end{cases}$$

Найденный коэффициент b называется коэффициентом регрессии y на x и обозначается ρ_{yx}.Аналогично определяется коэффициент регрессии x на y и обозначается ρ_{xy}.

Коэффициент корреляции является мерой наличия связи между переменными.

Дадим геометрическую интерпретацию коэффициенту корреляции (рис. 4).



OP – это линия регрессии y на x;

OQ – линия регрессии x на y; $\rho_{yx} = tg\alpha$; $\rho_{xy} = tg\beta$ $r_{xy} = \sqrt{tg\alpha\, tg\beta}$.

Если корреляции нет, то или линия OP, или OQ или обе вместе совпадают с осями координат, так как: α =0 или β =0 и, следовательно, r = 0.

Если корреляционная связь переходит в функциональную, то обе линии регрессии совпадают. Тогда $\alpha + \beta = 90^\circ$, $r = \sqrt{tg\alpha\, tg\beta} = \sqrt{tg\alpha\, tg\left(90^\circ - \alpha\right)} = \sqrt{tg\alpha\, ctg\alpha} = 1$, т.е. коэффициент корреляции равен 1.

Чем теснее связь между переменными, тем меньше угол между обеими линиями регрессии.

Рассмотренный коэффициент корреляции измеряет линейную связь между двумя количественными переменными. Этим, однако, не исчерпывается все возможное многообразие связей в социологии.

Во-первых, переменные могут иметь криволинейную регрессию: линия регрессии может быть параболой, кубической параболой, экспонентой и т.п. В каждом случае надо находить пути измерения связи между данными переменными.

Во-вторых, возможно наличие связи между более чем двумя переменными. Это проблема множественной корреляции, или многофакторного корреляционного анализа.

В-третьих, возможно существование связи между не только количественными переменными. В этом случае в статистике и социологии используются специальные показатели связи.

Лекция 10. Виды коэффициентов корреляции.

Задача корреляционного анализа сводится к установлению направления и формы между варьирующими признаками, измерению тесноты, и, наконец, к проверке значимости коэффициентов корреляции.

Переменные *x* и *y* могут быть измерены в разных шкалах. Это обстоятельство определяет выбор соответствующего коэффициента корреляции.

Тип шкалы		Мера связи
Переменная *x*	Переменная *y*	
Интервальная или отношений	Интервальная или отношений	Коэффициент Пирсона r_{xy}
Ранговая, интервальная или отношений	Ранговая, интервальная или отношений	Коэффициент Спирмена g_{xy}
Ранговая	Ранговая	Коэффициент τ Кендалла
Дихотомическая	Дихотомическая	Коэффициент ассоциации φ
Дихотомическая	Ранговая	Рангово-бисериальный $R_{гв}$
Дихотомическая	Интервальная или отношений	Бисериальный $R_{бис}$

Все коэффициенты по абсолютной величине не могут превосходить 1.

а) Коэффициент корреляции Пирсона вычисляется по формуле:

$$r_{xy} = \frac{\sum_i (x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 * \sum_i (y_i - \bar{y})^2}} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\left(\overline{x^2} - \left(\bar{x}\right)^2\right)} \cdot \sqrt{\left(\overline{y^2} - \left(\bar{y}\right)^2\right)}}$$

где x_i – значения, переменных принимаемые переменной x,

y_i - значения, переменных принимаемые переменной y,

$\bar{x}$ – средняя по x,

$\bar{y}$ - средняя по y.

Оценка значимости осуществляется при числе степеней свободы $k=n-2$.

б) Коэффициент корреляции рангов Спирмена вычисляется по формуле:

$$\rho = 1 - \frac{6 \cdot \sum_i (d_i)^2}{n \cdot (n^2 - 1)}$$

где n – количество ранжируемых признаков

d_i – разность между рангами по двум переменным для каждого испытуемого.

При наличии одинаковых рангов в числитель добавляются поправки на одинаковые ранги:

$$D_1 = \frac{n^3 - n}{12} \qquad D_2 = \frac{k^3 - k}{12},$$

где n – число одинаковых рангов в первом столбце,

k – число одинаковых рангов во втором столбце.

По каждой группе одинаковых рангов вводится своя поправка.

Критически значения определяются при уровне значимости равном числу значений признака, по таблице критических значений ρ Спирмена.

в) Коэффициент ассоциации φ вычисляется по формуле:

$$\varphi = \frac{p_{xy} - p_x \cdot p_y}{\sqrt{p_x \cdot (1 - p_x) \cdot p_y \cdot (1 - p_y)}},$$

где p_x – частота или доля признака, имеющего 1 по x,

$(1-p_x)$ - частота или доля признака, имеющего 0 по x,

p_y - частота или доля признака, имеющего 1 по y,

$(1-p_y)$ - частота или доля признака, имеющего 0 по y,

p_{xy} - частота или доля признака, имеющих 1 и по x и по y.

г) Коэффициент корреляции τ *Кендалла* вычисляется по формуле:

$$\tau = 1 - \frac{4 \cdot Q}{N \cdot (N - 1)},$$

где Q – число инверсий (подсчет инверсий осуществляется суммированием числа рангов второго признака меньше каждого из рангов второго признака, при условии, что ранги первого признака упорядочены по возрастанию).

д) Бисериальный коэффициент корреляции вычисляется по формуле:

$$R_{бис\ эмп} = \frac{\bar{x}_1 - \bar{x}_0}{\sigma_y} * \sqrt{\frac{n_1 * n_0}{N * (N - 1)}},$$

где $\bar{x}_1$ - среднее по тем элементам переменной *y*, которым соответствует признак 1 в переменной *x*,

$\bar{x}_0$ - среднее по тем элементам переменной *y*, которым соответствует признак 0 в переменной *x*,

n_1 – число единиц в переменной *x*,

n_0 – число нулей в переменной *x*,

$N = n_1 + n_0$,

σ_y - среднее квадратическое отклонение переменной *y*.

е) Рангово-бисернальный коэффициент корреляции вычисляется по формуле:

$$R_{rb\ эмп} = \frac{(\bar{x}_1 - \bar{x}_0)^2}{N},$$

где $\bar{x}_1$ - средний ранг по тем элементам переменной *y*, которым соответствует признак 1 в переменной *x*;

$\bar{x}_0$ - средний ранг по тем элементам переменной *y*, которым соответствует признак 0 в переменной *x*;

N – количество элементов в переменной *x*.

Для измерения нелинейной корреляционной связи между признаками используется *корреляционное отношение*.

Показатели корреляционного отношения вычисляется по формулам:

$$h_{yx} = \sqrt{\frac{\sum f_x (\bar{y}_x - \bar{y})^2}{\sum f_y (y_i - \bar{y})^2}},$$

$$h_{xy} = \sqrt{\frac{\sum f_y(\bar{x}_y - \bar{x})^2}{\sum f_x(x_i - \bar{x})^2}},$$

где $\bar{x}$ и $\bar{y}$ общие, а $\bar{x}_y$ и $\bar{y}_x$ - групповые средние арифметические, f_y и f_x частоты рядов *X* и *У*.

Для вычисления корреляционного отношения h_{xy} (*У* по *X*) или h_{xy} (*X* по *У*) необходимо:

1) расположить по порядку исходные данные по *X* от меньшей к большей, при этом сохранив значения соответствующих величин *У* по отношению к *X*.
2) Определить частоты переменной $X(f_x)$.
3) Подсчитать арифметическое средние по переменной *У* для соответствующей частоты $(\bar{y}_x)$;
4) Расположить по порядку исходные данные по *У* от меньшей величине к большей, при этом сохранив значения соответствующих величин *X* по отношению к *У*;
5) Определить частоты переменной $У(f_y)$;
6) Подсчитать арифметические средние по переменной *X* для соответствующей частоты $(\bar{x}_y)$;
7) Определить $\bar{x}$ и $\bar{y}$;
8) Произвести расчет по формулам;
9) Определить значимость полученных показателей.

Расчет уровней значимости коэффициентов корреляции.

При оценке значимости коэффициентов корреляции рассматриваются гипотезы:

H_0: коэффициент корреляции между признаками статистически значимо не отличается от нуля;

H_1: коэффициент корреляции между признаками статистически значимо отличается от нуля.

Все коэффициенты корреляции, не имеющие стандартных таблиц для нахождения критических значений, оценивают с помощью t – критерия Стьюдента по формуле:

$$t_{эмп} = |r_{эмп}| * \sqrt{\frac{n - 2}{1 - r^2_{эмп}}},$$

где $r_{эмп}$ – соответствующий коэффициент корреляции (корреляционное отношение),

n – число коррелируемых значений.

Критические значения $t_{кр}$ определяется по таблице значений для t – критерия Стьюдента. Число степеней свободы равно $k=n-2$.

Лекция 11. Многомерный анализ для исследования социальных процессов. Классификация многомерных методов.

Социально-экономические процессы и явления зависят от большого числа параметров, их характеризующих, что обусловливает трудности, связанные с выявлением структуры взаимосвязей этих параметров. В подобных ситуациях, т. е. когда решения принимаются на основании анализа стохастической, неполной информации, использование методов многомерного статистического анализа является не только оправданным, но и существенно необходимым.

Многомерные статистические методы среди множества возможных вероятностно-статистических моделей позволяют обоснованно выбрать ту, которая наилучшим образом соответствует исходным статистическим данным, характеризующий реальное поведение исследуемой совокупности объектов, оценить надежность и точность выводов, сделанных на основании ограниченного статистического материала.

К области приложения многомерных статистических методов могут быть отнесены задачи, связанные с исследованием поведения индивидуума, семьи или другой социально-экономической или производственной единицы, как представителя большой совокупности объектов.

Выделяют три центральные задачи, решаемые с помощью многомерных методов.

1. Статистическое исследование структуры и характера взаимосвязей, существующих между анализируемыми количественными переменными. При этом под переменными понимаются как регистрируемые на объектах признаки, так и время *t.*

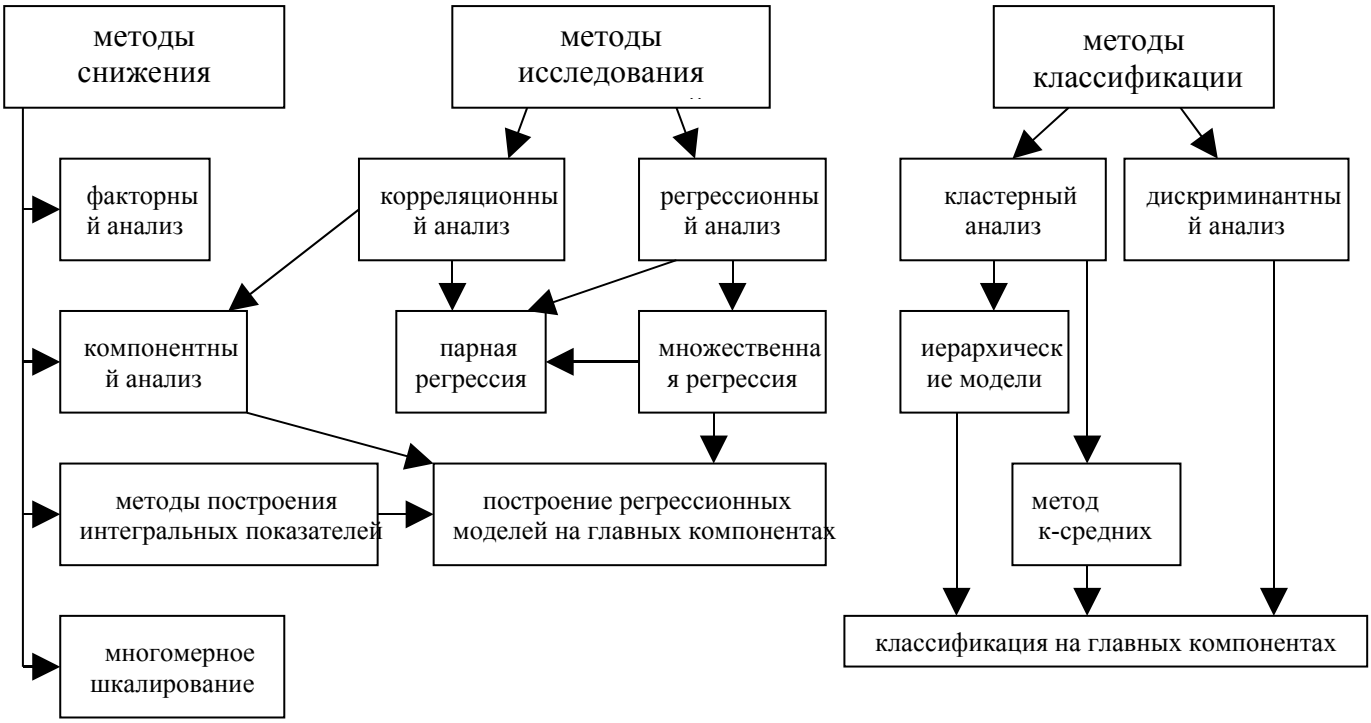
2. Разработка статистических методов классификации объектов и признаков.

3. Снижение размерности исследуемого признакового пространства с целью лаконичного объяснения природы анализируемых многомерных данных. Возможность лаконичного описания анализируемых многомерных данных основана на априорном допущении, в соответствии с которым существует небольшое число признаков с помощью которых могут быть достаточно точно описаны как сами наблюдаемые переменные анализируемых объектов, так и определяемые этими переменными свойства (характеристики) анализируемой совокупности. При этом упомянутые признаки могут находиться среди исходных признаков, а могут быть латентными, т.е. непосредственно статистически не наблюдаемыми, но восстанавливаемыми по исходным данным.

Эти задачи не исчерпывают всех возможностей многомерных статистических методов, но в настоящий момент являются наиболее распространенными.

В соответствии с задачами в структуре многомерных статистических методов выделяют методы снижения размерности, методы исследования зависимостей, методы классификации.

Классификацию многомерных методов представим на схеме:



Работая с многомерными статистическими методами важно, чтобы переменные изменялись в сравнимых шкалах. Из неоднородности единиц измерения вытекает невозможность обоснованного выражения значений различных показателей в одном

масштабе. Чтобы устранить неоднородность измерения исходных данных, все их значения предварительно нормируются, т.е. выражаются через отношение этих значений к некоторой величине, отражающей определенные свойства данного показателя. Нормирование исходных данных иногда проводится посредством деления исходных величин на среднеквадратичное отклонение соответствующих показателей. Другой способ сводится к вычислению, так называемого, стандартизованного вклада. Его еще называют *Z*-вкладом.

Z - вклад показывает, сколько стандартных отклонений отделяет данное наблюдение от среднего значения:

$Z_i = \frac{x_i - \bar{x}}{\sigma_i}$, где x_i - значение данного наблюдения, $\bar{x}$ - среднее, σ_i -стандартное отклонение.

Среднее для *Z*-вкладов является нулевым и стандартное отклонение равно 1.

Стандартизация позволяет сравнивать наблюдения из различных распределений.

Заметим, что методы нормирования означают признание всех признаков равноценными с точки зрения выяснения сходства рассматриваемых объектов. Признание равноценности различных показателей кажется оправданным отнюдь не всегда. Было бы желательным наряду с нормированием придать каждому из показателей вес, отражающий его значимость в ходе установления сходств и различий объектов.

В этой ситуации приходится прибегать к способу определения весов отдельных показателей – опросу экспертов.

Экспертные оценки дают известное основание для определения важности индикаторов, входящих в ту или иную группу показателей.

Довольно часто при решении подобных задач используют не один, а два расчета: первый, в котором все признаки считаются равнозначными, второй, где им придаются различные веса в соответствии со средними значениями экспертных оценок.

Лекция 12. Постановка задачи кластерного анализа.

Классификация является основой умозрительной человеческой деятельности и фундаментальным процессом научной практики. В ходе исследований, развития

науки и техники накоплено значительное количество материалов, которые необходимо систематизировать с целью выявления законов общественного развития, изучения эволюций и совершенствования технологий. Эта работа требует от исследователя детального изучения данных и их обобщения, в ходе которого отдельные факты складываются в закономерности, а закономерности в теории. Общий вопрос, задаваемый исследователями во многих областях, состоит в том, как организовать наблюдаемые данные в наглядные структуры. В настоящее время существует множество подходов к классификации объектов.

Среди них кластерный анализ - наиболее действенный количественный инструмент исследования социально - экономических процессов, описываемых большим числом характеристик.

Кластерный анализ наиболее ярко отражает черты многомерного анализа в классификации. Термин кластерный анализ включает в себя набор различных алгоритмов классификации.

Первое применение кластерный анализ нашел в социологии. Название кластерный анализ происходит от английского слова cluster - гроздь, скопление. Впервые в 1939 был определен предмет кластерного анализа и сделано его описание исследователем Трионом. Главное назначение кластерного анализа - разбиение множества исследуемых объектов и признаков на однородные в соответствующем понимании группы или кластеры. Это означает, что решается задача классификации данных и выявления соответствующей структуры в ней.

Методы кластерного анализа позволяют решать следующие задачи:

1. Проведение классификации объектов с учетом признаков, отражающих сущность, природу объектов. Решение такой задачи, как правило, приводит к углублению знаний о совокупности классифицируемых объектов;

2. Проверка выдвигаемых предположений о наличии некоторой структуры в изучаемой совокупности объектов, т. е. поиск существующей структуры;

3. Построение новых классификаций для слабоизученных явлений, когда

необходимо установить наличие связей внутри совокупности и попытаться привнести в нее структуру.

Большое достоинство кластерного анализа в том, что он позволяет производить

разбиение объектов не по одному параметру, а по целому набору признаков. Кроме того, кластерный анализ в отличие от большинства математико-статистических методов не накладывает никаких ограничений на вид рассматриваемых объектов, и позволяет рассматривать множество исходных данных практически произвольной природы. Это имеет большое значение, например, для прогнозирования конъюнктуры, когда показатели имеют разнообразный вид, затрудняющий применение традиционных эконометрических подходов.

Кластерный анализ позволяет рассматривать достаточно большой объем информации и резко сокращать, сжимать большие массивы информации, делать их компактными и наглядными.

Важное значение кластерный анализ имеет применительно к совокупностям временных рядов, характеризующих экономическое развитие. Здесь можно выделять периоды, когда значения соответствующих показателей были достаточно близкими, а также определять группы временных рядов, динамика которых наиболее схожа.

Кластерный анализ можно использовать циклически. В этом случае исследование производится до тех пор, пока не будут достигнуты необходимые результаты. При этом каждый цикл может давать информацию, которая способна сильно изменить направленность и подходы дальнейшего применения кластерного анализа. Этот процесс можно представить системой с обратной связью.

В задачах социально-экономического прогнозирования весьма перспективно сочетание кластерного анализа с другими количественными методами (например, с регрессионным анализом).

Как и любой другой метод, кластерный анализ имеет определенные недостатки и ограничения. В частности, состав и количество кластеров зависит от выбираемых критериев разбиения. При сведении исходного массива данных к более компактному виду могут возникать определенные искажения, а также могут теряться индивидуальные черты отдельных объектов за счет замены их характеристиками обобщенных значений параметров кластера. При проведении классификации объектов игнорируется очень часто возможность отсутствия в рассматриваемой совокупности каких-либо значений кластеров.

Поэтому необходимо сделать несколько предостережений общего характера.

1) Многие методы кластерного анализа - довольно простые эвристические процедуры, которые, как правило, не имеют достаточного статистического обоснования.

2) Разные кластерные методы могут порождать и порождают различные решения для одних и тех же данных. Это обычное явление в большинстве прикладных исследований.

3) Цель кластерного анализа заключается в поиске существующих структур. В то же время его действие состоит в привнесении структуры в анализируемые данные, т. е. методы кластеризации могут приводить к порождению артефактов.

Исследования, использующие кластерный анализ, характеризуют следующие пять основных шагов: 1) отбор выборки для кластеризации; 2) определение множества признаков, по которым будут оцениваться объекты в выборке, и способа их стандартизации; 3) вычисление значений той или иной меры сходства между объектами; 4) применение метода кластерного анализа для создания групп сходных объектов; 5) проверка достоверности результатов кластерного решения.

В кластерном анализе считается, что:

а) выбранные характеристики допускают в принципе желательное разбиение на кластеры;

б) единицы измерения (масштаб) выбраны правильно.

Задача кластерного анализа заключается в том, чтобы на основании данных, содержащихся во множестве *X,* разбить множество объектов *G* на *m* (*m* - целое) кластеров (подмножеств) $Q_1, Q_2, ..., Q_m$, так, чтобы каждый объект G_j принадлежал одному и только одному подмножеству разбиения и чтобы объекты, принадлежащие одному и тому же кластеру, были сходными, в то время, как объекты, принадлежащие разным кластерам были разнородными.

Лекция 13. Иерархические кластерные структуры.

Решением задачи кластерного анализа являются разбиения, удовлетворяющие некоторому критерию оптимальности. Этот критерий может представлять собой некоторый функционал, выражающий уровни желательности различных разбиений и

группировок, который называют целевой функцией. Например, в качестве целевой функции может быть взята внутригрупповая сумма квадратов отклонений:

$$W = \sum_{j=1}^{n} (x_j - \bar{x})^2 = \sum_{j=1}^{n} x_j^2 - \frac{1}{n}(\sum_{j=1}^{n} x_j)^2$$

где x_j - представляет собой измерения j - го объекта.

Наиболее трудным в задаче классификации является определение меры однородности объектов.

Понятно, что объекты *i*-ый и *j*-ый попадали бы в один кластер, когда расстояние (отдаленность) между точками X_i и X_j было бы достаточно маленьким и попадали бы в разные кластеры, когда это расстояние было бы достаточно большим. Таким образом, попадание в один или разные кластеры объектов определяется понятием расстояния между X_i и X_j из E_p, где E_p p - мерное евклидово пространство.

Неотрицательная функция $\rho(X_i, X_j)$ называется функцией расстояния (метрикой), если:

а) $\rho(X_i, X_j) \geq 0$, для всех X_i и X_j из E_p

б) $\rho(X_i, X_j) = 0$, тогда и только тогда, когда $X_i = X_j$

в) $\rho(X_i, X_j) = \rho(X_j, X_i)$

г) $\rho(X_i, X_j) \leq \rho(X_i, X_k) + \rho(X_{k,} X_j)$, где X_i; X_j и X_k - любые три вектора из E_p.

Значение $\rho(X_i, X_j)$ для X_i и X_j называется расстоянием между X_i и X_j и эквивалентно расстоянию между G_i и G_j соответственно выбранным характеристикам $(F_1, F_2, F_3, \ldots, F_p)$.

Наиболее часто употребляются следующие функции расстояний:

$\rho = \sqrt{\sum (x - y)^2}$ – евклидово расстояние, наиболее общий тип расстояния. Оно является геометрическим расстоянием в многомерном пространстве;

$\rho = \sum (x - y)^2$ – квадрат евклидова расстояния используется для того, чтобы придать большие веса более отдаленным друг от друга объектам;

$\rho = \sum |x - y|$ – расстояние городских кварталов (манхэттенское расстояние) для этой меры влияние отдельных больших разностей (выбросов) уменьшается;

$\rho = \max(x - y)$ – расстояние Чебышева полезно, когда желают определить два объекта как «различные», если они различаются по какой-либо одной координате (каким-либо одним измерением);

$\rho = \left(\sum |x - y|^p\right)^{\frac{1}{k}}$ степенное расстояние используют, когда хотят увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются. Параметры r и p определяются пользователем Параметр p ответственен за постепенное взвешивание разностей по отдельным координатам, параметр r ответственен за прогрессивное взвешивание больших расстояний между объектами;

$\rho = \frac{количество\ x_i \neq y_i}{i}$ – процент несогласия используется в тех случаях, когда данные являются категориальными.

Пусть n измерений $X_1, X_2, ..., X_n$ представлены в виде матрицы данных размером $p \times n$

$$X = \begin{pmatrix} x_{11} & x_{12} & ...x_{1n} \\ x_{21} & x_{22} & ...x_{2n} \\ x_{p1} & x_{p2} & ...x_{pn} \end{pmatrix} = \left(X_1, X_2, ..., X_n\right)$$

Тогда расстояние между парами векторов $\rho\left(X_i, X_j\right)$ могут быть представлены в виде симметричной матрицы расстояний:

$$D = \begin{pmatrix} 0 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 0 & \cdots & \rho_{n} \\ \rho_{n1} & \rho_{n2} & \cdots & 0 \end{pmatrix}$$

Понятием, противоположным расстоянию, является понятие сходства между объектами G_l и G_j. Неотрицательная вещественная функция $S\left(X_i; X_j\right) = s_{ij}$ называется мерой сходства, если:

1) $0 \leq S\left(X_i, X_j\right) < 1$ для $X_l \neq X_j$

2) $S(X_i, X_i) = 1$

3) $S(X_i, X_j) = S(X_j, X_i)$

Пары значений мер сходства можно объединить в матрицу сходства:

$$S = \begin{pmatrix} 1 & s_{12} & \dots & s_{1n} \\ s_{21} & 1 & \dots & s_{2n} \\ s_{n1} & s_{n2} & \dots & 1 \end{pmatrix}$$

Величину s_{ij} называют коэффициентом сходства.

Естественной мерой сходства характеристик объектов во многих задачах является коэффициент корреляции между ними

$r_{ij} = \dfrac{\sum_{h=1}^{N} (x_{hi} - m_i)(x_{hj} - m_j)}{\sigma_i \cdot \sigma_j}$, где $m_i, m_j, \sigma_i, \sigma_j$ - соответственно средние и среднеквадратичные отклонения для характеристик *i* и *j*. Мерой различия между характеристиками может служить величина *1-r*.

На первом шаге, когда каждый объект представляет собой отдельный кластер, расстояния между этими объектами определяются выбранной мерой. Однако когда связываются вместе несколько объектов необходимо правило объединения или связи для двух кластеров. Существует множество методов объединения кластеров, перечислим наиболее распространенные:

Одиночная связь (метод ближайшего соседа) – расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами.

Полная связь (метод наиболее удаленных соседей) – расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах.

Невзвешенное попарное среднее – расстояние между двумя различными кластерами вычисляется как среднее расстояние между всеми парами объектов в них.

Взвешенное попарное среднее – идентично методу невзвешенного попарного среднего, за исключением того, что при вычислениях размер соответствующих кластеров (т.е. число объектов, содержащихся в них) используется в качестве весового коэффициента.

Невзвешенный центроидный метод – расстояние между двумя кластерами определяется как расстояние между их центрами тяжести.

Взвешенный центроидный метод (медиана) – идентичен предыдущему, за исключением того, что при вычислениях используются веса для учёта разницы между размерами кластеров (т.е. числами объектов в них)

Метод Варда. – отличается от всех других методов, поскольку он использует методы дисперсионного анализа для оценки расстояний между кластерами. Метод минимизирует сумму квадратов (SS) для любых двух кластеров, которые могут быть сформированы на каждом шаге. В целом метод представляется очень эффективным, однако он стремится создавать кластеры малого размера.

Число алгоритмов методов кластерного анализа слишком велико. Все их можно подразделить на иерархические и неиерархические.

Иерархические алгоритмы связаны с построением дендограмм и делятся на:

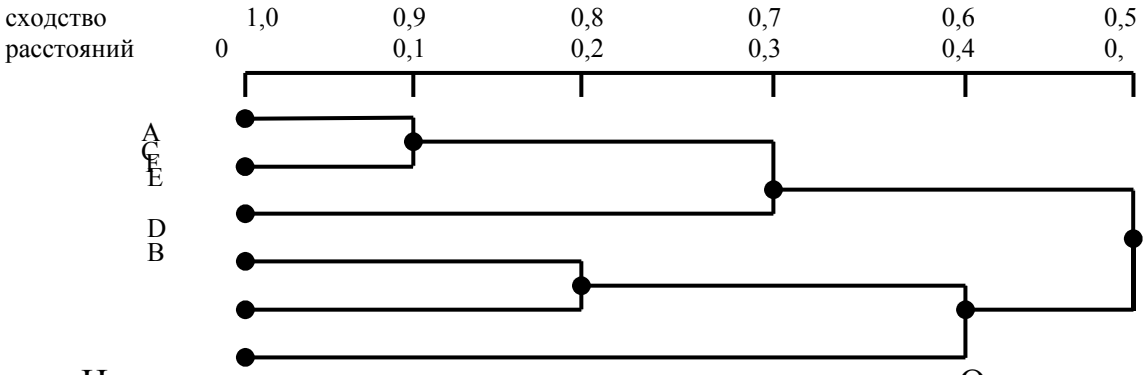
а) агломеративные, характеризуемые последовательным объединением исходных элементов и соответствующим уменьшением числа кластеров;

б) дивизимные (делимые), в которых число кластеров возрастает, начиная с одного, в результате чего образуется последовательность расщепляющих групп.

Иерархические агломеративные методы – многошаговые методы, работающие в такой последовательности: на нулевом шаге за разбиение принимается исходная совокупность n элементарных кластеров, матрица расстояний между которыми $\{\rho_{ij}\}_{n\times n} = \{l_{ij}\}_{n\times n}$; на каждом следующем шаге происходит объединение двух кластеров K_s и K_t, сформированных на предыдущем шаге, в один кластер $K_s \cup K_t$ (будем его обозначать $K_{s\oplus t}$, при этом размерность матрицы расстояний уменьшается, по сравнению с размерностью матрицы предыдущего шага, на единицу.

Наиболее известный метод представления матрицы расстояний или сходства основан на идее дендрограммы или диаграммы дерева. Дендрограмму можно определить как графическое изображение результатов процесса последовательной кластеризации, которая осуществляется в терминах матрицы расстояний. С помощью дендрограммы можно графически или геометрически изобразить процедуру кластеризации при условии, что эта процедура оперирует только с элементами матрицы расстояний или сходства.

Существует много способов построения дендрограмм. В дендрограмме объекты располагаются вертикально слева, результаты кластеризации - справа. Значения расстояний или сходства, отвечающие строению новых кластеров, изображаются по горизонтальной прямой поверх дендрограмм.



На рисунке показан один из примеров дендрограммы. Он соответствует случаю шести объектов *(n=6)* и *k* характеристик (признаков). Объекты *A* и *C* наиболее близки и поэтому объединяются в один кластер на уровне близости, равном 0,9. Объекты *D* и *E* объединяются при уровне 0.8. Теперь имеем 4 кластера:

$(A \oplus C), (F), (D \oplus E), (B)$.

Далее образуются кластеры *((A⊕ C)⊕ F)* и *((E⊕ D), B),* соответствующие уровню близости, равному 0,7 и 0,6. Окончательно все объекты группируются в один кластер при уровне 0,5.

Вид дендрограммы зависит от выбора меры сходства или расстояния между объектом и кластером и метода кластеризации.

Лекция 14. Метод к-средних.

Иерархические методы используются обычно в задачах классификации небольшого числа объектов (порядка нескольких десятков), где больший интерес представляет не число кластеров, а анализ структуры множества этих объектов и наглядная интерпретация проведенного анализа в виде дендрограммы. Если же число кластеров заранее задано или подлежит определению, то для классификации чаще всего используют *параллельные* кластер-процедуры – это итерационные алгоритмы,

на каждом шаге которых *используется* одновременно (параллельно) все наблюдения. Так как эти алгоритмы на каждом шаге работают со всеми наблюдениями, то основной целью их конструирования является нахождение способов сокращения числа перебора вариантов (даже при числе наблюдений порядка нескольких десятков), что приводит зачастую лишь к приближенному, но не слишком трудоемкому решению задач. В параллельных кластер-процедурах реализуется обычно идея оптимизации разбиения в соответствии с некоторым функционалом качества.

Наиболее распространенными являются при заданном числе k кластеров следующие функционалы качества разбиения:

- сумма внутрикластерных дисперсий $f(R) = \sum_{p=1}^{k} \sum_{i \in K_p} l^2\left(x_i, \bar{x}_{K_p}\right)$,
- сумма попарных внутрикластерных расстояний $f(R) = \sum_{p=1}^{k} \frac{1}{n_p} \sum_{i,j \in K_p} l^2\left(x_i, x_j\right)$,

а при неизвестном числе кластеров функционалы

$$f(R) = \alpha f_1(R) + \beta f_2(R) \quad \text{или} \quad f(R) = [f_1(R)]^{\alpha} [f_2(R)]^{\beta}$$

где $f_1(R)$ - некоторая не возрастающая функция числа классов, характеризующая средний внутриклассовый разброс наблюдений, $f_2(R)$ - некоторая неубывающая функция числа классов, характеризующая взаимную удаленность классов или меру «концентрации» наблюдений.

Схема работы алгоритмов, связанная с функционалами качества, такая: для некоторого начального разбиения R_0 вычисляют значение $f(R_0)$; затем каждую из точек x_i, поочередно перемещают во все кластеры и оставляют в том положении, которое соответствует наилучшему значению функционала качества. Работу заканчивают, когда перемещение точек не дает улучшения качества. Часто описанный алгоритм применяют несколько раз, начиная с разных начальных разбиений R_0, и выбирают наилучший вариант разбиения.

Очень важным вопросом является проблема выбора необходимого числа кластеров. Иногда можно число кластеров выбирать априорно. Однако в общем случае это число определяется в процессе разбиения множества на кластеры.

Проводились исследования Фортьером и Соломоном, и было установлено, что число кластеров должно быть принято для достижения вероятности *a* того, что найдено наилучшее разбиение. Таким образом, оптимальное число разбиений является функцией заданной доли β наилучших или в некотором смысле допустимых разбиений во множестве всех возможных. Общее рассеяние будет тем больше, чем выше доля β допустимых разбиений. Фортьер и Соломон разработали таблицу, по которой можно найти число необходимых разбиений $S(\alpha, \beta)$ в зависимости от α и β (где α - вероятность того, что найдено наилучшее разбиение, β - доля наилучших разбиений в общем числе разбиений) Причем в качестве меры разнородности используется не мера рассеяния, а мера принадлежности, введенная Хользенгером и Харманом. Таблица значении $S(\alpha, \beta)$ приводится ниже.

$\beta \setminus \alpha$	0.20	0.10	0.05	0.01	0.001	0.0001
0.20	8	11	14	21	31	42
0.10	16	22	29	44	66	88
0.05	32	45	59	90	135	180
0.01	161	230	299	459	689	918
0.001	1626	2326	3026	4652	6977	9303
0.0001	17475	25000	32526	55000	75000	100000

Довольно часто критерием объединения (числа кластеров) становится изменение соответствующей функции. Например, суммы квадратов отклонений:

$$E_j = \sum_{i=1}^{n} r_{ij}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} r_{ij}\right)^2$$

Процессу группировки должно соответствовать здесь последовательное минимальное возрастание значения критерия *E*. Наличие резкого скачка в значении *E* можно интерпретировать как характеристику числа кластеров, объективно существующих в исследуемой совокупности.

Итак, второй способ определения наилучшего числа кластеров сводится к выявлению скачков, определяемых фазовым переходом от сильно связанного к слабосвязанному состоянию объектов.

Иерархические и параллельные кластер-процедуры практически реализуемы лишь в задачах классификации не более нескольких десятков наблюдений. К решению задач с большим числом наблюдений применяют *последовательные* кластер-процедуры - это итерационные алгоритмы, на каждом шаге которых используется

одно наблюдение (или небольшая часть исходных наблюдений) и результаты разбиения на предыдущем шаге. Идею этих процедур поясним на представленном в ППП «STASTICA» *методе К-средних* («*K* – Means Clusyering») с заранее заданным числом *k* классов.

На нулевом шаге за центры искомых *k* кластеров принимают случайно выбранные *k* наблюдений – точки $x_1, x_2, \ldots x_k$; каждому кластеру присваивают единичный вес. На первом шаге находят расстояния точки x_{k+1} до центров кластеров и точку x_{k+1} относят к кластеру, расстояние до которого минимально; рассчитывают новый центр тяжести (как взвешенное среднее по каждому показателю) этого кластера и вес кластера увеличивают на единицу; все остальные кластеры остаются неизменными (с прежними центрами и весами). На втором шаге аналогичную процедуру выполняют для точки x_{k+2} и т. д. При достаточно большом числе *n* классифицируемых объектов или достаточно большом числе итерации пересчет центров тяжести практически не приводит к их изменению.

Если в какой-то точке не удается, прогнав все $x_{k+(n-1)}$ точек, достичь практически не изменяющихся центров тяжести, то либо используя получившееся разбиение *n* точек на *k* кластеров в качестве начального применяют изложенную процедуру к точкам x_1, x_2 и т. д.; либо в качестве начального разбиения принимают различные комбинации *k* точек из исходных *n* точек и в качестве окончательного берут наиболее часто встречающееся финальное разбиение.

Кластерный анализ методом к-средних дополняет и уточняет картину, полученную с помощью иерархического кластерного анализа. Однако конфигурация кластеров не поддается представлению в графическом виде.

Лекция 15. Факторный анализ, и его использование в исследовании связи.

Факторный анализ – статистический метод, который используется при обработке больших массивов экспериментальных данных. Задачами факторного анализа являются: сокращение числа переменных (редукция данных) и определение структуры взаимосвязей между переменными, т.е. классификация переменных, поэтому факторный анализ используется как метод сокращения данных или как метод структурной классификации.

В современной статистике под *факторным анализом* понимают совокупность методов, которые на основе реально существующих связей признаков (или объектов) позволяют выявлять латентные обобщающие характеристики организационной структуры и механизма развития изучаемых явлений и процессов.

Понятие латентности в определении ключевое. Оно означает неявность характеристик, раскрываемых при помощи методов факторного анализа. Вначале мы имеем дело с набором элементарных признаков *Xj,* их взаимодействие предполагает наличие определенных причин, особенных условий, т.е. существование которых скрытых факторов. Последние устанавливаются в результате обобщения элементарных признаков и выступают как интегрированные характеристики, или признаки, но более высокого уровня. Естественно, что коррелировать могут не только тривиальные признаки X_j, но и сами наблюдаемые объекты N_i, Поэтому поиск латентных факторов теоретически возможен как по признаковым, так и по объектным данным.

Идея метода состоит в сжатии матрицы признаков в матрицу с меньшим числом переменных, сохраняющую почти ту же самую информацию, что и исходная матрица, т.е. сконцентрировать исходную информацию, выражая большое число рассматриваемых признаков через меньшее число более емких внутренних характеристик явления, которые, однако, не поддаются непосредственному измерению.

Предположим, *n* наблюдаемых объектов (автомобилей) оценивается в двумерном признаковом пространстве R^2 с координатными осями: X_1 – стоимость автомобиля и X_2 – длительность рабочего ресурса мотора. При условии коррелированности X_1 и X_2 в системе координат появляется направленное и достаточно плотное скопление точек, формально отображаемое новыми осями (F_1 и F_2). Характерная особенность F_1 и F_2 заключается в том, что они проходят через плотные скопления точек и в свою очередь коррелируют с X_1 и X_2. Максимальное число новых осей F_r будет равно числу элементарных признаков.

Допуская линейную зависимость F_r *от* X_{jt} можем записать:

$F_1 = a_{11} x_1 + a_{21} x_2$ и $F_2 = a_{12} x + a_{22} x_2$.

Интерпретируем оси пусть F_1 – экономичность автомобиля, F_2 – его надежность в эксплуатации. Суждение об F_1, и F_2 базируется на оценке структуры латентных факторов, т.е. оценке весов X_1 и X_2 в F_r .

Если объекты характеризуются достаточно большим числом элементарных признаков *(m* > 3), то логично и другое предположение – о существовании плотных скоплений точек (признаков) в пространстве *n* объектов. При этом новые оси обобщают уже не признаки X_{jt} а объекты, соответственно и латентные факторы *F,* будут распознаны по составу наблюдаемых объектов.

Материалом для факторного анализа служат корреляционные связи, а точнее – коэффициенты корреляции Пирсона, которые вычисляются между переменными, включенными в обследование.

В зависимости от того, какой тип корреляционной связи – элементарных признаков или наблюдаемых объектов – исследуется в факторном анализе, различают *R* и *Q* – технические приемы обработки данных.

Название *R*-техники носит объемный анализ данных по *m* признакам, в результате него получают *r* линейных комбинаций (групп) признаков *(F_r=f(Xj);* r = l,n). Анализ по данным о близости (связи) *n* наблюдаемых объектов называется Q-техникой и позволяет определять *r* линейных комбинаций (групп) объектов:

В настоящее время на практике более 90% задач решается при помощи *R*-техники.

Толчком для развития методов факторного анализа изначально послужили задачи и проблемы из области психологии. Позже методы факторного анализа стали активно применяться в социологических исследованиях, медицине, затем в военной промышленности, экономике.

Методы факторного анализа целесообразно разделить на два класса: упрощенные и современные аппроксимирующие методы.

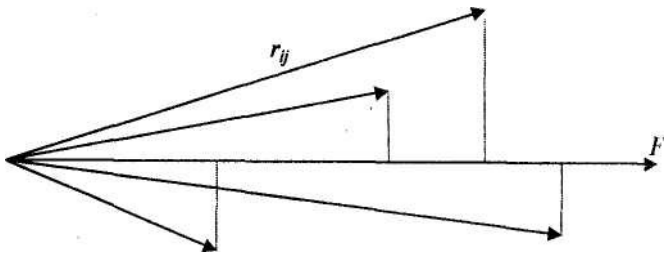
Простые методы факторного анализа в основном связаны с начальными теоретическими разработками. Они имеют ограниченные возможности в выделении латентных факторов и аппроксимации факторных решений. В числе этих методов следует назвать:

однофакторную модель Ч. Спирмена. Она позволяет выделить только один генеральный латентный и один характерный факторы. Для возможно существующих других латентных факторов делается предположение об их незначимости;

бифакторную модель Г. Хользингера. Допускает влияние на вариацию элемен-

тарных признаков не одного, а нескольких латентных факторов (обычно двух) и одного характерного фактора;

центроидный метод Л. Тэрстоуна. В нем корреляции между переменными рассматриваются как пучок векторов, а латентный фактор геометрически представляется как уравновешивающий вектор, проходящий через центр этого пучка Метод позволяет выделять несколько латентных и характерные факторы, впервые появляется возможность соотносить факторное решение с исходными данными, т.е. простейшем виде решать задачу аппроксимации.



Современные аппроксимирующие методы часто предполагают, что первое, приближенное решение уже найдено каким-либо из способов, последующими шагами это решение оптимизируется. Методы отличаются сложностью вычислений. К этим методам относятся:

- *групповой метод* Л. Гуттмана и П. Хорста. Решение базируется на предварительно отобранных каким-либо образомгруппах элементарных признаков;
- *метод главных факторов* Г. Томсона. Наиболее близок методу главных компонент, отличие заключается в предположении о существовании характерностей;
- *метод максимального правдоподобия* (Д. Лоули), *минимальных остатков* (Г. Харман), *а-факторного анализа* (Г. Кайзери И. Кэффри,), *канонического факторного анализа* (К. Рао),все *оптимизирующие.* Позволяют последовательно улучшить предварительно найденные решения на основе использования статистических приемов оценивания случайной величины или статистических критериев, предполагают большой объем трудоемких вычислений. Наиболее перспективным и удобным для работы в этой группе признается метод максимального правдоподобия.

Основной задачей, которую решают разнообразными методами факторного

анализа, является сжатие информации, переход от множества значений по *m* элементарным признакам с объемом информации $n \times m$ к ограниченному множеству элементов матрицы факторного отображения $(m \times r)$ или матрицы значений латентных факторов для каждого наблюдаемого объекта размерностью $n \times r$, причем обычно $r < m$.

Методы факторного анализа позволяют также визуализировать структуру изучаемых явлений и процессов, а это значит определять их состояние и прогнозировать развитие. Наконец, данные факторного анализа дают основания для идентификации объекта, т.е. решения задачи распознавания образа.

Методы факторного анализа обладают свойствами, весьма привлекательными для их использования в составе других статистических методов, наиболее часто в корреляционно-регрессионном анализе, кластерном анализе, многомерном шкалировании и др.

Лекция 16. Выделение латентных переменных, факторные нагрузки и факторные весы.

Главное понятие факторного анализа – фактор. Это искусственный статистический показатель, возникающий в результате специальных преобразований коэффициентов корреляции между изучаемыми признаками – латентная переменная.

Независимо от выбранного метода факторного анализа основные его результаты выражаются в наборах факторных нагрузок и факторных весов.

Факторные нагрузки - это значения коэффициентов корреляции каждого из исходных признаков с каждым из выявленных факторов. Чем теснее связь данного признака с рассматриваемым фактором, тем выше значение факторной нагрузки. Положительный знак факторной нагрузки указывает на прямую (а отрицательный знак - на обратную) связь данного признака с фактором. Таблица факторных нагрузок содержит *m* строк (по числу признаков) и k столбцов (по числу факторов).

Факторными весами называют количественные значения выделенных факторов для каждого из *n.* имеющихся объектов. Объекту с большим значением факторного веса присуща большая степень проявления свойств, определяемых данным фактором.

Поэтому положительные факторные веса соответствуют тем объектам, которые

обладают степенью проявления свойств больше средней, а отрицательные факторные веса соответствуют тем объектам, для которых степень проявления свойств меньше средней. Таблица факторных весов содержит n строк (по числу объектов) и k столбцов (по числу факторов).

Таким образом, данные о факторных нагрузках позволяют сформулировать выводы о наборе исходных признаков, отражающих тот или иной фактор, и об относительном весе отдельного признака в структуре каждого фактора. В свою очередь, данные о факторных весах определяют ранжировку объектов по каждому фактору.

В основе каждого метода факторного анализа лежит математическая модель, описывающая соотношения между исходными признаками и обобщенными латентными факторами.

Изучение факторных воздействий предполагает выявление взаимосвязей характерных признаков. Для многомерных объектов показателями связи являются оценки дисперсии и коэффициенты ковариации, которые обобщаются в матрице ковариаций (по выборочным данным – матрица *S)*. Когда исходные значения признаков нормированы, матрица ковариаций, переходит в матрицу парных корреляций *R*.

$$S = R = \frac{1}{n} Z^T Z$$

Симметрическая матрица *R* имеет собственную систему координат в пространстве R^m , где m – число анализируемых признаков. Допуская преобразования координатной системы в систему пространства латентных факторов, можно записать Z в виде линейной комбинации новых координат в матричной форме: Z = *AF*.

Воспользуемся возможностью подстановки в уравнение для *R* вместо Z произведения матриц *AF* и получим:

$$R = \frac{1}{n} AF(AF)^T = \frac{1}{n} AFF^T A^T$$

Изменив место расположения скаляра 1/n, выделим произведение $\frac{1}{n} FF^T$, результат произведения интерпретируется как матрица корреляций C, определяемая для латентных факторов F_r После замены 1/n *FF* на C запишем: $R = ACA^T$.

В предположении, что факторы F_r некоррелированы, т.е. $C = E$, где *E* – единичная матрица, приходим к равенству: $R = AA'$.

Л.Л. Тэрстоуном равенства типа: $R = ACA'$ и $R - AA'$ названы *фундаментальной*

факторной теоремой, A – здесь матрица факторного отображения, а ее элементы *a* – величины факторных нагрузок. Суть теоремы – в возможности воспроизведения исходной корреляционной матрицы *R* через матрицу факторного отображения *A*. При C = *E* связь матричных элементов г и *a* записывается в виде уравнения:

$$r_{ij} = a_{i1}a_{j1} + a_{i2}a_{j2} + \ldots + a_{ir}a_{jr}$$

Другими словами, корреляция пары характерных признаков опосредуется корреляцией каждого из признаков с некоторыми латентными факторами F_r. Латентные факторы определяют само существование связи i-го и j-го коррелирующих признаков.

Равенства Тэрстоуна допускаются гипотетически. Реально *AA′* и *ACA′* будут далеко не всегда в точности воспроизводить *R*. По крайней мере, это объясняется двумя причинами. Во-первых, в факторном анализе, позволяющем эффективно объяснять общую дисперсию данных, *r* – число латентных (обобщенных) признаков, как правило, значительно меньше числа исходных признаков *m*. И, во-вторых, в матрице *A* объединяются теоретические оценки факторных нагрузок. С учетом различий математических методов и специфичности вычислительных процедур следует допустить, что они не абсолютно истинны.

Таким образом, можно ожидать, что воспроизведенная из *AA′* или *ACA′* матрица корреляций R^+ будет отлична от R. Как следствие, на главной диагонали R^+ располагаются величины, обычно не равные, а меньшие единицы. На практике значения r^+_{ij} принимают за общности h_j, т.е. характеристики части дисперсии, поддавшейся объяснению через латентные факторы F_r, а $1 - a^+_{ij}$ – специфичность, т.е. необъясненная часть дисперсии. По степени расхождения R^+ и *R* судят о достаточности числа выделенных латентных факторов и адекватности аналитических выводов.

Матрица корреляций с общностями на главной диагонали называется редуцированной. Она является исходной для нахождения матрицы факторных нагрузок.

Существуют достаточно простые методы поиска общностей h_j:

метод наибольшей корреляции. На главной диагонали с положительным знаком записывается наибольший по величине коэффициент корреляции;

метод Барта. По каждому столбцу матрицы *R* вначале находят среднее значение коэффициентов корреляции $\overline{r_j}$, затем, если $\overline{r_j}$, сравнительно велико, за общ-

ность принимается значение, которое несколько выше наибольшего в столбце коэффициента корреляции и, если $\overline{r_j}$ – сравнительно малое значение, общность будет несколько меньше наибольшего в столбце коэффициента корреляции;

метод триад. Общности для каждого j -го столбца *R* вычисляют по формуле:

$$h_j^2 = \frac{r_{ik} r_{il}}{r_{kl}}$$

где r_{ik} r_{il} – коэффициенты корреляции, наибольшие в столбце;

метод малого центроида. Для каждой переменной *j* строится корреляционная матрица размерности 4x4. Включая саму переменную в эту матрицу, записывают оценки корреляции трех других переменных, особенно тесно связанных с первой. По данным малой матрицы корреляций и рассчитывают общности:

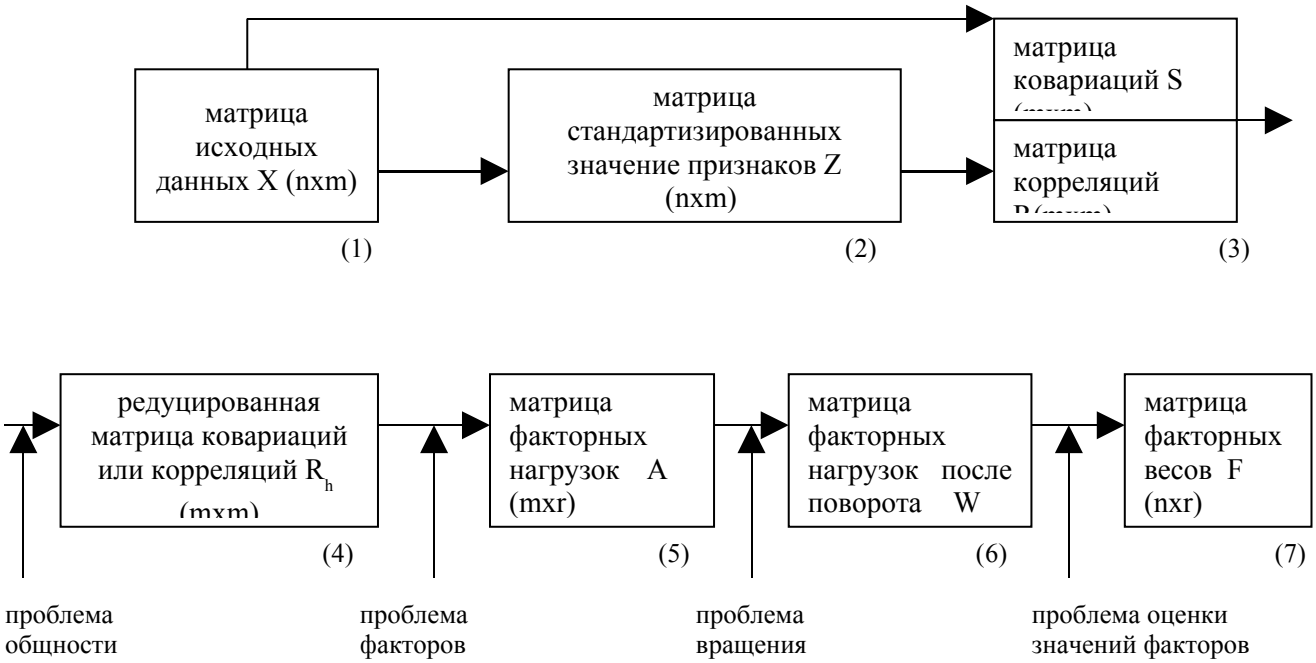
$$h_j^2 = \frac{\left(\sum r_{i1}\right)^2}{\sum r_{ij}}$$

где $\sum r_{i1}$ – сумма элементов первого столбца; $\sum r_{ij}$ – сумма всех элементов матрицы 4x4.

После определения редуцированной матрицы находят матрицу факторных нагрузок и по ее данным интерпретируют латентные факторы. Наилучшие решения находят при помощи современных методов факторного анализа: главных факторов, максимального правдоподобия и др. В общем случае выделенные факторы не обязательно ортогональны и тогда векторы (столбцы) матрицы факторных нагрузок будут линейно-зависимыми.

Лекция 17. Моделирование значений наблюдаемых переменных на основе выделенных латентных факторов.

Методы факторного анализа при всем их многообразии имеют общий алгоритм решения. Начинаясь построением матрицы исходных данных *X,* этот алгоритм завершается получением матриц факторного отображения (факторных нагрузок) и значений факторов (факторных весов) – *A* и *F.* С учетом принятых обозначении где *n* – число наблюдений, *m* – число аналитических признаков *X, r* – число значимых обобщенных признаков (латентных факторов), на схеме показана размерность матриц данных для каждого алгоритмического шага.



Первые шаги алгоритма 1–3 не вызывают каких-либо затруднений. Переход от матрицы исходных данных *X* к матрице стандартизованных данных *Z* осуществляется после пересчета всех элементов.

На следующем шаге простым перемножением скаляра 1/n и матриц Z^T и Z получаем матрицу парных корреляций: $R = 1/nZ^T Z$.

Шаг 2 может быть опущен и тогда последующее факторное решение находят не по матрице корреляций, а по матрице ковариаций, но тогда анализируемые признаки должны иметь одни и те же единицы измерения.

Выполнение четвертого шага алгоритма обуславливается решением первой проблемы – построения редуцированной матрицы корреляций с общностями на главной диагонали.

Вторая проблема возникает на этапе построения матрицы факторных нагрузок *A* и заключается в выборе оптимального метода для поиска весовых коэффициентов *a* элементов матрицы *A*.

Выполнение шага 6 алгоритма и решение проблемы вращения пространства общих факторов не обязательно.

На последнем этапе алгоритма необходимо получить значения каждого из выделенных факторов для каждого индивидуального объекта исследования, т.е. матри-

цу факторных весов.

На основе исходных данных в матрице значений Y и матрицы *A* возможно получить оценки элементов матрицы *F*. В зависимости от решаемой задачи по этим оценкам можно судить о каждом объекте исследования по *m* общим факторам.

Для уяснения методики приступим к оценке *F* в методе главных компонент.

$Y=AF$.

Y имеет размерность (nxN); порядок A равен n, а F - (nxN). Поскольку при извлечении всех главных компонент матрица A квадратная, то задача получения матрицы F не вызывает затруднений, если матрица R имеет ранг, равный n. Умножим обе части равенства слева на A^{-1}, получим

$$F = A^{-1}Y$$

По этой формуле получаются точно и однозначно индивидуальные значения главных компонент для каждого объекта исследования.

Чаще всего извлекаются не все главные компоненты, а только (m<n), поэтому матрица A не квадратная, а значит не имеет обратной матрицы. В этом случае для нахождения матрицы факторных весов необходимо в первоначальном равенстве обе части слева умножить на $(A^TA)^{-1}A^T$, получим

$\left(A^TA\right)^{-1}A^TY = \left(A^TA\right)^{-1}A^TAF = A^{-1}\left(A^T\right)^{-1}A^TAF = A^{-1}EAF = F$, т.е.

$$F = \left(A^TA\right)^{-1}A^TY$$

В этом выражении не надо обращать матрицу A. Если A не квадратная матрица, то $\left(A^TA\right)$ будет квадратной порядка m.

Дать точное определение индивидуальных значений факторов, как в случае выделения всех факторов не возможно. Поскольку задача не решается однозначно, то можно методом наименьших квадратов получить оценки индивидуальных значений общих факторов. Удобно обратиться к методу регрессионного анализа, когда имеется одна зависимая нормированная переменная и n независимых переменных, которые связаны между собой линейно.

$$y_{0i} = \beta_1 y_{1i} + \beta_2 y_{2i} + \ldots + \beta_n y_{ni} + e_i$$

Коэффициенты β_j выбираются таким образом, чтобы сумма квадратов ошибок оценок e^2 была минимальной.

Произведение корреляционной матрицы R на вектор-столбец коэффициентов регрессии β равен вектору-столбцу коэффициентов корреляции между оценками нормированных значений зависимой переменной и всеми исходными признаками V,т.е. $V = R\beta$, $\beta = R^{-1}V$, $\beta^T = (R^{-1}V)^T = V^T R^{-1}$

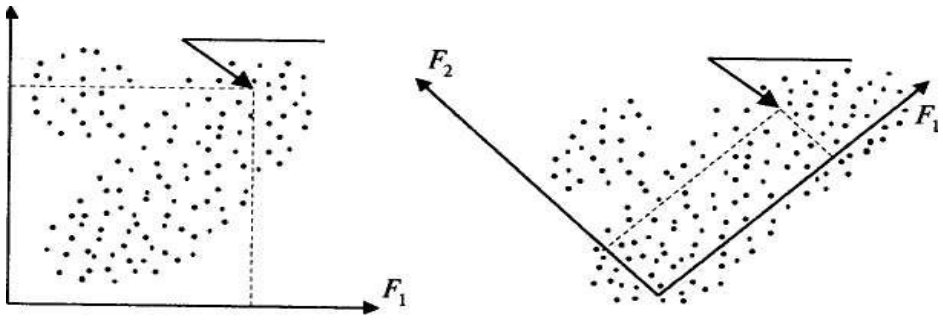
Матрица оценок индивидуальных значений факторов может быть определена по формуле: $F = \beta^T Y = V^T R^{-1} Y$,

где элементами матрицы V являются коэффициенты корреляции между переменными и факторами (матрица факторных нагрузок), R – матрица коэффициентов корреляции между переменными, Y – матрица исходных данных.

Как правило, проведение факторного анализа заканчивается оценкой индивидуальных значений факторов для каждого объекта исследования.

Лекция 18. Методы главных компонент и главных факторов, вращение факторов.

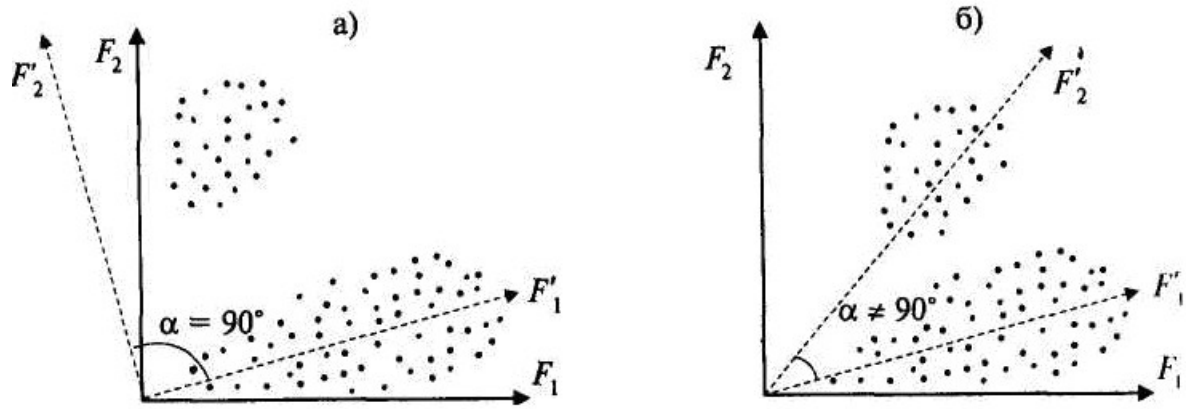
Потребность во вращении возникает, когда пространственное расположение общих факторов F_r нелогично или трудно поддается интерпретации. Возможность появления нелогичных первых результатов анализа объясняется не определяемым четко и не задаваемым положением факторных осей в пространстве, или, отсутствием изначально какой-либо пространственной привязки для осей F_r.



На рисунке показаны два различных положения в пространстве факторных осей *{F_1 и F_2)*. Легко заметить, что изменение положения F_1 и F_2 одновременно приводит к изменению координат исходных признаков X_j. Цель поворота – преобразование координат (факторных нагрузок) таким образом, чтобы факторобразующие признаки имели наибольшие нагрузки, близкие к единице, а остальные признаки – минимальные значения, близкие к нулю, т.е. добиваются экономичного описания данных.

Повороты осей могут быть ортогональными и косоугольными. Предпочтительно, хотя и более трудно выполнимо и интерпретируемо, косоугольное вращение, при этом, значительно повышаются возможности оптимального отображения сгущений признаков в пространстве R^F.

На рис. а ось F' после поворота F, очевидно, займет более рациональное положение, но из-за жесткости осевой конструкции положение F_2 удаляется от оптимального; на рис. б косоугольным вращением приходят к оптимизации положения сразу обеих осей F_1 и F_2



Вращение пространства общих факторов F_r не изменяет величин общностей h и по-прежнему $AA^T = R^+$, или $ACA^T = R^+$ при $R^+ \rightarrow R$.

Рассмотрим особенности наиболее часто применяющихся методов главных компонент и главных факторов, которые имеют много общего.

Метод главных компонент (Г. Хотеллинг) строго говоря, не относится к факторному анализу, хотя он имеет с ним много общего. Специфическим является, во-первых, то, что в ходе вычислительных процедур одновременно получают все главные компоненты и их число первоначально равно числу элементарных признаков; во-вторых, постулируется возможность полного разложения дисперсии элементарных признаков, другими словами, ее полное объяснение через латентные факторы (обобщенные признаки).

Метод главных факторов заключается в том, что дисперсия элементарных признаков здесь объясняется не в полном объеме, признается, что часть дисперсии остается нераспознанной как характерность. Факторы выделяются последовательно: первый, объясняющий наибольшую долю вариации элементарных признаков, затем второй, объясняющий меньшую, вторую после первого латентного фактора часть

дисперсии, третий и т.д. Процесс выделения факторов может быть прерван на любом шаге, если принято решение о достаточности доли объясненной дисперсии элементарных признаков или с учетом интерпретируемости латентных факторов.

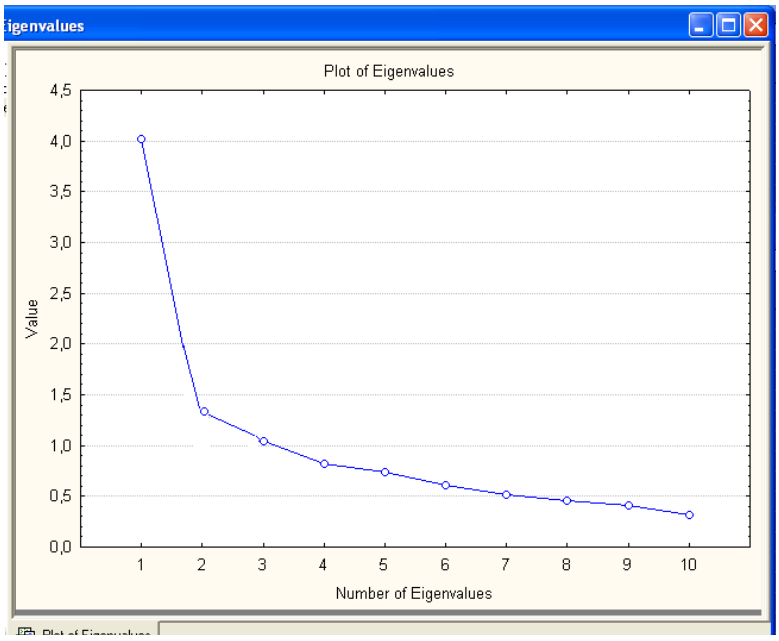
Алгоритм обоих методов начинается с получения матрицы парных коэффициентов корреляции с единицами на главной диагонали. После этого определяются общности и получают редуцированную матрицу.

Для определения латентных факторов находят собственные числа и собственные векторы редуцированной корреляционной матрицы.

Если использовать метод главных компонент, то необходимо определить собственные векторы для каждого собственного числа. В методе главных факторов первоначально отбирают значимые собственные числа, количество которых соответствует количеству главных факторов. Для этого используют критерий Кайзера или критерий факторной осыпи.

1. Критерий Кайзера предлагает отобрать только факторы, с собственными значениями, большими *1*.

2. Критерий факторной осыпи является графическим методом, впервые предложенным Кэттелем (Cattell). Необходимо изобразить собственные значения, расположенные в убывающем порядке в виде графика (по оси абсцисс порядковый номер числа, по оси ординат его значение).



Кэттель предложил найти такое место на графике, где убывание собственных значений слева направо максимально замедляется. Значения выше этой точки опре-

деляют оптимальное количество факторов. Предполагается, что справа от этой точки находится только «факториальная осыпь» – «осыпь» является геологическим термином, обозначающим обломки горных пород, скапливающиеся в нижней части скалистого склона. В соответствии с этим критерием можно оставить в приведенном примере 2 или 3 фактора.

Первый критерий (*критерий Кайзера*) иногда сохраняет слишком много факторов, в то время как второй критерий (*критерий каменистой осыпи*) иногда сохраняет слишком мало факторов; однако оба критерия вполне хороши при нормальных условиях, когда имеется относительно небольшое число факторов и много переменных. Обычно исследуется несколько решений с большим или меньшим числом факторов, и затем выбирается одно наиболее интерпретируемое.

Независимо от метода для выделенных собственных чисел определяются собственные векторы матрицы. Нормированные координаты собственных векторов, умноженные на весовой коэффициент, который равен корню из соответствующего собственного числа, являются столбцами матрицы факторных нагрузок.

Алгоритмы факторного анализа отличаются, трудоемкостью, их полное выполнение возможно при условии использования технических средств.

3.2. Задания для практических и домашних работ.

1. Основные вопросы математической статистики.

Типовые задания.

1. Выборка задана в виде распределения частот. Найти объем выборки, распределение относительных частот, построить эмпирическую функцию распределения, изобразить гистограмму и полигон частот и относительных частот.

2. Из генеральной совокупности извлечена выборка. Определить несмещенную оценку генеральной средней, смещенную и несмещенную оценки дисперсии генеральной совокупности.

3. Найти доверительный интервал для оценки с надежностью 0,95 неизвестного математического ожидания нормально распределенного признака генеральной совокупности.

4. Найти минимальный объем выборки, при котором с надежностью 0,95 точность оценки математического ожидания генеральной совокупности по выбороч-

ной средней будет равна δ.

Примечание: решение подобных заданий не вызывает затруднений, но способствует осознанному восприятию других разделов курса.

2. Параметрические критерии. Критерии Стьюдента.

Типовые задания.

1. Два университета (А и В) готовят специалистов аналогичных специальностей. Министерство образования решило проверить качество подготовки в обоих университетах, организовав для этого объемный тестовый экзамен для студентов пятого курса. Отобранные случайным образом студенты показали следующие результаты:

A: 41, 50, 35, 45, 53, 30, 57, 20, 50, 44, 36, 48, 55, 28, 40, 50;

B: 40, 57, 52, 38, 20, 25, 47, 52, 48, 55, 48, 53, 39, 49, 46, 45, 55, 43, 51, 55, 40.

Можно ли утверждать при уровне значимости α=0,05, что один из университетов обеспечивает лучшую подготовку.

Решение: обозначим результаты тестирования студентов университета А через *x,* а университета В – *y.*

Для ответа на вопрос задачи необходимо проверить гипотезу о равенстве средних значений успеваемости студентов:

H_0: $\bar{x} = \bar{y}$ - средняя успеваемость студентов в университетах А и В одинакова,

H_1: $\bar{x} \neq \bar{y}$ - средняя успеваемость студентов в университетах А и В различна.

Так как выборки независимы, то рассчитаем $t_{набл}$ по формуле:

$$t_{набл} = \frac{\bar{x} - \bar{y}}{\sqrt{(n_1 - 1)S_x^2 + (n_2 - 1)S_y^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}$$, с учетом того, что

объем выборки университета А - n_1=16,

объем выборки университета В - n_2 =21,

среднее значение успеваемости для университета А-

$$\bar{x} = \frac{41 + 50 + 35 + 45 + 53 + 30 + 57 + 20 + 50 + 44 + 36 + 48 + 55 + 28 + 40 + 50}{16} \approx 42{,}6;$$

среднее значение успеваемости для университета В-

$$\bar{y} = \frac{40 + 57 + 52 + 38 + 20 + 25 + 47 + 52 + 48 + 55 + 48 + 53 + 39 + 49 + 46 + 45 + 55 + 43 + 51 + 55 + 40}{21} \approx 45{,}6$$

$$S_x^2 = \frac{1}{n_1 - 1}\sum (x - \bar{x})^2 = \frac{1}{16 - 1}\left[(41 - 42{,}6)^2 + (50 - 42{,}6)^2 + (35 - 42{,}6)^2 + \ldots + (40 - 42{,}6)^2 + (50 - 42{,}6)^2\right] = 110{,}91$$

выборочная дисперсия 1 выборки;

$$S_y^2 = \frac{1}{n_2 - 1}\sum (y - \bar{y})^2 = \frac{1}{21 - 1}\left[(40 - 45{,}6)^2 + (57 - 45{,}6)^2 + \ldots + (55 - 45{,}6)^2 + (40 - 45{,}6)^2\right] = 91{,}85$$

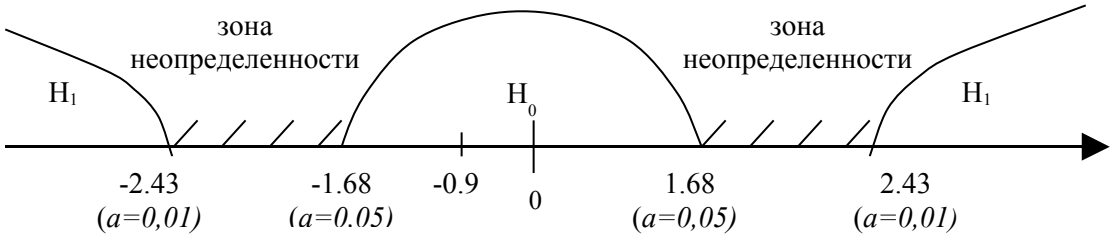
выборочная дисперсия 2 выборки.

$$t_{набл} = \frac{42{,}6 - 45{,}6}{\sqrt{(16 - 1)\cdot 110{,}91 + (21 - 1)\cdot 91{,}85}} \cdot \sqrt{\frac{16 \cdot 21(16 + 21 - 2)}{16 + 21}} = -0{,}9$$

Критическое значение $t_{кр}$ находим с помощью таблицы Стьюдента. Число степеней свободы $k = n_1 + n_2 - 2 = 16 + 21 - 2 = 35$

$$t_{кр} = \begin{cases} 2{,}43 & для\ \alpha = 0{,}01 \\ 1{,}68 & для\ \alpha = 0{,}05 \end{cases}$$

Построим доверительный интервал:



Так как значение $t_{набл} = -0{,}9$ принадлежит области принятия нулевой гипотезы, то нет основания ее отвергать, значит средняя успеваемость студентов в университетах А и В одинакова, т.е. университеты обеспечивают равнозначное качество подготовки специалистов.

2. На выполнение одного тестового задания предусмотрены по норме 2 мин. Группа на выполнение теста по математическим методам, состоящего из 30 заданий потратила более 1 часа. Для проверки того факта, соответствуют ли предложенные задания норме, были проведены измерения времени, затраченного на выполнение теста у 49 случайных студентов гуманитарных специальностей. Эксперимент показал следующие результаты: среднее время выполнения данного теста студентами 2 мин 10 с, среднее квадратическое отклонение затраченного времени 20 с. Можно

ли говорить о том, что вопросы теста соответствуют норме?

Решение: Обозначим время выполнения одного задания теста через t. Сформулируем гипотезы:

H_0: *t = 2 мин = 120 с* – время выполнения задания теста соответствует норме.

H_1: *t > 120 с* – время выполнения тестового задания больше нормы.

Так как среднее квадратическое отклонение, а, следовательно, и дисперсия времени, затраченного на выполнение одного задания известна, воспользуемся формулой: $t_{набл} = \frac{\bar{t} - t_0}{\sigma}\sqrt{n} = \frac{130 - 120}{20}\sqrt{49} = 3{,}5$

где $\bar{t}$ *= 2 мин 10 с = 130 с* – среднее время выполнения задания;

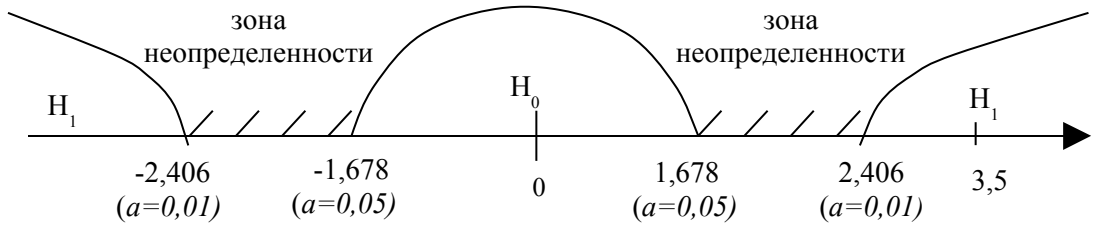
t_0 *= 120 с* – время, предусмотренное нормой;

σ *= 30 с* - известное среднее квадратическое отклонение;

n=49 - объем выборки.

Критическое значение $t_{кр}$ найдем по таблице Стьюдента, вычислив число степеней свободы *k = n – 1=49 – 1 = 48*.

$$t_{кр} = \begin{cases} 1{,}67 & для\ \alpha = 0{,}05 \\ 2{,}40 & для\ \alpha = 0{,}01 \end{cases}$$



Наблюдаемое значение попадает в область принятия альтернативной гипотезы, поэтому можно утверждать, что не все вопросы теста соответствуют предусмотренной норме.

Задания для домашней работы:

1. Для изучения влияния двухнедельной диеты и соответствующего комплекса упражнений на изменение веса спортивный клуб провел анализ по двум выборкам из 7 человек до и после диеты и упражнений. Отбор осуществлялся случайным образом по членским карточкам. Получены следующие результаты (буквы – инициалы испытуемого, цифры – вес, кг):

I выборка: АГ 85,5; ВТ 92,7; ДИ 79; КД 68,6; КЛ 102,5; МА 88,3; ТВ 82,7

II выборка: БП 90,5; ДК 77,5; ИВ 85,3; КР 72,5; ЛМ 108,7; МТ 80,3; ЯК 79.

а) Можно ли по имеющимся данным достаточно объективно оценить результаты диеты и упражнений? Обоснуйте свой ответ.

б) Для того же анализа в повторную выборку отобрали тех же людей, что и в первую выборку, и получили следующие данные:
АГ 83; ВТ 90,5; ДИ 77,5; КД 68; КЛ 94,5; МА 85; ТВ 80,5. Оцените результаты диеты и упражнений.

Есть ли основания не доверять рекламному проспекту клуба, обещающему потерю веса в 3 кг?

2. При исследовании способности к восприятию речи обнаружено, что каждый человек воспринимает лучше синхронную речь, (движение губ и воспроизводимые слова совпадают), т.е. у каждого человека в той или иной степени развито умение «читать по губам». Для проверки этого факта провели эксперимент. Испытуемым предлагалось несколько синхронизированных и несинхронизированных фраз, которые они должны были воспроизвести. Количество правильно воспроизведенных фраз указано в таблице. Сформулируйте возможные гипотезы и осуществите их проверку, если в эксперименте участвовало 12 человек.

	1	2	3	4	5	6	7	8	9	10	11	12
Не синхронно	20	17	16	25	5	29	2	36	25	8	34	18
Синхронно	50	37	25	28	16	36	1	43	44	10	29	27

3. Для анализа психологической адаптации первокурсников проведено измерение времени, которое потрачено студентами для установления дружеских отношений в группе. Результаты исследования представлены в таблице. Можно ли утверждать, что юноши быстрее привыкают к новым условиям по сравнению с девушками?

	Количество дней, необходимых для установления дружеских отношений с одногруппниками.									
юноши	10	4	7	8	12	5	6	2	23	7
девушки	20	13	5	9	4	12	10			

4. На основании наблюдений за работой 25 кандидатов на должность секретаря-референта установлено, что в среднем они тратили 7 минут на набор одной страницы сложного текста на компьютере при выборочном стандартном отклонении S - 2 минуты. При предположении, что время (X) набора текста имеет нормальный за-

кон распределения:

а) Оцените количество претендентов на работу, которые набрали текст быстрее, чем за 5 минут.

б) Предполагалось, что среднее время набора страницы текста должно составить 5,5 минуты. Не противоречат ли полученные данные этой гипотезе?

5. Считается, что у студентов, обучающихся на специальности «Психология» средний балл аттестата равен 4,5. Приемная комиссия отметила снижение оценок в аттестатах у некоторых абитуриентов. Для проверки этого факта вычислен средний балл аттестата у 20 случайно выбранных абитуриентов, поступающих на специальность «Психология». Результаты представлены в таблице. Можно ли утверждать, что средний балл аттестата соответствует норме?

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
ср.балл	4,2	4	5	4,5	4,3	4,6	4,8	4,1	4	4,5	4,7	4,2	4,1	4,3	4,2	4	4,6	4,5	4,2	5

3. Критерии Фишера для сравнения «разбросов» значений. Решение задач с комбинированным применением параметрических критериев.

Типовые задания.

1. Два университета (А и В) готовят специалистов аналогичных специальностей. Министерство образования решило проверить качество подготовки в обоих университетах, организовав для этого объемный тестовый экзамен для студентов пятого курса. Отобранные случайным образом студенты показали следующие результаты:

А: 41, 50, 35, 45, 53, 30, 57, 20, 50, 44, 36, 48, 55, 28, 40, 50;

В: 40, 57, 52, 38, 20, 25, 47, 52, 48, 55, 48, 53, 39, 49, 46, 45, 55, 43, 51, 55, 40.

Есть ли основания считать, что разброс оценок у студентов одного университета больше чем у другого.

Решение: обозначим результаты тестирования студентов университета А через *x,* а университета В – *y.*

Сформулируем гипотезы:

H_0: $S_x^2 = S_y^2$ - разброс оценок относительно среднего одинаковый в обеих университетах.

H_1: $S_x^2 \neq S_y^2$ - разброс оценок в университетах А и В различен.

Наблюдаемое значение F – критерия рассчитывается по формуле

$$F_{набл} = \frac{S_x^2}{S_y^2} = \frac{110{,}91}{91{,}85} \approx 1{,}21$$

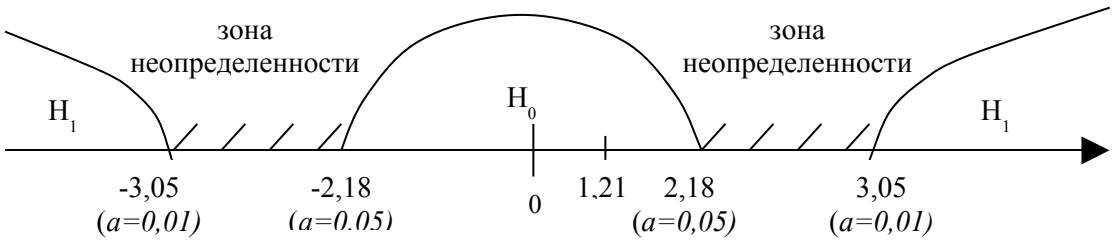
где S_x^2=110,91– большая (по величине) выборочная дисперсия,

S_y^2=91,85 – меньшая (по величине) выборочная дисперсия.

$F_{крит}$ найдем по таблице Фишера для числа степеней свободы $k_1 = n_1 - 1 = 16-1 = 15$ и $k_2 = n_2 - 1 = 21-1 = 20$.

$$F_{крит} = \begin{cases} 2{,}20 & для\ \alpha = 0{,}05 \\ 3{,}09 & для\ \alpha = 0{,}01 \end{cases}$$

Строим доверительный интервал:



Наблюдаемое значение критерия попадает в область принятия нулевой гипотезы, значит разброс в знаниях у студентов в университетах А и В одинаковый.

2. Сравните средний уровень знаний и диапазон в оценках по иностранному языку у студентов одной группы, если в группе 20 человек из них 9 студентов изучают английский язык, а 11 – немецкий, и на тестовом зачете по предмету показали следующие результаты:

английский язык - 35, 44, 34, 50, 52, 45, 44, 43, 38;

немецкий язык – 55, 45, 50, 46, 44, 53, 56, 45, 50, 40, 55.

Задания для домашней работы:

1. Сравните разброс уровня развития интеллекта юношей и девушек 18 лет, если данные теста по определению IQ даны в таблице:

	Уровень IQ.						
юноши	120	101	98	86	100	95	110
девушки	105	87	96	125	101	90	120

2. Два университета готовят специалистов по специальности бухгалтер. В один университет поступило 15 абитуриентов, в другой 12. Сравнить разброс оценок у

абитуриентов, поступивших на данную специальность в двух разных вузах.

A: 15,30,38,26,29,40,31,29,21,24,29,20,23,32,27

B: 26,30,32,19,32,35,36,27,31,31,33,29

3. Оцените разброс в ценах между фирмами, занимающимися продажей и изготовлением бытовой техники.

1 фирма: 29,15,30,75,69,86

2 фирма:31,39,46,59,70,75,74

4. Сравните разброс в знаниях иностранного языка у студентов одной группы если в группе 25 человек, 15 из них изучают английский, 10 французский и на тестовом зачете показали следующие результаты:

английский: 35,40,45,50,32,33,38,55,60,70

французский: 55,45,50,46,44,53,56,38,40,65,75

4. Однофакторный дисперсионный анализ.

Типовые задания.

Три различные группы из шести человек получили списки из десяти слов. Первой группе слова предъявлялись с низкой скоростью (1 слово в 5 секунд), второй группе – со средней скоростью (1 слово в 2 секунды), третьей – с высокой (1 слово в секунду). Количество воспроизведенных слов представлено в таблице. Определить влияет ли скорость предъявления слов на их воспроизведение.

№	1 группа	2 группа	3 группа
1	8	7	4
2	7	8	5
3	9	5	3
4	5	4	6
5	6	6	2
6	8	7	4

Решение: необходимо выяснить изменяется ли количество воспроизведенных слов в трех независимых выборках при разной скорости предъявления слов, т.е. решить задачу однофакторного дисперсионного анализа.

Сформулируем гипотезы:

H_0: скорость предъявления слов не влияет на количество воспроизведенных слов;

H_1: количество воспроизведенных слов зависит от скорости их предъявления.

Наблюдаемое значение критерия вычислим по формуле $F_{набл} = \frac{S^2_{факт}}{S^2_{сл}}$.

Предварительно определим значения всех параметров, входящих в формулу:

$n=6$ – количество наблюдений,

$m=3$ – количество выборок (групп),

$\bar{x}_{гр\,1} = \frac{8+7+9+5+6+8}{6} \approx 7{,}17$ - среднее значение признака в 1 группе,

$\bar{x}_{гр\,2} = \frac{7+8+5+4+6+7}{6} \approx 6{,}17$ - среднее значение признака во 2 группе,

$\bar{x}_{гр\,3} = \frac{4+5+3+6+2+4}{6} = 4$ - среднее значение признака в 3 группе,

$\bar{x} = \frac{8+7+9+5+6+8+7+8+5+4+6+7+4+5+3+6+2+4}{18} \approx 5{,}78$-среднее значение признака по всей совокупности,

$к$ – число степеней свободы:

$$к_{общ} = n \cdot m - 1 = 6 \cdot 3 - 1 = 17$$

$$к_{факт} = m - 1 = 3 - 1 = 2$$

$$S^2_{факт} = \frac{n \cdot \sum \left(\bar{x}_{гр} - \bar{x}\right)^2}{к_{факт}} = \frac{6 \cdot \left[(7{,}17 - 5{,}78)^2 + (6{,}17 - 5{,}78)^2 + (4 - 5{,}78)^2\right]}{2} = 15{,}7578$$

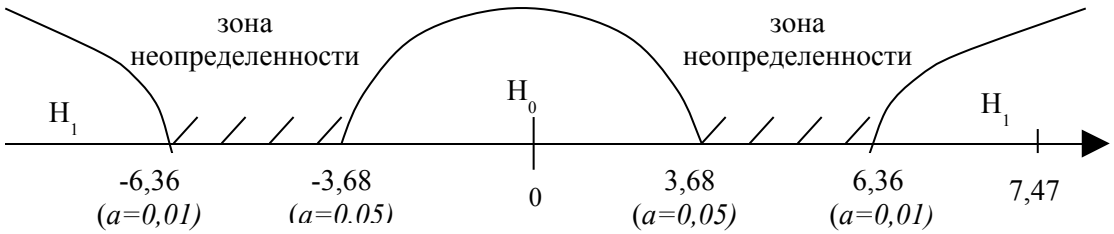
$$S^2_{сл} = \frac{\sum \left(x_{ij} - \bar{x}\right)^2 - n \cdot \sum \left(\bar{x}_{гр} - \bar{x}\right)^2}{к_{общ} - к_{факт}} =$$

$$= \frac{\left[(8 - 5{,}78)^2 + (7 - 5{,}78)^2 + \ldots + (2 - 5{,}78)^2 + (4 - 5{,}78)^2\right] - 6 \cdot \left[(7{,}17 - 5{,}78)^2 + (6{,}17 - 5{,}78)^2 + (4 - 5{,}78)^2\right]}{17 - 2} = 2{,}11$$

$$F_{набл} = \frac{S^2_{факт}}{S^2_{сл}} = \frac{15{,}7578}{2{,}11} \approx 7{,}47$$

Критические значения $F_{крит}$ найдем с помощью таблицы Фишера, число степеней свободы $k_1=k_{факт}=2$ и $k_2=k_{общ}-k_{факт}=17-2=15$.

$F_{крит} = \begin{cases} 3{,}68 & для\ \alpha = 0{,}05 \\ 6{,}36 & для\ \alpha = 0{,}01 \end{cases}$. Строим «ось значимости»:



Наблюдаемое значение критерия попадает в область принятия альтернативной

гипотезы, значит, скорость предъявления слов влияет на количество их воспроизведения.

Задания для домашней работы:

1. Группа из шести человек проходила интеллектуальный тест, состоящий из 30 равнозначных по сложности вопросов. На первые 10 вопросов теста группе пришлось ответить в 8 часов утра, на вторые 10 - в полдень, а на последние 10 вопросов в конце рабочего дня. Количество правильных ответов представлено в таблице. Определите, влияет ли время суток на количество верных ответов.

утро	полдень	вечер
8	10	6
9	7	7
7	7	6
8	10	8
9	9	7
7	9	6

2. При интервью чаще всего люди отказываются отвечать на вопросы личного характера (о состоянии здоровья, о размере заработной платы). Исследователь решил выяснить, зависит ли согласие отвечать на вопросы о личной жизни от стиля интервьюирования. Для этого проведен эксперимент: одним и тем же испытуемым предлагалось ответить на ряд вопросов личного характера. В первом случае вопросы задавались в восторженной манере, в дружественной беседе; во втором – стиль общения был формальным, а в третьем – происходила резкая незаинтересованная беседа. 10 вопросов личного характера были «перемешаны» с общими вопросами анкеты. Результаты эксперимента – количество ответов на вопросы личного характера представлены в таблице 2.7. Определить зависит ли согласие отвечать на вопросы от стиля интервьюирования, если разные способы апробировались на разных людях.

Способ 1	Способ 2	Способ 3
10	8	5
8	7	5
6	7	2
9	7	3
10	5	3
9	5	6
9	9	4
6	6	4
7	8	4
7	7	5
8	8	6
8	8	2
8	9	3
9	9	1
10	10	0

5. Многофакторный дисперсионный анализ.

Типовые задания.

Две группы студентов пишут тест по математической статистике. Результаты теста даны в процентах. Одна группа студентов выполняла предложенное задание на занятии, в присутствии преподавателя, а другая дома самостоятельно. Помимо этого полученные результаты теста разбиты на две части согласно половой принадлежности испытуемых и занесены в таблицу. Определить влияют ли на результаты теста представленные условия.

Пол	На занятии	Дома
муж	37	56
	47	57
	40	58
	60	53
жен	69	60
	55	58
	40	46
	32	40

Решение: сформулируем гипотезы:

1) H_o: место выполнения задания (условие А) не влияет на результаты теста.

H_1: место выполнения задания влияет на результаты теста.

2) H_o: пол испытуемого (условие В) не влияет на результаты теста.

H_1: пол испытуемого влияет на результаты теста.

3) H_o: влияние места выполнения задания одинаково для испытуемых разного пола (условие АВ).

H_1: влияние места выполнения задания для испытуемых разного пола различно.

Произведем необходимые расчеты:

$$SS_A = nb\sum (\overline{x_{грA}} - \bar{x})^2 = 8 \cdot 2\left((47{,}5 - 50{,}5)^2 + (53{,}5 - 50{,}5)^2\right) = 288$$

$$SS_B = na\sum (\overline{x_{грB}} - \bar{x})^2 = 8 \cdot 2\left((51 - 50{,}5)^2 + (50 - 50{,}5)^2\right) = 8$$

$$SS_{AB} = n\sum (\overline{x_{AB}} - \bar{x})^2 = 8\left((46 - 50{,}5)^2 + (56 - 50{,}5)^2 + (49 - 50{,}5)^2 + (51 - 50{,}5)^2\right) = 424$$

$$SS_{общ} = \sum (x_i - \bar{x})^2 = 1622$$

$$SS_{сл} = SS_{общ} - SS_A - SS_B - SS_{AB} = 1622 - 288 - 8 - 424 = 902$$

где $\overline{x_{грA}}$ – среднее по градациям условия А;

$\overline{x_{грB}}$ – среднее по градациям условия В;

$\bar{x}$ - общее среднее;

$\overline{x_{AB}}$ – среднее по ячейкам «пол – место выполнения»;

n – количество испытуемых;

a – количество градаций условия А

b – количество градаций условия В.

Вычислим число степеней свободы:

$k_A = a - 1 = 2 - 1 = 1$

$k_B = b - 1 = 2 - 1 = 1$

$k_{AB} = k_A k_B = 1$

$k_{общ} = N - 1 = 16 - 1 = 15$

$k_{сл} = k_{общ} - k_A - k_B - k_{AB} = 12$

Определим дисперсии: $S^2 = \frac{SS}{k}$

$S_A^2 = 288$, $S_B^2 = 8$, $S_{AB}^2 = 424$, $S_{сл}^2 = 75{,}2$.

Воспользуемся статистикой F-Фишера. Вычислим наблюдаемые значения статистики по формуле: $F = \frac{S^2}{S_{сл}^2}$, имеем $F_A = 3.83$, $F_B = 0.11$, $F_{AB} = 5.63$.

Определяем критические значения F-критерия для уровня значимости 0,05.

$F_{крит} = 4{,}74$

F_A и F_B меньше $F_{крит}$, значит ни одно из условий (А или В) не влияет на результаты теста, но взаимодействие условий оказывает достоверные влияния на результаты теста, т.к. F_{AB} больше $F_{крит}$.

Задания для домашней работы:

Группа студентов из 20 человек изучает иностранный язык в течение 1 года, причем 10 человек занимаются изучением английского языка в 1-ой подгруппе, а 10– во второй.

Для анализа успешности обучения проведен тестовый контроль знаний студентов каждой группы сразу после изучения, а затем через год после изучения языка. Проанализировать полученные результаты.

	№	1 подгруппа	2 подгруппа

После изучения	1		
	2		
	3		
	…		
	10		
Через год после изучения	1		
	2		
	3		
	…		
	10		

Введите произвольные данные результатов тестирования.

6. Непараметрические критерии различий

Типовые задания.

1. Используя тест Векслера психолог определил показатели интеллекта у двух групп учащихся из городской (100, 104, 120, 120, 96, 126, 130, 134, 120) и сельской школы (100, 82, 120, 118, 82, 84, 76, 110, 102, 104, 76, 88). Его интересует вопрос – будут ли обнаружены статистически значимые различия в показателях интеллекта, если в городской выборке 11 детей, а в сельской 12?

Решение: Сформулируем гипотезы:

H_0: Городские школьники не превосходят сельских школьников по уровню интеллекта.

H_1: Городские школьники превосходят сельских школьников по уровню интеллекта.

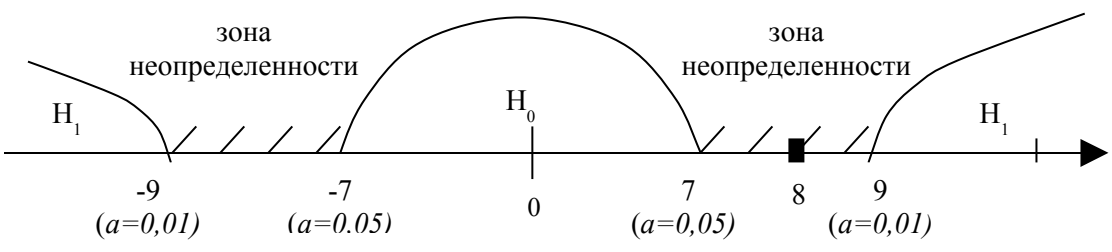
Проверку гипотез выполним с помощью критерия Q Розенбаума. Для решения задачи результаты измерений сразу представим в удобном для расчета критерия Q виде, т.е. расположив числа в порядке возрастания слева направо и одно измерение под другим (верхний ряд – городская школа, нижний- сельская):

S_2 |96, 100, 104, 104, 120, 120, 120, 120|, 126, 130, 134

76, 82, 82, 84, 88,|96, 100, 102, 104, 110, 118, 120 | S_1

В этом случае $S_1 = 3$, $S_2 = 5$, $Q_{эмп} = 3 + 5 = 8$.

Критические значения для критерия Q находим по таблице Розенбаума, по которой определяем, что для $n_1 = 11$ и $n_2 = 12$ при $\alpha = 0,05$ $Q_{кр} = 7$, а при $\alpha = 0,01$ $Q_{кр} = 9$.



Полученное значение $Q_{эмп}$ попало в зону неопределенности. Психолог поэтому может считать полученные различия между рядами значимыми на уровне 5% (т. е. принимать, что уровень интеллекта учащихся городской школы выше, чем у учащихся сельской школы) и незначимыми на уровне 1%, т.е. исходить из того, что показатели интеллекта не различаются в обеих школах.

2. Можно ли утверждать, что студенты-психологи превосходят студентов-физиков по уровню невербального интеллекта. Данные, полученные с помощью методики Д. Векслера, приведены в таблице

Студенты-физики		Студенты - психологи	
Код имени испытуемого	Показатель невербального интеллекта	Код имени испытуемого	Показатель невербального интеллекта
1. И.А.	111	1. Н.Т.	113
2. К.А.	104	2. ОБ.	107
Ъ. К.Е.	107	3. Е.В.	123
4. ПА.	90	4. Ф.О.	122
5. С.А.	115	5. И.Н.	117
6. Ст.А.	107	6. И.Ч.	112
7. Т.А.	106	7. И.В.	105
8. ФА.	107	8. К.О.	108
9. Ч.И.	95	9. Р.Р.	111
10. Ц.А.	116	10. Р.И.	114
11. См.А.	127	11. О.К.	102
12. К.Ан.	115	12. Н.К.	104
13. Б.Л.	102		
14. Ф.В.	99		

Решение. Проранжируем показатели в общей выборке

Студенты -физики (n_1=14)		Студенты-психологи (n_2=12)	
Показатель невербального интеллекта	Ранг	Показатель невербального интеллекта	Ранг
127	26		
		123	25
		122	24
		117	23
116	22		
115	20,5		
115	20,5		
		114	19
		113	18
		112	17
111	15,5	111	15.5
		108	14
107	11.5	107	11,5
107	11,5		
107	11,5		
106	9		
		105	8
104	6.5	104	6,5

	102 99 95 90	4,5 3 2 1	102	4,5
Суммы	1501	165	1338	186
Средние	107,2		111,5	

Мы видим, что по уровню невербального интеллекта более "высоким" рядом оказывается выборка студентов-психологов. Именно на эту выборку приходится большая ранговая сумма: 186.

Теперь мы готовы сформулировать гипотезы:

H_0: Группа студентов-психологов не превосходит группу студентов-физиков по уровню невербального интеллекта.

H_1: Группа студентов-психологов превосходит группу студентов-физиков по уровню невербального интеллекта.

Выполним проверку гипотез с помощью критерия U Манна-Уитни.

Определяем эмпирическую величину U:

$$U_{эмп} = (14 \cdot 12) + \frac{12(12+1)}{2} - 186 = 60.$$

По таблице Манна-Уитни определяем критические значения для $n_1=14, n_2=12$.

$$U_{кр} = \begin{cases} 51, & для\ \alpha \le 0{,}05 \\ 38, & для\ \alpha \le 0{,}01 \end{cases}.$$

$U_{эмп} > U_{кр}$, следовательно H_0 принимается. Группа студентов-психологов не превосходит группы студентов-физиков по уровню невербального интеллекта.

Обратим внимание на то, что для данного случая критерий Q Розенбаума неприменим, так как размах вариативности в группе физиков шире, чем в группе психологов: и самое высокое, и самое низкое значение невербального интеллекта приходится на группу физиков.

3. Четыре группы испытуемых выполняли тест Бурдона в разных экспериментальных условиях. Задача в том, чтобы установить – зависит ли эффективность выполнения теста от условий или, иными словами, существуют ли статистически достоверные различия в успешности выполнения теста между группами. В каждую группу входило четыре испытуемых. Число ошибок показателя переключаемости внимания в процентах дано в таблице.

№ испытуемых	1 группа	2группа	3группа	4 группа
1	23	45	34	21

2	20	12	24	22
3	34	34	25	26
4	35	11	40	27
Суммы	112	102	123	96

Решение. Сформулируем гипотезы:

H_0: Четыре группы, выполнявшие тест, не различаются в успешности выполнения теста.

H_1: Четыре группы, выполнявшие тест, различаются в успешности выполнения теста.

Для поверки гипотез будем использовать критерий Н Крускала-Уолиса.

Проранжируем показатели в независимости от того, к какой группе они принадлежат

Данные	Ранги	Данные	Ранги
11	1	26	9
12	2	27	10
20	3	34	12
21	4	34	12
22	5	34	12
23	6	35	14
24	7	40	15
25	8	45	16
Сумма рангов 136			

Следующий этап в подсчете $H_{эмп}$ состоит в распределении данных вновь на исходные группы, но уже с полученными рангами

№ испытуемых	1 группа	Ранги	2 группа	Ранги	3 группа	Ранги	4 группа	Ранги
1	23	6	45	16	34	12	21	4
2	20	3	12	2	24	7	22	5
3	34	12	34	12	25	8	26	9
4	35	14	11	1	40	15	27	10
Суммы	112	35	102	31	123	42	96	31

Теперь можно подсчитать величину $H_{эмп}$ по формуле (22)

$$H_{эмп} = \frac{12}{16+17}\cdot\left[\frac{35\cdot 35}{4}+\frac{31\cdot 31}{4}+\frac{42\cdot 42}{4}+\frac{28\cdot 28}{4}\right]-3\cdot 17 = 1{,}21.$$

При определении критических значений критерия *H* применительно к четырем и более выборкам используют таблицу для критерия хи-квадрат, подсчитав предварительно число степеней свободы для *c=4*. Тогда *v =4-1=3*.

$$H_{кр} = \begin{cases} 7{,}815 \text{ для } \alpha \le 0{,}05 \\ 11{,}345 \text{ для } \alpha \le 0{,}01 \end{cases}.$$

Полученное эмпирическое значение $H_{эмп}$ оказалось существенно меньше критического значения для 5% уровня. Следовательно, принимается нулевая гипотеза.

4**.** Выборка претендентов на должность коммерческого директора в Санкт-Петербургском филиале зарубежной фирмы была обследована с помощью Оксфордской методики экспресс-видеодиагностики, использующей диагностические ролевые игры. Были обследованы 20 мужчин в возрасте от 25 до 40 лет, средний возраст 31,5 года. Оценки производились по 15 значимым, с точки зрения зарубежной фирмы, психологическим качествам, обеспечивающим эффективную деятельность на посту коммерческого директора. Одним из этих качеств была "Авторитетность".

В конце 8-часового сеанса диагностических ролевых игр и упражнений проводился социометрический опрос участников группы, в котором они должны были ответить на вопрос: "Если бы я сам был представителем фирмы, я выбрал бы на должность коммерческого директора: 1).... 2).... 3)...." Участники знали, что каждый их шаг является материалом для диагностики, и что в данном случае, в частности, проверяется, помимо прочего, их способность к объективному суждению о людях. В результате этой процедуры каждый участник получил то или иное количество выборов от других участников, отражающее его социометрический статус в группе претендентов.

Номера испытуемых	Группа 1: 0 выборов (n_1=5)	Группа 2: 1 выбор (n_2=5)	Группа 3: 2-3 выбора (n_3=5)	Группа 4: 4 и более выборов (n_4=5)
1	5	5	5	9
2	5	6	6	9
3	2	7	7	8
4	5	6	7	8
5	4	4	5	7
Суммы	21	28	30	41

Можно ли считать, что группы с разным статусом различаются и по уровню авторитетности, определявшейся независимо от социометрии с помощью экспресс-видеодиагностики?

Решение.

Сформулируем гипотезы.

H_0: Тенденция повышения значений по шкале Авторитетности при переходе от группы к группе (слева направо) случайна.

H_1: Тенденция повышения значений по шкале Авторитетности при переходе от группы к группе (слева направо) неслучайна.

Для решения задачи будем использовать S – критерий тенденций Джонкира.

Чтобы нам было удобнее подсчитывать количества более инверсий (высоких значений), лучше упорядочить значения в каждой группе по их возрастанию.

Группа 1		Группа 2		Группа 3		Группа 4
Индивидуальные значения	Число инверсий	Индивидуальные значения	Число инверсий	Индивидуальные значения	Число инверсий	Индивидуальные значения
2	15	4	10	5	5	7
4	14	5	8	5	5	8
5	11	6	7	6	5	8
5	11	6	7	7	4	9
5	11	7	4	7	4	9
Суммы	62		36		23	

После того, как все индивидуальные значения расположены в порядке возрастания, легко подсчитать, сколько значений справа превышают данное значение слева.

Начнем с крайнего левого столбца. Значение "2" превышают все 15 значений из трех правых столбцов; значение "4" - 14 значений из трех правых столбцов; значение "5" превышают 11 значений из трех правых столбцов.

Расчет для второго столбца производим по тому же принципу. Мы видим, что значение "4" превышают все 10 значений из оставшихся столбцов справа; значение "5" - 8 значений из столбцов справа и т.д.

Сумма всех инверсий составит величину $A=62+36+23=121$, которую нам нужно будет подставить в формулу для подсчета критерия S.

Однако вначале определим максимально возможное значение A, которое мы получили бы, если бы все значения справа были больше значений слева. $B=4\cdot3\cdot25/2=150$.

Эмпирическое значение критерия вычисляется по формуле 23:

$S_{эмп} = 92$.

По таблице Джонкира определяем критические значения $S_{кр}$ для $c=4$, $n=5$:

$$S_{кр} = \begin{cases} 51 \text{ для } \alpha \le 0{,}05 \\ 72 \text{ для } \alpha \le 0{,}01 \end{cases}.$$

$S_{эмп} > S_{кр}$, следовательно H_0 отвергается. Принимается H_1. Тенденция повыше-

ния значений по шкале Авторитетности при переходе от группы к группе не случайна.

Отвечая на вопрос задачи, мы можем сказать, что группы с разным статусом различаются по показателю Авторитетности, определявшемуся независимо от социометрической процедуры.

Задания для домашней работы:

1. В группе слушателей ФПК по педагогике и психологии назрел глухой конфликт между иногородними слушателями и слушателями, проживавшими в Санкт-Петербурге, где и происходили занятия. В курсе психологического практикума по групповой психологии иногородним слушателям было предложено принять на себя роль петербуржцев и участвовать в споре на их стороне. 7 слушателей были протагонистами - активными игроками, перевоплотившимися в петербуржцев, а 7 других суфлировали им, подсказывая реплики и ссылки на те или иные факты. После этого сеанса социодраматической замены ролей участникам был задан вопрос: "Если принять за 100% психологическую дистанцию между Вами и петербуржцами до дискуссии, то на сколько процентов она сократилась или увеличилась после дискуссии?"

Результаты представлены в таблице. Могут ли эти данные использоваться как подтверждение идеи Д. Л. Морено о том, что принятие на себя роли оппонента способствует сближению с ним ?

№ испытуемых	Группа 1: протагонисты (п1=7)	Группа 2: суфлеры (п2=7)
1	75	10
2	30	10
3	25	15
4	10	20
5	30	30
6	20	25
7	50	5

2. В исследовании С.К.Скаковского (1990) изучалась проблема психологических барьеров при обращении в службу знакомств у мужчин и женщин.

В эксперименте участвовали 17 мужчин и 23 женщины в возрасте от 17 до 45 лет (средний возраст 32,5 года). Испытуемые должны были отметить на отрезке точку, соответствующую интенсивности внутреннего сопротивления, которое им пришлось преодолеть, чтобы обратиться в службу знакомств. Длина отрезка, отражающая максимально возможное сопротивление, составляла 100 мм. В таблице приведе-

ны показатели интенсивности сопротивления, выраженные в миллиметрах.

Можно ли утверждать, что мужчинам приходится преодолевать субъективно более мощное сопротивление?

Группа 1 - мужчины (n_1=17)		Группа 2 - женщины (n_2=23)	
1	81	1	70
2	80	2	66
3	73	3	66
4	72	4	63
5	72	5	63
6	69	6	61
7	69	7	60
8	65	8	54
9	65	9	47
10	62	10	43
11	60	11	41
12	54	12	40
13	54	13	39
14	43	14	38
15	30	15	38
16	26	16	35
17	26	17	30
		18	27
		19	25
		20	23
		21	17
		22	10
		23	9

3. В выборке из 28 мужчин-руководителей подразделений крупного промышленного предприятия Санкт-Петербурга перед началом курса тренинга партнерского общения проводилось обследование с помощью 16-факторного личностного опросника Р.Б.Кеттелла. В таблице приведены индивидуальные значения испытуемых по фактору N, отражающему житейскую искушенность и проницательность. Данные представлены в "сырых" баллах и сгруппированы по четырем возрастным группам. Можно ли утверждать, что есть определенная тенденция изменения значений фактора N при переходе от группы к группе?

№ испытуемых	Группа 1: 26-31 год(n_1=7)	Группа 2: 32-37 лет (n_2=7)	Группа 3: 38-42 года (n_3=7)	Группа 4: 46-52 года (n_4=7)
1	2	11	8	11
2	10	7	12	12
3	5	8	14	9
4	8	12	9	9
5	10	12	16	10
6	7	12	14	14
7	12	9	10	13
Суммы	54	71	83	78
Средние	7,71	10,14	11,86	11,14

7. Непараметрические критерии изменения.

Типовые задания.

1. Проводится групповой тренинг. Его задача – выяснить будет ли эффективен данный вариант тренинга для снижения уровня тревожности участников?

Решение. Для решения этой задачи психолог с помощью теста Тейлора дважды выявляет уровень тревожности у 14 участников до и после проведения тренинга.

Результаты измерения приведем в таблице, включив в нее столбец, необходимый для расчета по критерию знаков G.

№ испытуемых	Уровень тревожности «до» тренинга	Уровень тревожности «после» тренинга	Сдвиг
1	30	34	+ 4
2	39	39	0
3	35	26	-9
4	34	33	-1
5	40	34	-6
6	35	40	+5
7	22	25	+3
8	22	23	+1
9	32	33	+1
10	23	24	+1
11	16	15	-1
12	34	27	-7
13	33	35	+2
14	34	37	+3

В столбце, обозначенном словом «Сдвиг», для каждого участника отдельно определяют, насколько изменился его уровень тревожности после проведения тренинга. Сдвиг – это величина разности между уровнями тревожности одного и того же участника «после» и «до» тренинга. Но не наоборот! Величины сдвигов обязательно должны быть даны в соответствующем столбце таблице с учетом знаков.

В критерии знаков по результатам, полученным в столбце таблицы, обозначенном словом «Сдвиг», подсчитываются суммы нулевых, положительных и отрицательных сдвигов. При использовании критерия знаков необходимо учитывать только сумму положительных и отрицательных сдвигов, а сумму нулевых – отбрасывать.

Проведем необходимый подсчет для нашей задачи: общее число (сумма) нулевых сдвигов = 1; общее число (сумма) положительных сдвигов = 8; общее число (сумма) отрицательных сдвигов = 5.

Таким образом, отбросив нулевые сдвиги, получаем 13 ненулевых сдвигов.

При этом подсчет показал, что сдвиги имели место и что большая часть из них положительна.

Напомним, что критерий знаков G предназначен для установления того, как изменяются значения признака при повторном измерении связной выборки: в сторону увеличения или уменьшения. Поэтому, анализируя соотношение положительных и отрицательных сдвигов в нашей задаче, мы проверяем гипотезы:

H_0: После проведения тренинга не наблюдается достоверный сдвиг в сторону уменьшения уровня тревожности участников.

H_1: После проведения тренинга наблюдается достоверный сдвиг в сторону уменьшения уровня тревожности участников.

Сумма сдвигов, получившаяся наименьшей - нетипичный сдвиг: $G_{эмп}=5$.

По таблице G-знаков, в которой приводятся критические величины 5% и 1% уровней значимости данного критерия, определяем критические значения критерия знаков. Поскольку в нашем примере $n = 13$, (это число ненулевых сдвигов), $G_{кр}=3$ для $p=0,05$ и $G_{кр}=1$ для $p=0,01$.

$G_{эмп} > G_{кр}$, следовательно принимаем гипотезу H_1.

2. Проводится с младшими школьниками коррекционная работа по формированию навыков внимания, используя для оценки результатов корректурную пробу. Задача состоит в том, чтобы определить, будет ли уменьшаться количество ошибок внимания у младших школьников после специальных коррекционных упражнений? Для решения этой задачи психолог у 19 детей определяет количество ошибок при выполнении корректурной пробы до и после коррекционных упражнений. В таблице приведены соответствующие экспериментальные данные.

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
До	24	12	42	30	40	55	50	52	50	22	33	78	79	25	28	16	17	12	25
После	22	12	41	31	32	44	50	32	21	34	56	78	23	22	12	16	17	18	25

Решение. Воспользуемся парным критерием Т – Вилкоксона. Данные внесем в расчетную таблицу

№ испытуемых п/п	До	После	Сдвиг (значение разности с учетом знака)	Абсолютные величины разностей	Ранги абсолютных величин разностей	Символ нетипичного сдвига
1	24	22	- 2	2	10,5	
2	12	12	0	0	2	
3	42	41	-1	1	6,5	
4	30	31	+1	1	6,5	*
5	40	32	- 8	8	15	
6	55	44	- 11	11	16	

7	50	50	0	0	2	
8	52	32	-20	20	18	

9	50	32	- 18	18	17	
10	22	21	- 1	1	6,5	
11	33	34	+ 1	1	6,5	*
12	78	56	-22	22	19	
13	79	78	- 1	1	6,5	
14	25	23	-2	2	10,5	
15	28	22	-6	6	13,5	
16	16	12	-4	4	12	
17	17	16	- 1	1	6,5	
18	12	18	+ 6	6	13,5	*
19	25	25	0	0	2	
Сумма					190	26,5

Сформулируем гипотезы:

H_0: Интенсивность сдвигов в сторону уменьшения количество ошибок внимания у младших школьников после специальных коррекционных упражнений не превосходит интенсивности сдвигов в сторону их увеличения.

H_1: Интенсивность сдвигов в сторону уменьшения количество ошибок внимания у младших школьников после специальных коррекционных упражнений превосходит интенсивности сдвигов в сторону их увеличения.

Символом * мы отметили все имеющиеся в таблице нетипичные сдвиги. В нашем случае - это три положительных сдвига. Сумма рангов нетипичных сдвигов и будет искомая величина $T_{эмп}$. В нашем случае эта сумма равна: $T_{эмп}=6,5 + 6,5+ 13,5 = 26,5$.

По таблице определяем критические значения $T_{кр}$ для $n=19$.

Нужная нам строка таблицы Вилкоксона выделена ниже в таблицу:

n	α	
19	0,05	0,01
	53	38

$T_{эмп} < T_{кр}$, следовательно гипотеза H_0 отвергается. Интенсивность сдвигов в сторону уменьшения количество ошибок внимания у младших школьников после специальных коррекционных упражнений превосходит интенсивности сдвигов в сторону их увеличения.

3. Шести школьникам предъявляют тест Равена. Фиксируется время решения каждого задания.

Выясняется вопрос – будут ли найдены статистически значимые различия между временем решения первых трех заданий теста?

№ испытуемых	Время решения первого	Время решения второго за-	Время решения третьего за-

	задания теста в сек.	дания теста в сек.	дания теста в сек.
1	8	3	5
2	4	1	12
3	6	23	15
4	3	6	6
5	7	12	3
6	15	24	12

Для нахождения различий можно было бы воспользоваться двумя предыдущими критериями, но тогда нужно было бы сравнивать между собой данные второго столбца с третьим и четвертым, а также данные третьего столбца с четвертым, т.е. выполнить три однотипных операции. Критерий Фридмана позволяет сразу сравнить между собой три и большее число столбцов, что возможность существенно ускорить процесс решения.

Применение критерия Фридмана требует выполнения следующих операций.

1. Ранжирование экспериментальных данных по строкам таблицы. Обратите внимание, что в этом случае ранжирование проводится не по столбцам (вертикально), т.е. по одному показателю в группе испытуемых, а по строкам (горизонтально), т.е. по всем показателям для одного испытуемого. Выполняя эту операцию в нашей задаче, перепишем еще раз таблицу, добавив в нее необходимые столбцы для значений рангов.

№ испытуемых п/п	Время решения первого задания теста в сек.	Ранги времени решения первого задания теста	Время решения второго задания теста в сек.	Ранги времени решения второго задания теста	Время решения третьего задания теста в сек.	Ранги времени решения третьего задания теста
1	8	3	3	1	5	2
2	4	1	15	3	12	2
3	6	1	23	3	15	2
4	3	1	6	2,5	6	2,5
5	7	2	12	3	3	1
6	15	2	24	3	12	1
Сумма рангов		10		15,5		10,5

2. Суммирование полученных рангов по столбцам таблицы.

В столбце 3 получена сумма рангов равная 10, в столбце 5 – равная 15,5 и в столбце 7 – равная 10,5.

Сформулируем гипотезы.

H_0: Различия во времени решение трех первых заданий теста у шести испытуемых являются случайными.

H_1: Различия во времени решение трех первых заданий теста у шести испытуемых не являются случайными.

Проведем расчет эмпирического значения критерия Фридмана:

$$\chi^2_{r\text{эмп}} = \left(\frac{12}{6 \cdot 3 \cdot 4} \cdot \left(10^2 + 15{,}5^2 + 10{,}5^2\right) \right) - 3 \cdot 6 \cdot 4 = 3{,}08 .$$

По таблице Фридмана определяем величины критических значений $\chi_r^2{}_{кр}$ для числа испытуемых равному 6. Соответствующий блок таблицы Фридмана представлен ниже

n	α	
	0,052	0,012
6	6,33	8,33

Подчеркнем, что с целью стандартизации предъявления табличных значений, критические значения, взятые из таблицы приложения, даны в виде, соответствующем ранее использованному варианту.

Следует особо подчеркнуть, что таблицы для поиска критических значений критерия Фридмана очень специфичны и отличаются от стандартных статических таблиц. Здесь уровни значимости α – даны нетрадиционно, и поэтому каждый раз следует выбирать наиболее близкие значения к 0,05 и 0,01. В нашем случае эти значения составляют 0,052 и 0,012.

$\chi_r^2{}_{эмп} > \chi_r^2{}_{кр}$, следовательно H_0 отклоняется.

Различия во времени решение трех первых заданий теста у шести испытуемых не являются случайными.

4. Социолог высказывает предположение о наличии тенденции: время решения заданий теста будет возрастать по мере увеличения их сложности. Для выявления этой тенденции исследователь сравнивает время решения пяти заданий теста.

№ испытуемых	Время решения 1-го задания.	Ранги времени решения 1-го задания	Время решения 2-го задания.	Ранги времени решения 2-го задания	Время решения 3-го задания.	Ранги времени решения 3-го задания	Время решения 4-го задания.	Ранги времени решения 4-го задания	Время решения 5-го задания.	Ранги времени решения 5-го задания
1	8	3	3	1	5	2	12	4	24	5
2	4	1	15	4	12	2	13	3	35	5
3	6	1	23	5	15	2	20	4	18	3
4	3	1	6	2,5	6	2,5	12	4	43	5
5	7	2	12	4,5	3	1	8	3	12	4,5
6	15	3	24	5	12	2	7	1	22	4
Сумма рангов		11		22		11,5		18		26,5

Решение. Сформулируем гипотезы.

H_0: Тенденция возрастания времени решения заданий теста, по мере увеличения их сложности, является случайной.

H_1: Тенденция возрастания времени решения заданий теста, по мере увеличения их сложности, не является случайной.

Определяем эмпирическое значение критерия

$L_{эмп}=11\cdot1+11,5\cdot2+19\cdot3+22\cdot4+26,5\cdot5=311,5.$

По таблице критических значений находим для $n=6$ и $c=5$:

$$L_{кр} = \begin{cases} 291 \ для\ \alpha = 0,05 \\ 299 \ для\ \alpha = 0,01 \end{cases}.$$

$L_{эмп} > L_{кр}$, следовательно принимается гипотеза H_1.

Задания для домашней работы:

1. В исследовании Г.А.Бадасовой было установлено, что испытуемые по-разному относятся к наказаниям, которые совершают по отношению к их детям разные люди. Например, наказание со стороны самого родителя считается более приемлемым, чем наказание со стороны бабушки и тем более воспитательницы или учительницы. Оценки степени согласия с утверждениями о допустимости телесных наказаний по семибальной шкале в экспериментальной группе приведены в таблице (больший балл соответствует большей степени согласия).

Испытуемые	Условие 1: "Я сам наказываю"	Условие 2: "Бабушка наказывает"	Условие 3: "Учительница наказывает"
1	4	2	1
2	1	1	1
3	5	4	4
4	4	3	2
5	3	3	2
6	4	5	1
7	3	3	1
8	5	5	3
9	6	5	3
10	2	2	2
11	6	3	2
12	5	3	4
13	7	5	4
14	5	5	2
15	5	5	4
16	6	6	4
Суммы	71	60	40

Можно ли говорить о достоверной тенденции в оценках?

8. Анализ данных с помощью критериев согласия

Типовые задания.

1. Определить будет ли удовлетворенность работой на данном предприятии распределена равномерно по следующим альтернативам (градациям): 1 - Работой вполне доволен; 2 - Скорее доволен, чем не доволен; 3-Трудно сказать, не знаю, безразлично; 4-Скорее недоволен, чем доволен; 5 - Совершенно недоволен работой. Для решения этой задачи производится опрос случайной выборки из 65 респондентов (испытуемых) об удовлетворенности работой: «В какой степени Вас устраивает Ваша теперешняя работа?», причем ответы должны даваться согласно вышеозначенным альтернативам.

Альтернативы	1	2	3	4	5
Частота выбора	8	22	14	9	12

Решение. Сформулируем гипотезы.

H_0: распределение выбора альтернатив не отличается от равномерного.

H_1 : распределение выбора альтернатив отличается от равномерного.

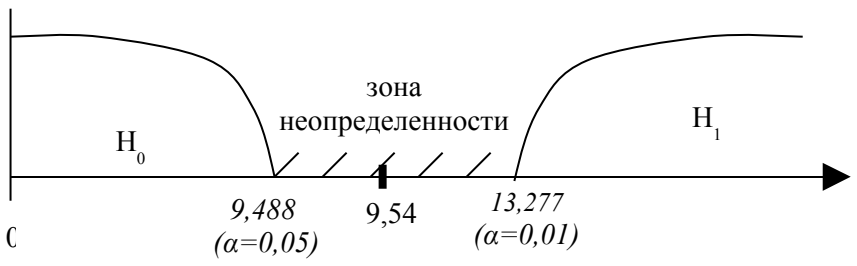
Проверку гипотез проведем с помощью критерия χ^2. Рассчитаем теоретические частоты для равномерного распределения и занесем значения в таблицу

Альтернативы	$f_э$	$f_т$	$f_э - f_m$	$(f_э - f_m)^2$	$\frac{(f_э - f_m)^2}{f_m}$
1	8	13	-5	25	1,92
2	22	13	+9	81	6,23
3	14	13	+ 1	1	0,08
4	9	13	-4	16	1,23
5	12	13	-1	1	0,08
Суммы	65	65	0		$\chi^2_{эмп}$=9,54

Число степеней свободы *v=5-1=4*. По таблице χ^2 находим: $\chi^2_{кр}$=

$$\begin{cases} 9{,}488 \text{ для } \alpha \le 0{,}05 \\ 13{,}277 \text{ для } \alpha \le 0{,}01 \end{cases}.$$

Строим «ось значимости»:



Величина $\chi^2_{эмп}$ попала в зону неопределенности. Можно считать, однако, что полученные различия значимы на уровне 0,05 и принять гипотезу H_1 о различии теоретического и эмпирического распределений.

2. У 267 человек был измерен рост. Будет ли полученное в этой выборке распределение роста близко к нормальному?

Интервалы роста	[156,5-159,5)	[159,5-162,5)	[162,5-165,5)	[165,5-168,5)	[168,5-171,5)	[171,5-174,5)	[174,5-177,5)	[177,5-180,5)	[180,5-183,5]
Число человек	3	9	31	71	82	46	19	5	1

Решение. Сформулируем гипотезы.

H_0: распределение роста не отличается от нормального.

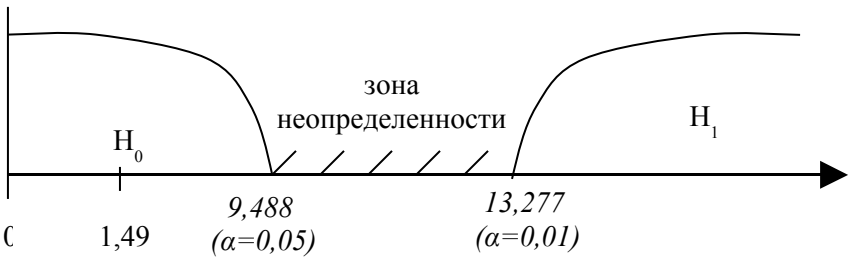
H_1 : распределение роста отличается от нормального.

Проверку гипотез проведем с помощью критерия χ^2. Рассчитаем теоретические частоты. Длина интервала h=3. Так как теоретическая частота в первом и последнем интервалах меньше пяти, т. е. не выполняется ограничение (4) критерия χ^2, объединим первый интервал со вторым, а последний с предпоследним.

Альтернативы	Центры интервалов X_i	Эмпирические частоты $f_э$		$u_i = \frac{x_i - \bar{x}}{\sigma}$	Теоретические частоты $f_m = \frac{nh}{\sigma} \cdot \varphi(u_i)$		$f_э - f_m$	$(f_э - f_m)^2$	$\frac{(f_э - f_m)^2}{f_m}$
1	158	3	12	-2,76	1,74	12,18	-0,18	0,03	0,01
2	161	9		-2,02	10,44				
3	164	31		-1,29	34,25		-3,25	10,56	0,31
4	167	71		-0,55	67,65		3,35	11,22	0,17
5	170	82		0,19	77,30		4,70	22,10	0,29
6	173	46		0,93	51,08		-5,08	25,79	0,50
7	176	19		1,67	19,51		-0,51	0,26	0,01
8	179	5	6	2,41	4,40	5	1	1	0,20
9	182	1		3,15	0,60				
Суммы	1530	267		-	267				$\chi^2_{эмп}$ =1,49

Число степеней свободы рассчитаем как *k-3=7-3=4*.

$\chi^2_{кр} = \begin{cases} 9{,}488 \text{ для } \alpha \le 0{,}05 \\ 13{,}277 \text{ для } \alpha \le 0{,}01 \end{cases}$. Строим «ось значимости»:



Полученная величина эмпирического значения хи-квадрат попала в зону не-

значимости, поэтому необходимо принять гипотезу H_0 об отсутствии различий. Следовательно, есть основания утверждать, что распределения роста близко к нормальному.

3. По данным аттестационной комиссии из 53 студентов энергетического факультета 23 справились на отлично, а из 45 студентов экономического факультета с тем же заданием справились 23 человека. Можно ли утверждать, что различия в успешности решения аттестационной работы экономистами и энергетиками достоверны.

Решение. Сформулируем гипотезы.

H_0: Различия в успешности решения недостоверны.

H_1: Различия в успешности решения достоверны.

Для проверки гипотез будем использовать φ-критерий Фишера.

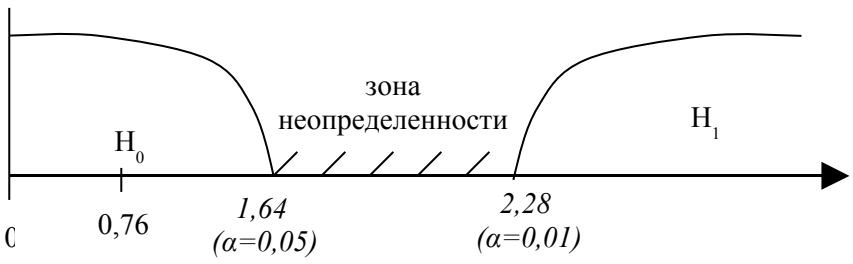
Рассчитаем сначала процентные доли успешных решений для экономического и энергетического факультетов:

*(23/53)*100%=43,4%; (23/45)*100%=51,1%.*

По таблице найдем величины φ_1 и φ_2, соответствующие процентным доля в каждой группе. Для *51,1%* - $\varphi_1=1,593$, а для *43,4%* - $\varphi_2=1,438$. Эмпирическое значение $\varphi_{эмп}=0,76$.

Определяем критические значения критерия:

$$\varphi_{кр}=\begin{cases}1{,}64 \text{ для } \alpha \le 0{,}05\\ 2{,}28 \text{ для } \alpha \le 0{,}01\end{cases}.$$ Строим «ось значимости»:



Так как эмпирическое значение попала в зону незначимости, то необходимо принять гипотезу H_0. Иными словами, различия в успешности решения аттестационной работы экономистами и энергетиками не достоверны.

Задания для домашней работы.

1. Используя критерий Пирсона, при уровне значимости 0,05, проверить, согласуется ли гипотеза о нормальном распределении генеральной совокупности X с эмпирическим распределением выборки объема n=200, представленным в таблице

x_i	5	7	9	11	13	15	17	19	21
n_i	15	26	25	30	26	21	24	20	13

2. Используя критерий Пирсона, при уровне значимости 0,05, проверить, согласуется ли гипотеза о нормальном распределении генеральной совокупности X с заданным эмпирическим распределением.

X_i; X_{i+1}	-20;-10	-10;0	0;10	10;20	20;30	30;40	40;50
n_i	20	47	80	89	40	16	8

3. Дан ряд распределения значений амплитуды установившихся колебаний руки человека. Проверить отличается ли данное распределение от нормального.

Среднее значение интервала	41,5	45,5	49,5	53,5	57,5	61,5	65,5	69,5	73,5	77,5	81,5
Частость	0,010	0,015	0,045	0,125	0,145	0,290	0,185	0,110	0,065	0,000	0,010

4. Одинаков ли уровень профессиональной ориентации на экономическом и социальном факультетах, если из закончивших экономический факультет 98 человек, стали работать по специальности 64 человека, а из закончивших факультет социальных наук 63 человек – 39 стали работать по специальности.

5. Психолог сравнивает два эмпирических распределения, в каждом из которых было обследовано 70 человек по тесту интеллекта. Различаются ли между собой эти два распределения?

Уровни интеллекта	60	70	80	90	100	110	120	130	140
1-е распределение	1	3	8	17	26	7	5	2	1
2-е распределение	0	1	2	19	20	13	10	4	1

7. Используя критерий Колмогорова – Смирнова проверить, отличается ли приведенное в таблице распределение от равномерного.

x_i	1	2	3	4	5	6
n_i	18	23	15	21	25	18

8. В проективной методике Х. Хекхаузена (модификация ТАТ) испытуемому последовательно предъявляются 6 картин. Всякий раз он сначала рассматривает картину в течение 20 сек, а затем, в течение 5 минут, пишет по ней рассказ, стараясь, в соответствии с инструкцией, проявить "максимум фантазии и воображения".

После того, как испытуемый закончит писать первый рассказ, ему предъявляется вторая картина, и т. д.

В данном исследовании разным испытуемым картины предъявлялись в разном

порядке, так что каждая картина оказывалась первой, второй, третьей и т.д. примерно одинаковое количество раз.

При обследовании 113 человек в возрасте от 20 до 35 лет (средний возраст 23,2 года, 67 мужчин, 46 женщин) было установлено, что в рассказах по картинам с условными названиями "Преподаватель и ученик" и "Мастер измеряет деталь" словесные формулировки, отражающие "боязнь неудачи", встречаются гораздо чаще, чем в рассказах по другим картинам, в особенности по картине "Улыбающийся юноша"

Название картины		Количество вербальных реакций, отражающих “надежду на успех”		Количество вербальных реакций, отражающих “боязнь неудачи”		Всего
1	"Мастер измеряет деталь"	Л	106	138	Б	244
2	"Преподаватель и ученик"	В	102	180	Г	282
3	"В цехе у машины"	А	108	34	Е	142
4	"У двери директора"	Ж	50	87	З	137
5	"Человек в бюро"	И	99	57	К	156
6	"Улыбающийся юноша"	Л	115	20	м	135
Всего		580		516		1096

Можно ли утверждать, что картины методики обладают разной побудительной силой в отношении мотивов: а)"надежда на успех"; б)"боязнь неудачи"?

Как следует из таблицы, нет почти ни одной картины, которая в равной мере стимулировала бы мотив "надежда на успех" и мотив "боязнь неудачи". Можно ли считать стимульный набор методики Хекхаузена неуравновешенным по направленности воздействия.

9. В процессе проведения транзактно-аналитических сессий установлено, что запреты на «психологические поглаживания» встречаются с неодинаковой частотой. Например, многие участники тренинга признают у себя запрет “Не проси психологических поглаживания у других людей”, а запрет “Не давай психологических поглаживаний самому себе” встречается гораздо реже.

Можно ли считать, что распределение запретов не является равномерным? Можно ли утверждать, что запрет "Не проси" встречается достоверно чаще остальных?

Запрет		Частота	Доля по отношению к общему количеству
1	Не давай психологических поглаживаний	44	15,66%

2	Не принимай психологических поглаживаний	45	16,01%
3	Не проси психологических поглаживаний	98	34.88%
4	Не отказывайся от психологических поглаживаний, даже если они тебе не правятся	58	20.64%
5	Не давай психологических поглаживании самому себе	36	12,81%
Всего		281	100,00%

9. **Таблицы сопряженности и их анализ**; вычисление эмпирических и теоретических частот.

Типовые задания.

1. По данным диагностики темперамента, с помощью опросника Г. Айзенка, и диагностики профессиональных предпочтений, с помощью опросника ДДО Климова, 93 подростков определить влияет ли тип темперамента на профессиональные предпочтения.

Профессиональные предпочтения		Человек – техника	Человек - знаковая система	Человек – человек	Всего
Тип темперамента	№ столбца / № строки	1	2	3	-
Холерик	1	1	5	15	21
Сангвиник	2	13	5	9	27
Меланхолик	3	2	17	3	22
Флегматик	4	16	3	4	23
Итого	-	32	30	31	93

Решение. Сформулируем гипотезы.

H_0: Тип темперамента не влияет на профессиональные предпочтения.

H_1: Тип темперамента влияет на профессиональные предпочтения.

Для проверки гипотез будем использовать критерий хи-квадрат и расчетную формулу ().

Вычислим сначала теоретические частоты по формуле 31:

$f_{m11}=(21*32)/93=7,2;$ $f_{m12}=(21*30)/93=6,8;$ $f_{m13}=(21*31)/93=7;$ $f_{m21}=(27*32)/93=9,3;$ $f_{m22}=(27*30)/93=8,7;$ $f_{m23}=(27*31)/93=9;$ $f_{m31}=(22*32)/93=7,6;$ $f_{m32}=(22*30)/93=7,1;$ $f_{m33}=(22*31)/93=7,3;$ $f_{m41}=(23*32)/93=7,9;$ $f_{m42}=(23*30)/93=7,4$ $f_{m43}=(23*31)/93=7,7.$

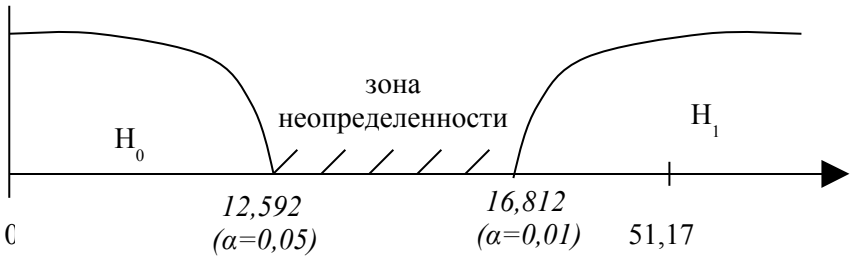
Дальнейшие расчеты оформим в таблице.

Альтернативы	$f_э$	$f_т$	$f_э - f_m$	$(f_э - f_m)^2$	$\frac{(f_э - f_m)^2}{f_m}$
n11n	1	7,2	-6,2	38,44	5,34
n12	5	6,8	-1,8	3,24	0,48

n13	15	7	8	64	9,14
n21	13	9,3	3,7	13,69	1,47
n22	5	8,7	-3,7	13,69	1,57
n23	9	9	0	0	0
n31	2	7,6	-5,6	31,36	4,13
n32	17	7,1	9,9	98,01	13,80
n33	3	7,3	-4,3	18,49	2,53
n41	16	7,9	8,1	65,61	8,31
n42	3	7,4	-4,4	19,36	2,62
n43	4	7,7	-3,7	13,69	1,78
Суммы	93	93	0	-	$\chi^2_{эмп}$ =51,17

Число степеней свободы будет равно *v=(4-1)*(3-1)=6*.

По таблице находим: $\chi^2_{кр}=\begin{cases}12{,}592 \text{ для } \alpha \le 0{,}05\\ 16{,}812 \text{ для } \alpha \le 0{,}01\end{cases}$. Строим «ось значимости»:



Полученная эмпирическая величина критерия хи-квадрат попала в зону значимости. Иными словами следует принять гипотезу H_1 о том, что тип темперамента влияет на профессиональные предпочтения.

Задания для домашней работы:

1. По данным наблюдения определить, имеется ли у людей сопряженность между цветом волос и цветом глаз.

	светлые	русые	черные	рыжие
голубые	177	71	17	14
серые	95	119	75	25
карие	12	44	23	8

2. В социально-психологическом исследовании стереотипов мужественности Н. В. Стан (1992) выборке из 31 женщин с высшим образованием в возрасте от 22 до 49 лег (средний возраст 35 дет) предъявлялись напечатанные на отдельных карточках перечни качеств, характеризующих один из четырех типов мужественности: мифологический, национальный, современный и религиозный. Испытуемым предлагалось внимательно ознакомиться с предложенными описаниями и выбрать из них то, которое в большей степени соответствует их представлению об идеальном мужчине. Затем испытуемым предлагалось выбрать одну из 3 оставшихся карточек, а затем одну из двух оставшихся.

Тип мужественности	Эмпирические позиции				Всего
	1	2	3	4	
1. *Мифологический тип:* Мощный сильный, стройный, ловкий, бесстрашный, гордый, непокорный, уверенный, дерзкий, непреклонный, вспыльчивый, гневный, борец.	2 А	6 Б	4 В	19 Г	31
2. *Национальный тип:* Ловкий, решительный, сдержанный, великодушный, преданный, открытый, бесхитростный, милосердный, уверенный, честный, доверчивый, защитник.	19 Д	4 Е	7 Ж	1 З	31
3. *Современный тип:* Сильный, властный, сдержанный, уверенный, рассудочный, постоянный, агрессивный, практичный, эрудированный, самостоятельный, решительный, деятельный, энергичный, волевой	7 И	10 К	12 Л	2 М	31
4. *Религиозный тип:* Мягкий, миролюбивый, спокойный, кроткий, уступчивый, искренний, внимательный, выносливый, терпеливый, чувствительный	3 Н	11 О	8 П	9 Р	31
Всего	31	31	31	31	124

Различаются ли распределения предпочтений, выявленные по каждому из 4-х типов, между собой? Можно ли утверждать, что предпочтение отдается какому-то одному или двум из типов мужественности? Наблюдается ли какая-либо групповая тенденция предпочтений?

3. В выборке студентов факультета психологии Санкт-Петербургского университета с помощью известного "карандашного" теста определялось преобладание правого или левого глава в прицельной способности глаз. Совпадают ли эти данные с результатами обследования 100 студентов медицинских специальностей.

	Количество испытуемых с преобладанием левого глаза	Количество испытуемых с преобладанием правого глаза	Суммы
Студенты-психологи	7	8	14
Студенты -медик и	19	81	100
Суммы	25	89	114

10. **Измерение связи между переменными в интервальной шкале.**

Типовые задания.

1. 20 школьником были даны тесты на наглядно-образное и вербальное мышление. Измерялось среднее время решения заданий теста в секундах. Психолога интересует вопрос: существует ли взаимосвязь между временем решения этих задач.

№ испытуемого	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Среднее время решения наглядно-образных задач	19	32	33	44	28	35	39	39	44	44	24	37	29	40	42	32	48	42	33	47
Среднее время решения вербальных	17	7	17	28	27	31	20	17	35	43	10	28	13	43	45	24	45	26	16	26

Решение.

Сформулируем гипотезы:

H_0: связь между временем решения наглядно-образных и вербальных задач статистически незначима.

H_1: связь между временем решения наглядно-образных и вербальных задач статистически значима.

Обозначим переменные *x*-среднее время решения наглядно-образных задач, переменная *y*-среднее время решения вербальных заданий тестов.

Для удобства расчеты будем проводить в таблице

№	x_i	y_i	$x_{i*} y_i$	x_i^2	y_i^2
1	19	17	323	361	289
2	32	7	224	1024	49
3	33	7	561	1089	289
4	44	28	1232	1936	784
5	28	27	756	784	729
6	35	31	1085	1225	961
7	39	20	780	1521	400
8	44	17	663	1521	289
9	44	35	1540	1936	1225
10	44	43	1892	1936	1849
11	24	10	240	576	100
12	37	28	1036	1369	784
13	29	13	377	841	169
14	40	43	1720	1600	1849
15	42	45	1890	1764	2025
16	32	24	768	1024	5760
17	48	45	2160	2304	2025
18	42	26	1092	1764	676
19	33	16	528	1089	256
20	47	26	1222	2209	676
Сумма	731	518	20089	27873	16000

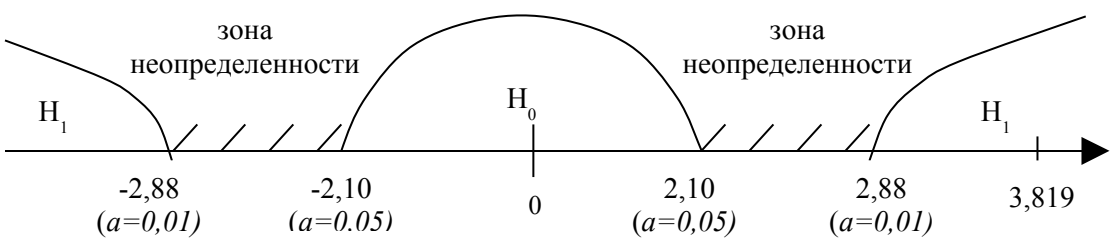
Рассчитаем эмпирическую величину коэффициента корреляции

$$r_{xy\ эмп} = \frac{20 * 20089 - 731 * 518}{\sqrt{(20 * 27873 - 731 * 731) * (20 * 16000 - 518 * 518)}} = 0{,}669,$$

$$t_{эмп} = 0{,}669 * \sqrt{\frac{20 - 2}{1 - 0{,}669^2}} = 3{,}819.$$

Определим критические значения $t_{кр}$ для числа степеней свободы $k=n-2=18$.

$$t_{кр} = \begin{cases} 2{,}10 & для\ \ \alpha \le 0{,}05 \\ 2{,}88 & для\ \ \alpha \le 0{,}01 \end{cases}.$$ Строим «ось значимости»:



Ввиду того, что величина $t_{эмп}$ попало в зону значимости – гипотеза H_0 отвергает и принимается гипотеза H_1. Иными словами, связь между временем решения наглядно-образных и вербальных задач статистически значима на 1% уровне и положительна. То есть увеличению времени решения наглядно-образных задач соответствует увеличение времени вербальных задач.

2. Психолог у 8 подростков сравнивает баллы по третьему, математическому, субтесту теста Векслера и оценки по алгебре. Данные представлены в таблице

№ испытуемого	1	2	3	4	5	6	7	8	сумма
Баллы по тесту Векслера	8	18	18	10	16	10	8	14	102
Оценки по алгебре	2	3	4	5	4	4	3	5	30

Связана ли успешность решения субтеста Векслера с оценкой по алгебре? Связаны ли оценки по алгебре с успешностью решения третьего субтеста Векслера?

Решение. Обозначим переменные: x -баллы полученные испытуемыми по третьему субтесту теста Векслера; y - оценки по алгебре.

Коэффициент линейной корреляции, рассчитанный по формуле (9) $r_{xy}=0,244$ (проверить самостоятельно). Этот коэффициент незначим и, следовательно, линейной связи x и y нет. Нужно выяснить существует ли между этими переменными другой тип связи?

Расставим по порядку x от меньшей к наибольшей.

x	8	8	10	10	14	16	18	18
y	2	3	4	5	5	4	3	4

Определим частоты переменной x.

f_x	2	2	1	1	2
x_i	8	10	14	16	18

Подсчитаем арифметические частные для переменной y по отношению к переменной x.

f_x	2	2	1	1	2
x_i	8	10	14	16	18
$\overline{y}_x$	$\frac{2+3}{2}$	$\frac{4+5}{2}$	5	4	$\frac{3+4}{2}$

Аналогично для переменой y.

У	2	3	3	4	4	4	5	5
Х	8	8	18	10	16	18	10	14

f_y	1	2	3	2
y_i	2	3	4	5

$\bar{x}_y$	8	$\frac{8+18}{2}$	$\frac{10+16+18}{3}$	$\frac{10+14}{2}$

Теперь подсчитаем общие средние.

$$\bar{x} = \frac{102}{8} = 12{,}75\,;\ \bar{y} = \frac{30}{8} = 3{,}75$$

Найдем корреляционные отношения:

$$h_{yx} = \sqrt{\frac{2(2{,}5-3{,}75)^2 + 2(4{,}5-3{,}75)^2 + 1(5-3{,}75)^2 + 1(4-3{,}75)^2 + 2(3{,}5-3{,}75)^2}{1(2-3{,}75)^2 + 2(3-3{,}75)^2 + 3(4-3{,}75)^2 + 2(5-3{,}75)^2}} =$$

$$= \sqrt{\frac{6}{7{,}5}} = \sqrt{0{,}8} = 0{,}89$$

$$h_{xy} = \sqrt{\frac{1(8-12{,}75)^2 + 2(13-12{,}75)^2 + 3(14{,}7-12{,}75)^2 + 2(12-12{,}75)^2}{2(8-12{,}75)^2 + 2(10-12{,}75)^2 + (14-12{,}75)^2 + (16-12{,}75)^2 + (18-12{,}75)^2}} =$$

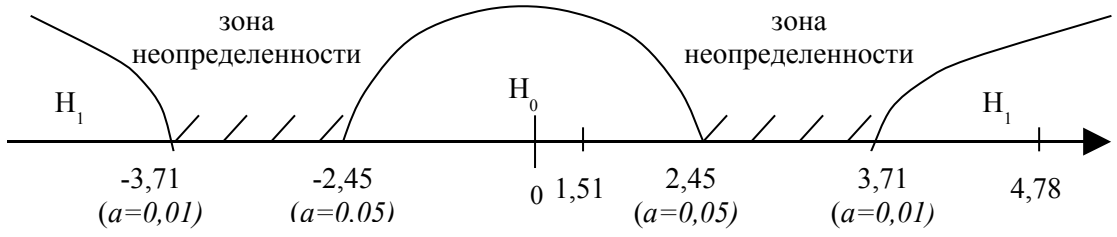
$$= \sqrt{\frac{35{,}22}{127{,}5}} = \sqrt{0{,}276} = 0{,}525$$

Проверим значимость показателей

$$t_{эмпух} = |0{,}89| * \sqrt{\frac{8-2}{1-0{,}89^2}} = 4{,}78\,,\ t_{эмпху} = |0{,}525| * \sqrt{\frac{8-2}{1-0{,}525^2}} = 1{,}51.$$

k=n-2=8-2=6.

$t_{кр} = \begin{cases} 2{,}45 & для\ \ \alpha \le 0{,}05 \\ 3{,}71 & для\ \ \alpha \le 0{,}01 \end{cases}$. Строим «ось значимости»:



Показатель h_{yx} значим, а h_{xy} – незначим. Таким образом можно сделать вывод о том, что в данном случае есть значимое влияние *Y* на *X*, а обратное влияние *X* на *Y* незначимо. Следовательно, хорошее значение алгебры влияет на эффективность с третьим субтестом теста Векслера, и, напротив, успешное решение третьего субтеста Векслера никак не сказывается на успешности овладения учащимися алгеброй.

Задания для домашней работы:

1. Опрос случайно выбранных 10 студентов, проживающих в общежитии уни-

верситета, позволяет выявить зависимость между средним баллом по результатам предыдущей сессии и числом часов в неделю, затраченных студентами на самостоятельную подготовку.

Средний балл	4,6	4,3	3,8	3,8	4,2	4,3	3,8	4,0	3,1	3,9
Число часов	25	22	9	15	15	30	20	30	10	17

2. Рассчитайте выборочный коэффициент линейной корреляции Пирсона, проверти его значимость при $\alpha = 0{,}05$.

3. В плане комплексного исследования личности у студентов психологического факультета определялся социометрический статус на курсе (У) и в своей учебной группе (Х). Результаты представлены в таблице. Имеется ли связь между показателями Х и У.

Х	1	-9	22	-11	-6	-0,5	5	9	0	2	11	15	-8	0	9
У	48	-9	100	-1	14	14	14	11	-28	46	42	27	19	83	45

4. При изучении амплитудно-частотных характеристик руки человека в условиях направленного уменьшению и увеличению амплитуды совместно регистрировались значения полупериодов (У) и амплитуд (Х) колебаний руки. Результаты одного из испытуемых представлены в таблице

Х	15	30	45	60	75	90	105	120	135	150	165	180
У	3,3	4,0	4,7	5,4	6,1	6,8	7,5	8,2	8,9	9,6	10,3	11,0

Проверить значимость корреляции между Х и У.

5. У студентов первого курса в комплексном исследовании личности были получены оценки (в баллах) Нейротизма по Айзенку и эмоциональной экспансивности, представлены в таблице

Нейротизм	18	20	18	22	9	12	13	16	14	16	16	17	21	22	7	23
Эмоциональной экспансивность	3	-15	15	-1	-26	11	2	10	-4	13	-17	9	-27	25	61	33

6. Выяснить имеется ли статистически значимая связь между показателями нейротизма и эмоциональной экспансивности.

7. Оценить силу корреляционной связи по выборочному корреляционному отношению.

х	2	2	2	3	3	3	3	3	3	3	3	5	5	5	5	5	5
У	25	25	25	45	45	45	45	110	110	110	110	45	110	110	110	110	110

8. Можно ли говорить о наличии криволинейной корреляционной связи между

временем выполнения тестовых заданий и количеством верных ответов, за соответствующий промежуток времени.

Время (мин)	10	20	30	40	50	60	70	80
Количество верных ответов (за последние 10 мин)	3	4	5	5	4	4	3	3

11. Непараметрические коэффициенты корреляции.

Типовые задания.

1. Психолог выясняет, как связаны между собой индивидуальные показатели готовности к школе, полученные показатели готовности к школе, полученные до начала обучения в школе у 11 первоклассников и их средняя успеваемость в конце учебного года. Проранжированные значения показателей школьной готовности и итоговые показатели успеваемости в конце года у этих же учащихся в среднем приведены в таблице

№ учащегося	1	2	3	4	5	6	7	8	9	10	11
Ранги показателей школьной готовности	3	5	6	1	4	11	9	2	8	7	10
Ранги среднегодовой успеваемости	2	7	8	3	4	6	11	1	10	5	9

Решение. Сформулируем гипотезы:

H_0: связь между показателями готовности к школе и средней успеваемостью первоклассников статистически незначима.

H_2: связь между показателями готовности к школе и средней успеваемостью первоклассников статистически значима.

Вычислим разность между рангами и их квадраты.

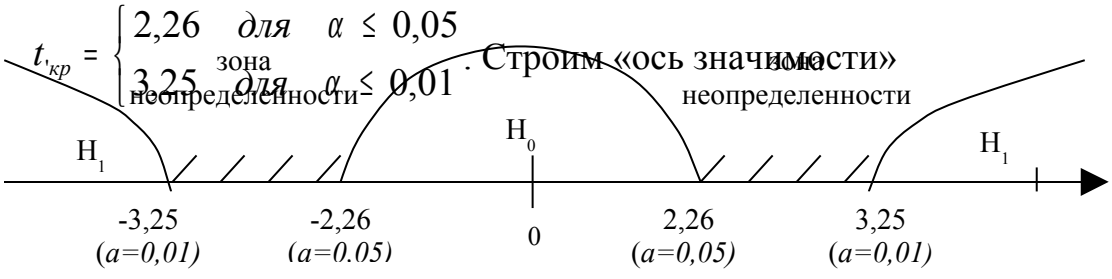
№	1	2	3	4	5	6	7	8	9	10	11	Сумма
d_i	1	-2	-2	-2	0	5	-2	1	-2	2	1	
d_i^2	1	4	4	4	0	25	4	1	4	4	1	52

Подставляем полученные данные в формулу (10), и производим расчет. Получаем:

$$\rho_{эмп} = 1 - \frac{6 * 52}{11 * (11^2 - 1)} = 0{,}76. \quad \text{найдем} \quad t_{эмп} = |0{,}76| * \sqrt{\frac{11 - 2}{1 - 0{,}76^2}} = 3{,}25.$$

При числе степеней свободы $k=n-2=11-2=9$ находим по таблице критические значения t-критерия Стьюдента.

$$t_{кр} = \begin{cases} 2{,}26 & для \quad \alpha \le 0{,}05 \\ 3{,}25 & для \quad \alpha \le 0{,}01 \end{cases}$$. Строим «ось значимости»



Полученное эмпирическое значение t-критерия совпало с критическим для уровня значимости в 1%. Следовательно, можно утверждать, что показатели школьной готовности и итоговые оценки первоклассников связаны положительной корреляцией - иначе говоря, чем выше показатель школьной готовности, тем лучше учится первоклассник. В терминах статистических гипотез психолог должен отклонить нулевую гипотезу и принять альтернативную.

2. Психолог, используя тест умственного развития (ШТУР) проводит исследование интеллекта у 12 учащихся 9 класса. Одновременно с этим он просит учителей литературы и математики провести ранжирование этих же учащихся по показателям умственного развития. Определить, как связаны между, собой объективные показатели умственного развития (данные ШТУРа) и экспертные оценки учителей.

Решение. Данные таблицы дополним столбцами для расчетов.

№ учащихся	Ранги тестирования с помощью ШТУРа	Экспертные оценки учителей по математике	Экспертные оценки учителей по литературе	d_{i1} (второй столбец минус третий)	d_{i2} (второй столбец минус четвертый)	d_{i1}^2	$(d_{i2})^2$
1	6	5	5	-1	1	1	1
2	7	10	8	-3	-1	9	1
3	4	8	7	-4	-3	16	9
4	5	4	11	1	-6	1	36
5	9	6	3	-3	6	9	36
6	12	8	6	4	6	16	36
7	2,5	2	11	0,5	-8,5	0,25	72,25
8	2,5	3	11	-0,5	-8,5	0,25	72,25
9	10	8	1	2	9	4	81
10	8	11	3	-3	5	9	25
11	11	12	3	-1	5	1	64
12	1	1	9	0	8	0	64
Сумма	78	78	78	0	-8	66,5	497,5

Сформулируем гипотезы:

H_0: $\rho_1 = 0$ (связь величины рангов по тесту ШТУРа с экспертными оценками учителей математики статистически не значима).

H_1: $\rho_1 \neq 0$ (связь величины рангов по тесту ШТУРа с экспертными оценками учителей математики статистически значима).

H_0: $\rho_2 = 0$ (связь величины рангов по тесту ШТУРа с экспертными оценками учителей литературы статистически не значима).

H_1: $\rho_2 \neq 0$ (связь величины рангов по тесту ШТУРа с экспертными оценками учителей литературы статистически значима).

Во втором столбце есть два одинаковых ранга, следовательно, величина поправки (по формуле ( )) $D_1 = \frac{2^3 - 2}{12} = 0{,}5$.

В третьем и четвертом столбцах по три одинаковых ранга:

$$D_3 = \frac{3^3 - 3}{12} = 2 \quad \text{и} \quad D_4 = \frac{3^3 - 3}{12} = 2.$$

Считаем первый ранговый коэффициент $\rho_{эмп2}$ с учетом поправок D_1, D_3 и D_4:

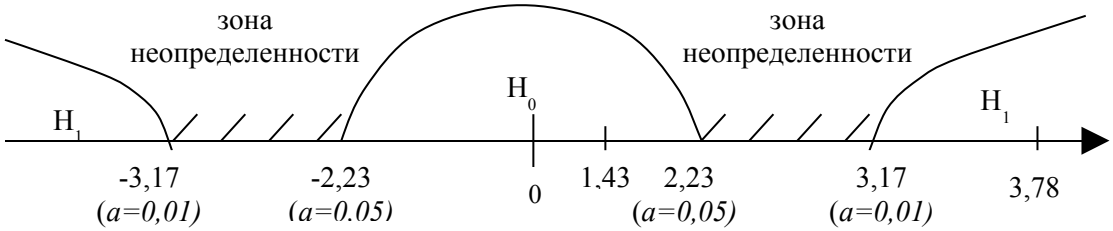
$$\rho_{эмп2} = 1 - \frac{6 * 497{,}5 + 0{,}5 + 2}{12 * 143} = 1 - 1{,}419 = -0{,}412.$$

Найдем эмпирические значения t- критерия Стьюдента

$$t_{эмп1} = 0{,}767 * \sqrt{\frac{12 - 2}{1 - 0{,}767^2}} = 3{,}78, \; t_{эмп2} = |-0{,}412| * \sqrt{\frac{12 - 2}{1 - (-0{,}412)}} = 1{,}43.$$

Критические значения находим при уровне значимости $k=n-2=12-2=10$

$t_{кр} = \begin{cases} 2{,}23 & для \quad \alpha \leq 0{,}05 \\ 3{,}17 & для \quad \alpha \leq 0{,}01 \end{cases}$. Отложим значения на «оси значимости»



В первом случае H_0 отвергается и принимается H_1: связь величины рангов по тесту ШТУРа с экспертными оценками учителей математики статистически значима.

Во втором случае H_0 принимается, т.е. оценки учащихся по тесту ШТУРа не связаны с экспертными оценками учителей по литературе.

3. Выяснить связана ли между собой семейное положение (0-холост, 1-женат) и успешность учебы (0-успешно, 1-неуспешно) студентов мужчин.

№ испытуемого	1	2	3	4	5	6	7	8	9	10	11	12
Семейное положение	0	1	0	0	1	1	0	1	0	0	0	1
Успешность обучения	0	1	1	0	1	0	0	1	0	1	0	1

Сформулируем гипотезы:

H_0: семейное положение и успешность учебы студентов мужчин статистически не значима $(\varphi = 0)$.

H_1: семейное положение и успешность учебы студентов мужчин статистически значима $(\varphi \neq 0)$.

По условию задачи *N=12*.

Доля единиц в переменной *x:* $p_x = \frac{5}{12} = 0{,}4167$.

Доля нулей в переменной *x:* $1\text{-}p_x = 1\text{-}0{,}4167 = 0{,}5833$.

Доля единиц в переменной *y:* $p_y = \frac{6}{12} = 0{,}5$.

Доля нулей в переменной *y:* $1\text{-}p_y = 1\text{-}0{,}5 = 0{,}5$.

Доля единиц по обеим переменным*:* $p_{xy} = \frac{4}{12} = 0{,}3333$

Рассчитаем коэффициент ассоциации:

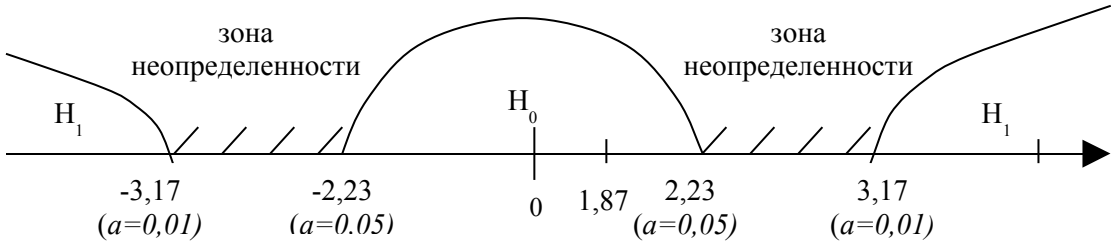
$$\varphi_{эмп} = \frac{0{,}3333 - 0{,}41667 * 0{,}5}{\sqrt{0{,}4167 * 0{,}5833 * 0{,}5 * 0{,}5}} = 0{,}507$$

Эмпирическое значение t- критерия Стьюдента равно:

$$t_{эмп} = |0{,}507| * \sqrt{\frac{12 - 2}{1 - 0{,}507^2}} = 1{,}87$$

Число степеней свободы в нашем случае будет равно: *k=n-2=12-2=10.* Соответствующие критические значения:

$t_{кр} = \begin{cases} 2{,}23 & для \quad \alpha \leq 0{,}05 \\ 3{,}17 & для \quad \alpha \leq 0{,}01 \end{cases}$. Строим «ось значимости»:



Гипотеза H_1 отклоняется и принимается гипотеза H_0: связь между семейным положением и успеваемостью обучения студентов мужчин отсутствует.

4. Психолог просит супругов проранжировать семь личностных черт, имеющих

определяющее значение для семейного благополучия. Задача заключается в том, чтобы определить, в какой степени совпадают оценки супругов по отношению к ранжированию качеств.

Черты личности	муж	жена
ответственность	7	1
общительность	1	5
сдержанность	3	7
выносливость	2	6
жизнерадостность	5	4
терпеливость	4	3
решительность	6	2

Решение. Так как данные задачи нумерованы по ранговой школе, то будем использовать коэффициент корреляции τ Кендала.

Сформулируем гипотезы

$H_0: \tau = 0,$

$H_1: \tau \neq 0.$

Упорядочим оценки мужа по возрастанию рангов. При этом оценки жены также переместятся.

муж	жена	Инверсии
1	5	4
2	6	4
3	7	4
4	3	2
5	4	2
6	2	1
7	1	0
Сумма		17

Подсчитаем количество инверсии для оценок жены. «5» встречается четыре ранга меньших 5. Ниже «6» - четыре ранга меньших 6. Ниже «7» - 4 ранга меньших 7. Ниже «3» - 2 ранга меньших 3. Ниже «4» - 2 ранга меньших 4. Ниже «2» - 1 ранг меньших 2.

Таким образом, число инверсий Q=17.

Рассчитаем значение τ

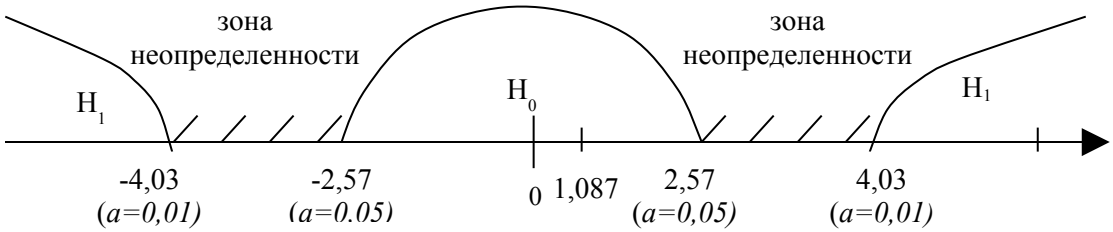
$$\tau_{эмп} = 1 - \frac{4*17}{7*(7-1)} = -0,619$$

$$t_{эмп} = |-0,619| * \sqrt{\frac{7-2}{1-(-0,619)^2}} = 1,087$$

Число степеней свободы: $к=n-2=7-2=5$.

Критические значения t-критерия Стьюдента:

$t_{кр} = \begin{cases} 2{,}57 & для \quad \alpha \le 0{,}05 \\ 4{,}03 & для \quad \alpha \le 0{,}01 \end{cases}$. Строим «ось значимости»:



Значение $t_{кр}$ попало в зону незначимости. H_1 отклоняется и принимается гипотеза H_0 о том, что коэффициент корреляции τ Кендалла достоверно не отличается от нуля. Иными словами, согласованности между мужем и женой в оценке значимых для семейного благополучия личностных черт нет.

5. Данные обследования 15 подростков разного пола по методике Айзенка приведены в таблице

№ испытуемого	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Пол	1	0	1	1	0	1	0	0	1	1	1	0	1	1	0
IQ	102	110	86	90	120	78	95	103	105	93	123	89	109	100	105

Код "1" - юноша, код "0" - девушка.

Имеются ли тендерные различия в показателях интеллекта?

Решение. Обозначим переменные: x - пол, y - IQ. Так как x измерена по дихотомической школе, а y по интервальной, то будем использовать бисериальный коэффициент корреляции.

Сформулируем гипотезы:

H_0: $R_{бис}=0$,

H_1: $R_{бис} \neq 0$.

Число единиц в переменной равно $n_1=9$, а число нулей - $n_0=6$. $N=n_1+n_0=15$ - общее число испытуемых.

Найдем среднее значение IQ отдельно для юношей и девушек:

$$\overline{x}_1 = \frac{102+86+90+78+105+93+123+109+100}{9} = 98{,}4$$

$$\overline{x}_0 = \frac{110+120+95+103+89+105}{6} = 103{,}67$$

Среднее квадратическое отклонение найдем по формуле:

$$\sigma_y = \sqrt{\frac{\sum_i (y_i - \bar{y})^2}{N}} = 12{,}374$$

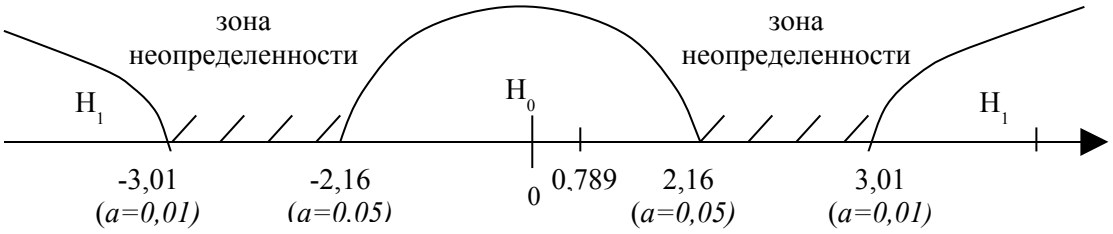
Вычислим $R_{бис\ эмп} = \frac{98{,}4 - 103{,}67}{12{,}374} * \sqrt{\frac{9*6}{15*14}} = -0{,}216$

Число степеней свободы *к=n-2=15-2=13*

$$t_{эмп} = |-0{,}216| * \sqrt{\frac{15-2}{1-(-0{,}216)^2}} = 0{,}789$$

Критические значения Критерия Стьюдента равны:

$t_{кр} = \begin{cases} 2{,}16 & для\ \ \alpha \le 0{,}05 \\ 3{,}01 & для\ \ \alpha \le 0{,}01 \end{cases}$. Строим «ось значимости»:



Результат попал в зону незначимости. Поэтому принимается гипотеза H_0. Таким образом, тендерных различий по интеллекту на данной выборке испытуемых не обнаружено.

6. Пятнадцать подростков разного пола были проранжированы учителем литературы по степени выраженности вербальных способностей. Данные представлены в таблице (код «1» - юноша, код «0» - девушка). Существуют ли тендерные различия в вербальных способностях.

№ испытуемого	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
пол	1	0	1	1	0	1	0	0	1	1	1	0	1	1	0
Ранги вербальных способностей	1	10	6	9	15	7	8	13	4	3	5	11	12	2	14

Решение. Обозначим переменные: *x* - пол, *y* - ранги вербальных способностей. Будем использовать рангово-бисериальный коэффициент корреляции.

Сформулируем гипотезы:

H_0: $R_{rb}=0;$

H_1: $R_{rb} \neq 0$.

Найдем среднее значения рангов отдельно для юношей и девушек:

$$\bar{x}_1 = \frac{1+6+9+4+3+5+12+2}{9} = 5{,}44 \qquad \bar{x}_0 = \frac{10+15+8+13+11+14}{6} = 11{,}83$$

Вычислим $R_{rb\ эмп}$

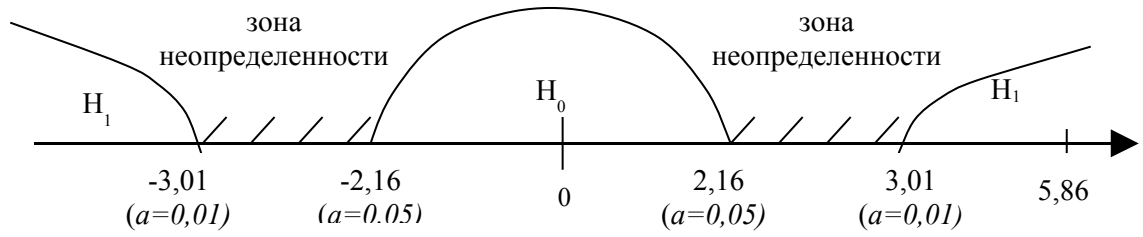
$$R_{rb\ эмп} = \frac{(5{,}44+11{,}83)*2}{15} = 0{,}852.$$

Проверим значимость полученного коэффициента корреляции при числе степеней свободы *к=n-2=15-3=13*.

$$t_{эмп} = |0{,}852| * \sqrt{\frac{15-2}{1-0{,}852^2}} = 5{,}86.$$

По таблице значений t-критерия Стьюдента находим критические значения:

$t_{кр} = \begin{cases} 2{,}16 & для \quad \alpha \le 0{,}05 \\ 3{,}01 & для \quad \alpha \le 0{,}01 \end{cases}$. Строим "ось значимости":



Результат попал в зону значимости. Поэтому принимается гипотеза H_1 согласно которой полученный рангово-бисериальный коэффициент корреляции значимо отличается от нуля. Иными словами, на данной выборке подростков отражены значимые тендерные различия по степени выраженности вербальных способностей.

Задания для домашней работы:

1. В результате анкетного обследования для выявления важнейших видов оборудования, используемого судоводителями во время вахты, получены два ряда ранговых оценок: «по важности» оборудования и «по частоте» его использования.

Важность	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Частота	1	4	2	6	3	5	12	9	15	7	11	8	10	14	17	13	16	18	19

Взаимосвязаны ли эти ряды?

2. По данным анкетного обследования получены два ряда групп работников, упорядоченных в соответствии с интересом к выполняемой работе (У1) и по соответствию образования и работе (У2). Определить есть ли корреляция между этими

переменными.

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$У_1$	3	3	3	3	3	6,5	6,5	8	9	10	11	12	13	14
$У_2$	1	5,5	9	10	11,5	3	8	4	2	7	5,5	11,5	13	14

3. Двум студентом с разных факультетов предполагалось проранжировать 10 качеств преподавателя математики. Данные представлены в таблице.

№	Качества	Ранг присвоенный первым студентом	Ранг присвоенный вторым студентом
1	Дружелюбие	4	8
2	Занимательность	1	3
3	Строгость	5	4
4	Увлеченность	6	5
5	Справедливость	7	2
6	Компетентность	2	1
7	Корректность	3	6
8	Стиль	8	7
9	Требовательность	9	9
10	Успешность	10	10

Определить коррелируют ли оценки проставленные преподавателю студентами с разных факультетов.

4. В группе спортсменов (футболистов и гимнастов) сравнивалось время реакции выбора в мс. Связаны ли различия во времени со специализацией спортсменов?

Специализация	ф	ф	г	ф	г	г	г	ф	ф	г
Время реакции	203	184	213	169	246	184	282	216	209	190

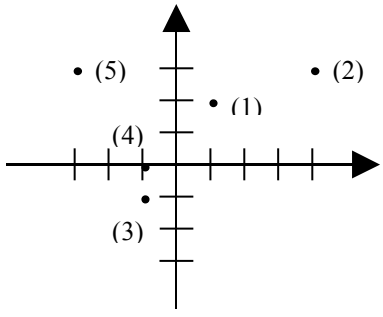
5. Связаны ли успешность обучения на первом курсе с выбором профессии по специальности в дальнейшем?

	Профессия по специальности	Профессия не по специальности
Учился плохо	2 человека	4 человека
Учился хорошо	5 человек	1 человек

12. Меры сходства и связи, вычисление расстояний.

Типовые задания.

1. Провести классификации пяти точек: (1,2), (4, 3), (-1, -1), (-1, 0), (-3, 3):



За метрику расстояний примем квадратичное евклидово расстояние. Матрица расстояний:

	1	2	3	4	5
1	0	10	13	8	17
2	10	0	41	34	49
3	13	41	0	1	20
4	8	34	1	0	13
5	17	49	20	13	0

Начальное разбиение: 1, 2, 3, 4, 5. Минимальное расстояние $\rho_{3,4} = 1$, переходим к разбиению: $1, 2, 3 \oplus 4, 5$.

Ближний сосед

	1	2	$3 \oplus 4$	5
1	0	10	8	17
2	10	0	34	49
$3 \oplus 4$	8	34	0	13
5	17	49	13	0

→

	$1 \oplus (3 \oplus 4)$	2	5
$1 \oplus (3 \oplus 4)$	0	10	13
2	10	0	49
5	13	49	0

→

	$(1 \oplus (3 \oplus 4)) \oplus 2$	5
$(1 \oplus (3 \oplus 4)) \oplus 2$	0	13
5	13	0

Последовательность разбиений: $1, 2, 3, 4, 5 \to 1, 2, 3 \oplus 4, 5 \to 1 \oplus (3 \oplus 4), 2, 5 \to (1 \oplus (3 \oplus 4)) \oplus 2, 5 \to 1 \oplus 2 \oplus 3 \oplus 4 \oplus 5$.

Поскольку этот метод объединяет кластеры, в которых расстояние между ближайшими элементами минимально по сравнению с другими кластерами, то два объекта попадают в один кластер, если существует соединяющая их цепочка ближайших друг к другу объектов («цепочечный эффект»). Поэтому метод «ближний сосед» называют методом «одиночной связи». Для устранения «цепочечного эффекта» можно, например, ввести ограничение на максимальное расстояние между объектами кластера: в первый кластер включить два наиболее близких объекта, затем в этот кластер включить объект, который имеет минимальное расстояние с одним из объектов кластера, а его расстояние до другого объекта кластера не больше числа l_0 и т. д.; формирование первого кластера продолжают до тех пор, пока нельзя будет найти объект, расстояние которого до любого объекта кластера, не превзойдет l_0 ; формирование 2-го и последующих кластеров осуществляется из оставшихся объектов аналогичным образом.

Дальний сосед

	1	2	$3 \oplus 4$	5
1	0	10	13	17
2	10	0	41	49
$3 \oplus 4$	13	41	0	20
5	17	49	20	0

→

	$1 \oplus 2$	$3 \oplus 4$	5
$1 \oplus 2$	0	41	49
$3 \oplus 4$	41	0	20
5	49	20	0

→

	$1 \oplus 2$	$(3 \oplus 4) \oplus 5$
$1 \oplus 2$	0	49
$(3 \oplus 4) \oplus 5$	49	0

Последовательность разбиений: $1, 2, 3, 4, 5 \to 1, 2, 3 \oplus 4, 5 \to 1 \oplus 2, 3 \oplus 4, 5 \to 1 \oplus 2, (3 \oplus 4) \oplus 5 \to 1 \oplus 2 \oplus 3 \oplus 4 \oplus 5$.

В этом методе объединяются кластеры, в которых минимально расстояние между самыми далекими друг от друга объектами. Это означает, что все остальные объекты в полученном после объединения кластере связаны друг с другом еще теснее, чем «соседи». Поэтому метод «дальнего соседа» называют методом полной связи.

Средняя связь

	1	2	3⊕4	5
1	0	10	10.5	17
2	10	0	37.5	49
3⊕4	10.5	37.5	0	16.5
5	17	49	16.5	0

→

	1⊕2	3⊕4	5
1⊕2	0	24	33
3⊕4	24	0	16.5
5	33	16.5	0

→

	1⊕2	(3⊕4)⊕5
1⊕2	0	27
(3⊕4)⊕5	27	0

В последней таблице $\rho_{1\oplus 2,(3\oplus 4)\oplus 5} = \frac{2}{3}\rho_{1\oplus 2,3\oplus 4} + \frac{1}{3}\rho_{1\oplus 2,5} = \frac{2}{3}\cdot 24 + \frac{1}{3}\cdot 33 = 27$

Последовательность разбиений: $1, 2, 3, 4, 5 \to 1, 2, 3\oplus 4, 5 \to 1\oplus 2, 3\oplus 4, 5 \to 1\oplus 2, (3\oplus 4)\oplus 5 \to 1\oplus 2\oplus 3\oplus 4\oplus 5.$

Центроидный метод

	1	2	3⊕4	5
1	0	<u>10</u>	10.25	17
2	<u>10</u>	0	37.25	49
3⊕4	10.25	37.25	0	16.25
5	17	49	16.25	0

→

	1⊕2	3⊕4	5
1⊕2	0	21.25	30.5
3⊕4	21.25	0	<u>16.25</u>
5	30.5	<u>16.25</u>	0

→

	1⊕2	(3⊕4)⊕5
1⊕2	0	$\frac{373}{18}$
(3⊕4)⊕5	$\frac{373}{18}$	0
	1⊕2	(3⊕4)⊕5

Задания для домашней работы:

1. Используя не менее двух методов кластер – процедур провести классификацию точек (3; 4) , (-3; 8), (-2;-6), (5; 7), (6; 0), (8; 4), (-3; 9), (2;-6), (5; -7), (5; 0). В качестве расстояний выбрать евклидово расстояние и расстояние городских кварталов.

2. Для представленных стран и характеризующих их признаков провести стандартизацию данных и рассчитать всевозможные расстояния между объектами..

страна	доля экспорта (%)	доля импорта (%)	население (млн.чел.)	% безработного населения
Россия	50	80	146	20
Украина	20	50	120	35
Белоруссия	30	25	100	15
Польша	10	60	95	10
Бельгия	60	20	70	15

13. Построение иерархических дендрограмм.

Типовые задания.

1. Построить дендрограмму для классификации стран по уровню экономического развития, если известны доли импорта и экспорта данных стран, общее количество населения, а также процент безработного населения каждой страны.

страна	доля экспорта (%)	доля импорта (%)	население (млн.чел.)	% безработного населения
Россия	50	80	146	20
Украина	20	50	120	35
Белоруссия	30	25	100	15
Польша	10	60	95	10
Бельгия	60	20	70	15

Решение: данные необходимо стандартизировать, т.к. они имеют разные единицы измерения:

страна	доля экспорта (%)	доля импорта (%)	население (млн.чел.)	% безработного населения
Россия	0,77	1,33	1,40	0,10
Украина	-0,68	0,12	0,48	1,66
Белоруссия	-0,19	-0,88	-0,22	-0,42
Польша	-1,16	0,52	-0,39	-0,94
Бельгия	1,25	-1,08	-1,27	-0,42

1 шаг. Все объекты располагаем на графике по оси OY, на оси OX расположим расстояния. За метрику расстояний примем евклидово расстояние. Матрица расстояний имеет вид:

	Россия	Украина	Белоруссия	Польша	Бельгия
Россия	0	2,61	2,95	2,94	3,66
Украина	2,61	0	2,46	2,81	3,55
Белоруссия	2,95	2,46	0	1,79	1,80
Польша	2,94	2,81	1,79	0	3,07
Бельгия	3,66	3,55	1,80	3,07	0

Из данных расстояний выбираем наименьшее 1,79 и объединяем объекты в один класс. Получается 4 объекта: Россия, Украина, (Польша+Белоруссия), Бельгия.

2 шаг. В качестве меры связи выберем центроидный метод. Найдем значения для кластера (Польша+Белоруссия) как среднее арифметическое:

страна	доля экспорта (%)	доля импорта (%)	население (млн.чел.)	% безработного населения
Россия	0,77	1,33	1,40	0,10
Украина	-0,68	0,12	0,48	1,66
Белоруссия+Польша	-0,68	-0,18	-0,31	-0,68
Бельгия	1,25	-1,08	-1,27	-0,42

Вновь рассчитаем расстояния:

	Россия	Украина	Белоруссия+Польша	Бельгия

Россия	0,00	2,61	2,81	3,66
Украина	2,61	0,00	2,49	3,54
Белоруссия+Польша	2,81	2,49	0,00	2,35
Бельгия	3,66	3,55	2,35	0,00

Самое минимальное расстояние 2,35 между кластером (Белоруссия+Польша) и Бельгией. Объединяем их в одну группу и отображаем на графике. Процесс продолжаем до тех пор пока все страны не объединятся в один кластер.

3 шаг.

страна	доля экспорта	доля импорта	население	Безработ. населения
Россия	0,77	1,33	1,40	0,10
Украина	-0,68	0,12	0,48	1,66
Белоруссия+Польша+Бельгия	0,29	-0,63	-0,79	-0,55

Матрица расстояний имеет вид:

	Россия	Украина	Белоруссия+Польша+Бельгия
Россия	0,00	2,61	3,05
Украина	2,61	0,00	2,83
Белоруссия+Польша+Бельгия	3,05	2,83	0,00

Минимальное расстояние 2,61, на котором объединяются Россия и Украина.

4 шаг:

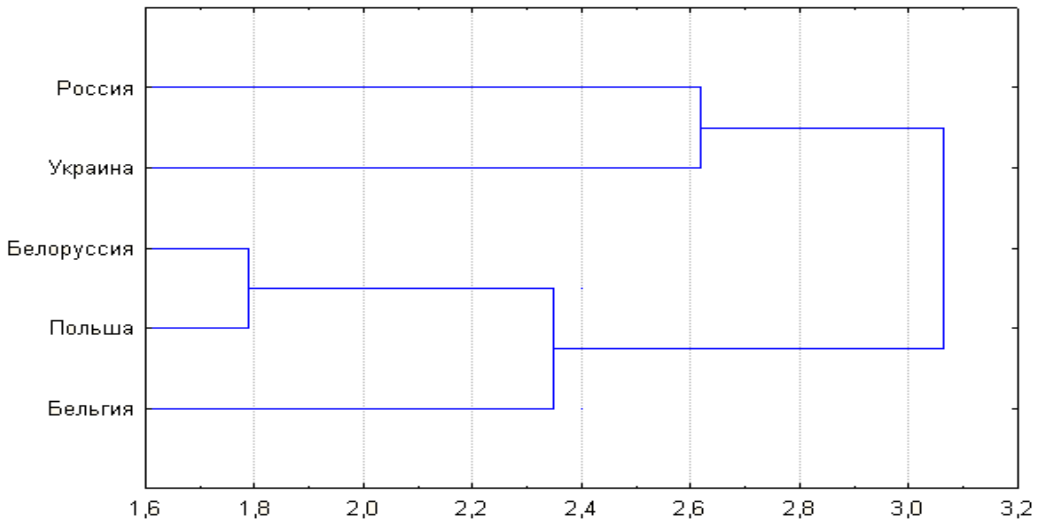
страна	доля экспорта	доля импорта	население	Безработ. населения
Россия+Украина	0,05	0,73	0,94	0,88
Белоруссия+Польша+Бельгия	0,29	-0,63	-0,79	-0,55

Матрица расстояний:

	Россия+Украина	Белоруссия+Польша+Бельгия
Россия+Украина	0,00	3,05
Белоруссия+Польша+Бельгия	3,05	0,00

Итак все объекты объединяются на расстоянии 3,05.

Дендрограмма имеет вид:



Задания для домашней работы:

1. Исследуются автомобили различных марок с целью разбить их по группам в зависимости от представленных в таблице 9 характеристик. Построить дендрограмму для данных автомобилей.

Марка авто	Время разгона (с)	Максимальная скорость (км/ч)	Средняя цена (тыс.руб)	Престиж-ность (0-нет,1-да)	Число моделей (шт.)	Гарантийный срок (в годах)
Мазда	7	170	170	0	20	4
Тойота	8	180	120	1	28	5
Сузуки	13	160	130	0	15	3
Мицубиси	9	190	150	0	35	7
Форд	4	245	600	1	18	10
Рено	7	220	500	1	22	15
Джип	4	200	350	1	17	12
ВАЗ	20	140	60	0	31	2

2. Объединить 8 Амурских фирм, занимающихся производством и установкой окон в несколько схожих совокупностей путем построения дендрограммы.

фирма	возраст фирмы	кол-во сотрудников	количество установленных окон за месяц	средняя цена окна (тыс.руб)	Количество филиалов по области	Скидки(%)	популярность (0-нет, 1да)
Home master	11	200	30	17	20	25	1
Уют	3	14	12	13,2	2	33	1
РосОкна	3	50	8	14,8	5	30	1
Ремикс	5	80	14	15	10	10	1
Ванда	8	35	17	14,2	12	15	1
Ремстрой	2	10	5	11,5	1	5	0
СМУ-17	7	20	3	13,5	1	0	0
Оникс	3	20	1	12,9	1	0	0

14. **Последовательный кластерный анализ, метод к – средних**

Типовые задания.

Разбить представленные точки на два кластера A(1,2), B(4, 3), C(-1, -1), D(-1, 0), E(-3, 3).

Решение: так как количество кластеров задано в условии воспользуемся методом к – средних, положив к=2.

На первом этапе рассчитаем евклидово расстояние между точками, расположим его в таблице

	A	B	C	D	E
A	0,00	3,16	3,61	2,83	4,12
B	3,16	0,00	6,40	5,83	7,00
C	3,61	6,40	0,00	1,00	4,47

D	2,83	5,83	1,00	0,00	3,61
E	4,12	7,00	4,47	3,61	0,00

Выберем две наиболее удаленные точки в качестве кластерных центров, ими являются точки E и B. Пусть точка E отвечает за первый кластер, B – за второй. Определяем расстояния от точек E и B до каждой из оставшихся точек.

	A	C	D
E	4,12	4,47	3,61
B	3,16	6,4	5,83

Точки C, D ближе к точке E чем к B, поэтому определяем их в первый кластер, а точку A во второй и рассчитываем для них кластерные центры (M), как среднее арифметическое между всеми точками, входящими в кластер:

$$M_1 = \left(\frac{-1-1-3}{3}; \frac{-1+0+3}{3}\right) = (-1{,}67; 0{,}67) \qquad M_2 = \left(\frac{1+4}{2}; \frac{2+3}{2}\right) = (2{,}5; 2{,}5)$$

Вычисляем расстояния между кластерными центрами, а также от каждой точки до ее кластерного центра

	A	B	C	D	E	M_2
M_1			1,80	0,95	2,68	**4,55**
M_2	1,58	1,58				0

Анализ показывает, что расстояние между кластерными центрами гораздо больше чем расстояние от каждой из точек до своего кластерного центра, что говорит о верной классификации объектов.

Итак, в один кластер необходимо поместить точки A и B, а в другой – C, D и E.

Задания для домашней работы:

18 претендентов прошли 10 различных тестов в кадровом отделе предприятия.

Номер теста	обозначение	Предмет теста
1	t1	Память на числа
2	t2	Математические задачи
3	t3	Находчивость при прямом диалоге
4	t4	Тест на составление алгоритмов
5	t5	Уверенность во время выступления
6	t6	Командный дух
7	t7	Находчивость
8	t8	Сотрудничество
9	t9	Признание в коллективе
10	t10	Сила убеждения

Максимальная оценка, которую можно было получить на каждом из тестов, составляет 10 балов. Результаты теста для 18 претендентов находятся в таблице в переменных t1-t10. Каждое наблюдение является характеристикой тестируемых кандидатов.

инициалы	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10
А.П.	10	9	3	10	4	5	1	7	6	5
М.К.	6	5	8	6	9	6	2	6	5	9
С.Т.	7	4	5	6	6	7	5	5	4	6
Т.Т.	8	8	5	6	6	4	6	7	7	5
А.А.	5	4	8	3	9	2	9	9	8	9
М.Р.	6	4	4	5	6	8	7	4	2	6
А.С.	9	10	7	8	6	9	3	6	5	7
Д.Н.	4	3	7	3	6	5	6	10	9	6
Ю.Т.	8	5	4	9	5	8	4	6	7	7
К.А.	8	6	4	6	5	7	8	6	8	5
Ч.Н.	7	7	6	7	7	9	9	3	4	7
М.Т.	5	5	9	5	10	5	10	6	5	9
Л.Г.	4	5	10	4	10	4	5	8	9	10
С.П.	10	10	6	9	5	8	4	8	10	5
Р.Д.	6	5	4	6	4	4	6	5	6	4
Ф.А.	6	7	3	7	2	3	8	4	3	3
К.С.	7	8	5	7	3	6	7	4	4	4
Г.Т.	8	9	6	8	5	10	7	10	8	5

С использованием результатов теста соответствия провести кластерный анализ и обнаружить 3 группы кандидатов, близких по своим качествам.

15. Определение оптимального количества факторов, критерий Кайзера, критерий факторной осыпи.

Типовые задания.

Задача по теме «Факторный анализ» может быть сформулирована на первом занятии по теме и по мере знакомства с понятийным аппаратом факторного анализа постепенно решена в течение 3-4 последующих занятий. Пример задачи приведен ниже.

При переходе детей из начальной школы в классы среднего звена решено организовать несколько профильных классов (математический, гуманитарный, спортивный и др.). В связи с этим социологу поручено разработать систему тестов для младших школьников, с помощью которых можно определить способности ребенка. Фрагмент результатов тестов приведен в таблице:

	рост	вес	Память на числа	Логическое мышление	Техника чтения	фантазия	Прыжки в длину
Ира	150	35	7	1	180	0	70
Катя	156	34	6	0	156	1	120
Маша	168	32	5	1	189	0	140
Даша	160	39	2	0	190	1	80
Юля	162	45	4	0	200	1	90
Таня	159	30	5	0	230	0	100
Кира	145	28	9	1	250	1	120
Оля	150	32	7	1	200	0	110
Поля	160	25	7	1	180	0	100

Задание 1. Сжать матрицу исходных данных путем выделения оптимального количества факторов.

Решение: на первом этапе необходимо стандартизировать исходные данные,. Матрица стандартизированных данных имеет вид:

	рост	вес	Память на числа	Логическое мышление	Техника чтения	фантазия	Прыжки в длину
Ира	-0,93	0,28	0,60	0,84	-0,61	-0,84	-1,53
Катя	-0,09	0,11	0,11	-1,05	-1,47	1,05	0,76
Маша	1,58	-0,22	-0,38	0,84	-0,29	-0,84	1,68
Даша	0,47	0,95	-1,84	-1,05	-0,26	1,05	-1,07
Юля	0,74	1,96	-0,87	-1,05	0,10	1,05	-0,61
Таня	0,33	-0,56	-0,38	-1,05	1,17	-0,84	-0,15
Кира	-1,63	-0,90	1,57	0,84	1,88	1,05	0,76
Оля	-0,93	-0,22	0,60	0,84	0,10	-0,84	0,31
Поля	0,47	-1,40	0,60	0,84	-0,61	-0,84	-0,15

Далее по матрице стандартизированных данных надо определить коэффициенты корреляции между всеми переменными:

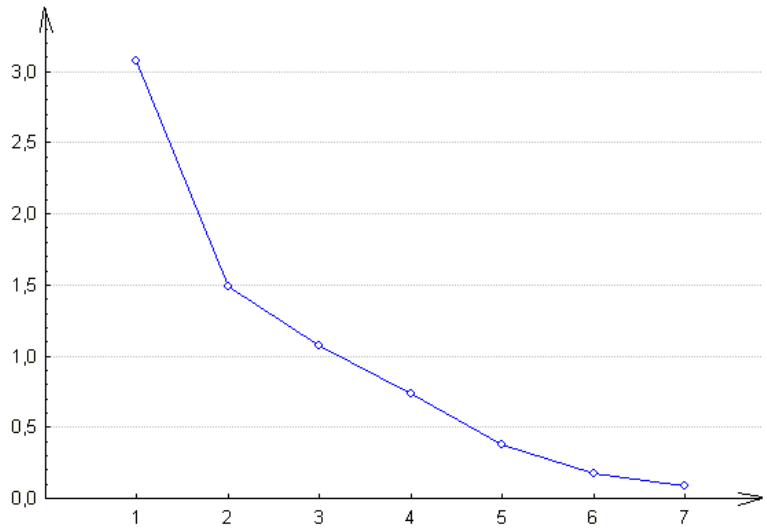
	рост	вес	Память на числа	Логическое мышление	Техника чтения	фантазия	Прыжки в длину
Рост	1,00	0,26	-0,70	-0,34	-0,36	-0,12	0,18
Вес	0,26	1,00	-0,67	-0,58	-0,23	0,50	-0,42
Память	-0,70	-0,67	1,00	0,71	0,27	-0,24	0,30
Логическое мышление	-0,34	-0,58	0,71	1,00	0,11	-0,55	0,25
Техника чтения	-0,36	-0,23	0,27	0,11	1,00	0,06	0,12
Фантазия	-0,12	0,50	-0,24	-0,55	0,06	1,00	-0,04
Прыжки в длину	0,18	-0,42	0,30	0,25	0,12	-0,04	1,00

Для определения оптимального количества латентных факторов необходимо найти собственные числа корреляционной матрицы путем решения уравнения: $|R - \lambda E| = 0$. Определить сколько % дисперсии объясняет каждое собственное число: $\frac{\lambda_i}{\sum \lambda} 100\%$ и вычислить накопленный процент.

Собственные числа	% объясненной дисперсии	Накопленный %
3,071116	43,87308	43,8731
1,484121	21,20173	65,0748
1,072478	15,32112	80,3959
0,738453	10,54932	90,9453
0,373828	5,34040	96,2857
0,172748	2,46783	98,7535
0,087256	1,24652	100,0000

Согласно критерию Кайзера значимыми являются три фактора (с собственными числами больше 1), они суммарно объясняют чуть более 80% всей дисперсии.

Для того чтобы воспользоваться критерием Кеттеля, необходимо изобразить собственные числа в системе координат. По оси OX располагаем порядковый номер собственного числа, а по оси OY – значения собственных чисел.



Согласно графику факторной осыпью можно считать собственные числа, начиная с пятого, так как именно с этой точки график начинает плавно приближаться к оси OX. Причем четыре оставшихся собственных числа объясняют всю дисперсию на 90,9 %.

Итак, оптимальное количество факторов данной задачи 3 или 4, более точный результат можно получить путем возможности интерпретации факторов.

Домашнее задание.

Определите удовлетворенность, какой стороной жизни преобладает у разных респондентов. Пусть в эксперименте участвовало 10 человек, которые отвечали на вопросы, представленные в таблице:

возраст	образование	Стаж непрерывной работы	Работа по специальности	Зарплата, тыс.руб	Кол-во человек в семье	жилплощадь	Наличие хобби	Посещение учреждение вне работы

Образование: 2 – высшее, 1 – специальное, 0 – среднее.

Работа по специальности: 1 – да, 0 – нет.

Хобби: 1 – да, 0 – нет.

Посещение учреждение вне работы: 3 – часто, 2 – иногда, 1 – нет.

Проведите стандартизацию данных и определите оптимальное количество факторов.

16. Выделение латентных переменных, нахождение и интерпретация матри-

цы факторных нагрузок.

Типовые задания.

В условиях задачи прошлого практического занятия найти матрицы факторных нагрузок для трех и четырех латентных факторов.

Решение: коэффициентами матрицы факторных нагрузок являются нормированные собственные векторы корреляционной матрицы, соответствующие выделенным раннее собственным числам.

Собственные векторы определяются путем решения систем линейных однородных уравнений вида $(R - \lambda E)X = 0$, а их нормирование производится по формуле $x_i^н = \frac{x_i}{\sqrt{x_1^2 + x_2^2 + ... + x_n^2}}$. Кроме того, после нормирования каждую координату вектора необходимо умножить на весовой коэффициент, в качестве которого выступает $\sqrt{\lambda}$.

Итак, для найденных на предыдущем занятии собственных чисел λ_1=3,071116; λ_2=1,484121; λ_3=1,072478; λ_4=0,738453 нормированные взвешенные собственные векторы равны: x_1=(-0,55; -0,44; 0,91; 0,84; 0,36; -0,5; 0,39);

x_2=(-0,24; 0,23; 0,23; -0,20; 0,52; 0,63; -0,36);

x_3=(0,74; 0,12; 0,03; 0,22; -0,38; -0,41; -0,8);

x_4=(-0,19; 0,07; 0,25; 0,08; -0,67; 0,37; 0,22).

Располагая координаты полученных собственных векторов по столбцам, а данные переменные по строке получим матрицу факторных нагрузок.

	Фактор 1	Фактор 2	Фактор 3	Фактор 4
Рост	-0,55	-0,24	**0,74**	-0,19
Вес	-0,44	0,23	0,12	0,07
Память	**0,91**	0,23	0,03	0,25
Логическое мышление	**0,84**	-0,20	0,22	0,08
Техника чтения	0,36	**0,52**	-0,38	-0,67
Фантазия	-0,50	**0,63**	-0,41	0,37
Прыжки в длину	0,39	-0,36	**-0,80**	0,22

В матрице факторных нагрузок содержатся коэффициенты корреляции между выделенными факторами и исходными переменными, анализ которых показывает, что первый фактор сильнее всего зависит от тестов на память и логическое мышление, т.е. может отвечать за математические способности школьников; второй фактор – гуманитарной направленности; третий – спортивной, а четвертый фактор не поддается интерпретации. В связи с этим в матрице факторных нагрузок оставляем три

фактора, а четвертый отбрасываем.

Задания для домашней работы:

Для задачи, сформулированной на прошлом занятии определить матрицу факторных нагрузок.

17. Нахождение и интерпретация матрицы факторных весов.

Типовые задания.

Определить матрицу факторных весов для задачи, сформулированной на 15 практическом занятии.

Решение:

Для нахождения матрицы факторных весов достаточно найти произведение матриц:

Элементы матрицы факторных весов являются значениями выделенных факторов для каждого испытуемого. Матрица факторных весов для данной задачи имеет вид:

факторы	математический	гуманитарный	спортивный
Марина	-0,74	-0,94	-1,45
Катя	1,31	-0,70	1,02
Маша	-1,03	1,44	1,60
Даша	0,26	0,90	-1,09
Юля	0,27	0,70	-0,38
Таня	1,79	0,28	-0,15
Кира	-0,80	-1,19	0,55
Оля	-0,29	-1,15	0,52
Поля	-0,78	0,66	-0,61

Анализируя матрицу можно сказать, что все выделенные способности у Марины ниже среднего (факторные весы <0), но преобладают математические.

В математический класс целесообразно направить Катю и Таню; в гуманитарный – Дашу, Юлю, Полю; а в спортивный – Машу, Киру, Олю.

Задания для домашней работы:

Для полученной ранее задачи найти матрицу факторных весов.

18. Вращение факторов.

Типовые задания.

К решению проблемы вращения прибегают не часто, но если в этом возникает необходимость, то достаточно ограничится варимаксным вращением.

Дана матрица А первых двух факторов для семи признаков. Преобразовать дан-

ную матрицу путем поворота на угол 30° по часовой стрелке.

$$A = \begin{pmatrix} 0,9 & -0,3 \\ 0,8 & -0,3 \\ 0,6 & 0,3 \\ 0,5 & 0,2 \\ 0,8 & 0,6 \\ -0,8 & 0,5 \\ 0,2 & 0,8 \end{pmatrix}$$

Решение: обозначим матрицу, полученную при вращении A_1. При варимаксном вращении $A_1=AT$, где T – ортогональная матрица преобразования.

В случае двух факторов $T = \begin{pmatrix} cos\alpha & \sin\alpha \\ -\sin\alpha & cos\alpha \end{pmatrix}$, т.к. sin 30°=0,5; cos 30°=0,866, то

$$T = \begin{pmatrix} 0,866 & 0,5 \\ -0,5 & 0,866 \end{pmatrix}.$$

$$A_1 = \begin{pmatrix} 0,9 & -0,3 \\ 0,8 & -0,3 \\ 0,6 & 0,3 \\ 0,5 & 0,2 \\ 0,8 & 0,6 \\ -0,8 & 0,5 \\ 0,2 & 0,8 \end{pmatrix} \cdot \begin{pmatrix} 0,866 & 0,5 \\ -0,5 & 0,866 \end{pmatrix} = \begin{pmatrix} 0,929 & 0,19 \\ 0,843 & 0,14 \\ 0,367 & 0,56 \\ 0,333 & 0,423 \\ 0,393 & 0,92 \\ -0,943 & 0,033 \\ 0,227 & 0,793 \end{pmatrix}$$

Задания для домашней работы:

Для первых двух факторов найденной Вами матрицы факторных нагрузок произвести варимаксное вращение на 30°, 45° и 60°.

3.3. Типовой вариант контрольной работы.

1. У учеников 11 класса решили проверить знания по русскому языку. Им предложили тест по предмету в середине учебного года и в конце всего периода обучения. Получены следующие результаты: в середине года: 26,25,21,28,30,27,28,26,30,21; в конце года: 20,23,20,28,28,26,25,26,30,20.

Можно ли утверждать, что в середине года ученики обладали лучшими знаниями.

2. Оцените разброс в ценах на однотипную продукцию между фирмами, занимающимися продажей бытовой техники. 1 фирма: 2900, 1500, 3000, 7500, 6900, 8600; 2 фирма: 3100, 3900, 4600, 5900, 7000, 7050, 7400.

3. Определить оказывает ли влияние на объем продаж бытовой техники реклама.

Месяц	Объем продаж, тыс.руб.	
	До рекламы	После рекламы
Январь	16	28
Февраль	5	6
Март	15	37
Апрель	3	6
Май	5	5
Июнь	10	2
Июль	18	26
Август	16	23
Сентябрь	14	16
Октябрь	9	10
Ноябрь	8	6
Декабрь	15	23

3.4. Задание для индивидуальной самостоятельной работы студентов.

1) *Расчетно-графическая работа*.

Работа рассчитана на 6-8 недель, ее целью является проверка знаний студентов по темам: «Непараметрические критерии», «Корреляционный анализ», «Критерии согласия». Выбор варианта осуществляется согласно порядковому номеру студента в списке группы.

1. 12 участников комплексной программы тренинга партнерского общения, продолжавшегося 7 дней, дважды оценивали у себя уровень владения тремя важнейшими коммуникативными навыками. Первое измерение производилось в первый день тренинга, второе – в последний. Участники должны были также наметить для себя реально достижимый, с их точки зрения, индивидуальный идеал в развитии каждого из навыков. Данные представлены в таблице. n – номер варианта.

Ощ[illegible] навыков разные сдвиги, или эти сдвиги для разных навыков примерно одинаковы. Уменьшается ли [illegible] после

код имени участника		1 измерение						2 измерение					
		Активное слушание		Снижение эмоционального напряжения		Аргументация		Активное слушание		Снижение эмоционального напряжения		Аргументация	
		Реал.	Идеал.	Реал.	Идеал.	Реал.	Идеал.	Реал.	Идеал.	Реал.	Идеал.	Реал.	Идеал.
1	И.	6	9n	5n	8	5	8	7	10n	6n	10	7	9
2	Я.	3	5n	1n	3	4	5	5	7n	4n	6	5	7
3	Ин.	4	6n	4n	6	5	8	8	10n	7n	8	6	8
4	Р.	4	6n	4n	5	5n	7	6	7n	5n	7	5n	7
5	К.	6	9n	4n	9	4	8	4	10n	5n	10	5	10
6	Н..	6	8n	5n	8	3n	6	8	9n	7n	9	6n	8
7	Е.	3	8n	5n	10	2	6	7	8n	8n	10	5	7
8	Ле.	6	9n	5n	8	3	7n	5	8n	7n	10	5	9n
9	Ли.	6	8n	5n	9	5	9	7	8n	6n	9	5	9
10	Т.	5	8n	6n	9	5n	8	7	10n	7n	10	6n	10
11	Ет.	6	8n	6n	10	3	9n	5	10n	4n	9	3	9n
12	Б.	6	8n	3n	10	4	7	7	9n	6n	8	5	8

2. Определить, влияет ли пол и тип темперамент человека на предпочтение им жанра кинофильма. n – номер варианта.

пол	Тип темпера-мента	жанр кинофильма			
		комедия	боевик	Триллер	мелодрама
мужской	Холерик	15+n	5n	8	2
	Сангвиник	20	12+2n	12+n	3+n
	Меланхолик	n	3n	n	20
	Флегматик	n-1	n	10	15
женский	Холерик	n	25	2	2n
	Сангвиник	n	n	12	n
	Меланхолик	5+n	15	5	100-n
	Флегматик	15+n	10	n-1	50

3. Определить будет ли удовлетворенность работой на данном предприятии распределена равномерно по следующим альтернативам: 1 – Работой вполне доволен; 2 – Скорее доволен, чем не доволен; 3–Трудно сказать, не знаю, безразлично; 4–Скорее недоволен, чем доволен; 5 – Совершенно недоволен работой. Для решения этой задачи производится опрос случайной выборки респондентов (испытуемых) об удовлетворенности работой.

альтернативы	1	2	3	4	5
частоты	100-n	52+n/3	30+n/2	75-n/5	n+n/3

Если удовлетворенность работой не подчиняется равномерному закону, то проверить, не является ли она нормальной? n – номер варианта, переписывая исходные данные результаты в таблице округлить до ближайшего целого.

4. При интервью чаще всего люди отказываются отвечать на вопросы личного характера (о состоянии здоровья, о размере заработной платы). Исследователь решил выяснить, зависит ли согласие отвечать на вопросы о личной жизни от стиля интервьюирования. Для этого проведен эксперимент: одним и тем же испытуемым предлагалось ответить на ряд вопросов личного характера. В первом случае вопросы задавались в восторженной манере, в дружественной беседе; во втором – стиль общения был формальным, а в третьем – происходила резкая незаинтересованная беседа. 10 вопросов личного характера были «перемешаны» с общими вопросами анкеты. Результаты эксперимента – количество ответов на вопросы личного характера представлены в таблице. Определить зависит ли согласие отвечать на вопросы от стиля интервьюирования, если разные способы апробировались на разных людях.

Способ 1	10	8	6	9	10	9	9	6	7	7	8	8	8	9	10
Способ 2	8	7	7	7	5	5	9	6	8	7	8	8	9	9	10
Способ 3	5	5	2	3	3	6	4	4	4	5	6	2	3	1	0

5. Дана стоимость квартир в некотором городе. Наряду с ценой квартиры и ее некоторыми второстепенными характеристиками представлена информация (переменная ранг) о качестве квартиры по сравнению с другими собранная по оценкам экспертов.

Рассчитать коэффициенты корреляции Пирсона и с их помощью оценить наличие связи между ценой на квартиру и ее второстепенными характеристиками и определить есть ли связь между оценкой экспертов и характеристиками квартир.

В связи с тем, что большинство представленных данных измерены не в интервальной шкале, определить непараметрические коэффициенты корреляции, рассчитать их значимость с помощью t-статистики и сделать вывод о наличии связи между переменными.

Цена	Ранг	Тип дома (о-панельный, 1-кирпичный)	Наличие балкона (0-нет, 1- есть)	Наличие горячей воды (0-нет, 1- есть)	Наличие газовой плиты (0-нет, 1- есть)	Планировка квартиры (0- старая, 1-новая)
700	5	0	1	1	0	0
750	13	0	1	1	1	1
750	4	0	1	1	1	1
750	6	0	1	1	0	0
775	14	0	1	1	1	1
780	3	0	1	1	1	1
800	11	0	1	1	1	1
820	7	0	1	1	1	1
910	16	1	1	1	1	1
930	2	0	1	1	0	1
950	15	0	1	1	1	1
960	8	1	1	1	1	1
999	10	1	1	1	1	1
1380	12	1	1	1	1	1
1560	9	1	1	1	1	1
2460	1	1	1	1	1	1
670	17	0	1	1	0	0
700	19	1	1	1	0	1
700	20	0	0	1	1	1
700	18	1	0	1	1	1

Защита расчетно-графической работы проводится по усмотрению преподавателя в устной или письменной форме.

2) *Творческая работа* по теме «Многомерные статистические методы иссле-

дования»: составить таблицу статистических данных, которая содержит 10-15 различных переменных и не менее 50 наблюдений (воспользоваться данными курсовых проектов). «Сжать» матрицу наблюдений, используя кластерный анализ. Выделить несколько латентных факторов, объединяющих исходные переменные, используя средства факторного анализа. Защита работы осуществляется в последнюю неделю семестра в устной форме.

3.5. Примерный вариант экзаменационных билетов

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 1

1. Матрица факторных весов: понятие, алгоритм нахождения, содержательный смысл элементов матрицы.

2. Критерий тенденций Джонкира для проверки гипотез

3. Задача. Определить есть ли связь между возрастом человека и наличием у него высшего образования в выборке из 10 человек:

возраст	40	27	30	25	40	32	22	20	19	35
высшее образование (есть «+»; нет «-»)	-	+	+	+	-	-	+	-	-	-

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 2

1. Матрица сходств: понятие, содержательный смысл элементов матрицы.

2. Критерий Пейджа для проверки гипотез

3. Задача. Подготовка детей к школе может осуществляться различными способами: дома индивидуально и в детском саду в коллективе под руководством педагога-воспитателя. На тестировании при поступлении в первый класс различные дети показали следующие результаты (из 20 возможных баллов):

«Домашние» дети	16	15	14	18	20	9	2	12	10						
Дети, посещающие детский сад	20	12	15	4	8	16	19	17	20	2	5	3	15	10	10

Определить воздействует ли место подготовки ребенка на его успехи при тестировании.

3.6. Карта обеспеченности дисциплины кадрами профессорско-преподавательского состава

шифр специальности	Наименование дисциплин в соответствии с учебным планом	Обеспеченность преподавательским составом							
		Ф.И.О. должность по штатному расписанию	Какое образовательное учреждение профессионального образования окончил, специальность по диплому	Ученая степень и ученое звание (почетное звание)	Стаж научно педагогической работы			Основное место работы, должность	Условия привлечения к трудовой деятельности (штатный, совместитель (внутренний или внешний с указанием доли ставки), иное
					Всего	В т. ч. педагогический			
						Всего	В том числе по преподаваемой дисциплине		
1	2	3	4	5	6	7	8	9	10
040201	Мат.методы в социологии	Двоерядкина Н.Н., ст.преподаватель	БГПУ, учитель математики	к.п.н.	8	8	4	АмГУ	1 ставка