Министерство образования Республики Беларусь
Белорусский государственный университет
Факультет философии и социальных наук
Кафедра социологии

СОГЛАСОВАНО	СОГЛАСОВАНО
Заведующий кафедрой	Декан факультета
____________	____________
А.Н. Данилов	В.С. Сайганова
«23» июня 2023 г.	«27» июня 2023 г.

Статистический анализ социологической информации

Электронный учебно-методический комплекс
для специальности: 6-05-0314-01 «Социология»
профилизация специальности
«Методология и методы социологического исследования»

Регистрационный № 2.4.2-24/338

Автор:
Елсукова Н. А., кандидат социологических наук, доцент

Рассмотрено и утверждено на заседании Научно-методического совета БГУ 03.05.2023 г., протокол № 7.

Минск 2023

УДК 303.71(075.8)
Е 552

Утверждено на заседании Научно-методического совета БГУ
Протокол № 7 от 03.05.2023 г.

Решение о депонировании вынес:
Совет факультета философии и социальных наук
Протокол № 11 от 27.06.2023 г.

А в т о р:
Елсукова Наталья Альбертовна, кандидат социологических наук, доцент БГУ, кафедра социологии, факультет философии и социальных наук.

Рецензенты:
кафедра философии и методологии университетского образования ГУО «Республиканский институт высшей школы» (заведующий кафедрой Лемешова Т. В., кандидат политических наук);
Шкурова Е. В., старший научный сотрудник Института философии Национальной академии наук кандидат социологических наук, доцент.

Елсукова, Н. А. Статистический анализ социологической информации : электронный учебно-методический комплекс для специальности: 6-05-0314-01 «Социология», профилизация специальности «Методология и методы социологического исследования» / Н. А. Елсукова ; БГУ, Фак. философии и социальных наук, Каф. социологии. – Минск : БГУ, 2023. – 78 с. : ил., табл. – Библиогр.: с. 77–78.

Электронный учебно-методический комплекс (ЭУМК) «Статистический анализ социологической информации» подготовлен в соответствии с требованиями образовательного стандарта специальности «Социология», учебной программой по дисциплине «Статистический анализ социологической информации» в целях учебно-методического обеспечения студентов специальности 6-05-0314-01 «Социология». ЭУМК предназначен для студентов учреждений высшего образования, обучающихся по социологическим специальностям очной и заочной формы получения первого высшего образования.

Электронный учебно-методический комплекс содержит конспект лекций, контрольные вопросы по темам, темы семинарских занятий, примерный перечень первоисточников, перечень контрольных мероприятий, вопросы к экзамену, список рекомендуемой литературы.

ОГЛАВЛЕНИЕ

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА 5
1. ТЕОРЕТИЧЕСКИЙ РАЗДЕЛ 8
1.1. Статистика как наука. Данные социологического исследования, матрица «объект-признак» 8
1.2. Измерение в социологии, измерительные шкалы 11
1.3. Одномерное частотное распределение 14
1.4. Группировки количественных признаков в интервалы 16
1.5. Графическое представление социологических данных 21
1.6. Характеристики центра распределения признака 26
1.7. Показатели вариации признака 28
1.8. Анализ формы распределения признака 30
1.9. Стандартизация количественных переменных. Z-оценки 31
1.10. Теоретические распределения и их статистические таблицы 32
1.11. Статистический вывод. Оценка параметров генеральной совокупности 36
1.12. Простая случайная выборка из генеральной совокупности. Ошибка простой случайной репрезентативной выборки и ее объем 39
1.13. Понятие статистической гипотезы. Процедура проверки гипотезы 42
1.14. Виды статистических гипотез: гипотезы о долях, гипотезы о средних, гипотезы о дисперсиях 45
1.15. Статистический анализ взаимосвязи признаков 46
1.16. Таблица сопряженности. Проверка гипотезы о наличии связи в таблице сопряженности 48
1.17. Таблицы сопряженности размером 2x2 51
1.18. Теоретико-информационные меры связи 52
1.19. Ранжированные ряды. Меры парной связи ранжированных рядов 54
1.20. Корреляционный анализ 57
1.21. Регрессионный анализ. Парная линейная регрессия 60
1.22. Дисперсионный анализ. Однофакторная дисперсионная модель 61
2. ПРАКТИЧЕСКИЙ РАЗДЕЛ 63
Тематика семинарских занятий по учебной дисциплине «Статистический анализ социологической информации» 63
3. РАЗДЕЛ КОНТРОЛЯ ЗНАНИЙ 68
3.1. Методические рекомендации и примерный перечень заданий по УСР ... 68
3.2. Примерные варианты тестовых заданий 71

3.3. Примерный перечень вопросов к зачету и экзамену 73
4. ВСПОМОГАТЕЛЬНЫЙ РАЗДЕЛ.. 77
4.1. Рекомендуемая литература 77
4.2. Электронные ресурсы... 78

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Электронный учебно-методический комплекс «Статистический анализ социологической информации» подготовлен в соответствии с требованиями Положения об учебно-методическом комплексе на уровне высшего образования, утвержденного Постановлением Министерства образования Республики Беларусь от 26.07.2011 № 167, образовательным стандартом специальности «Социология», учебной программой по дисциплине «Статистический анализ социологической информации» в целях учебно-методического обеспечения студентов специальности 1-23 01 05 «Социология». ЭУМК предназначен для студентов высших учебных заведений социологических специальностей, очной и заочной формы получения первого высшего образования. Содержание ЭУМК направлено на углубленное понимание статистического подхода к анализу данных социологических исследований и развитие навыков практического применения основ математической статистики в эмпирической социологии.

Цель ЭУМК – получение студентами теоретических и практических знаний математической статистики, которые позволяют профессионально работать с числовыми базами данных, полученными в ходе проведения количественных социологических исследований.

Задачи ЭУМК:

- Обозначить статистическую природу данных социологических исследований.
- Рассмотреть процедуру статистического измерения, определив понятия признака и измерительной шкалы.
- Обучить расчету одномерных распределений, процедурам построения группировок, представления данных в графическом виде, нахождению характеристик центра распределения и показателей вариации признака.
- Ознакомить студентов с основами теории статистического вывода.
- Научить процедурам оценивания ошибки простой случайной выборки, расчету объема выборки, обеспечивающего репрезентативность данных исследования.
- Рассмотреть вопросы, связанные с проверкой статистических гипотез и определить их роль в цикле статистического анализа данных социологического исследования.
- Определить основные модели парной статистической связи, уделив особое внимание вопросам построения и анализа таблиц сопряженности.

В результате освоения учебной дисциплины студент должен ***знать*:**

- что представляет собой статистическое измерение и измерительные шкалы;
- формализацию данных социологического исследований в виде матрицы объект-признак;
- природу одномерных распределений изучаемых признаков и графического представления данных;
- принципы построения статистических группировок данных;

- основы статистического вывода;
- подходы к анализу парных статистических связей;

В результате изучения учебной дисциплины студенты должны ***уметь***:

- измерить изучаемые признаки и выразить их в шкальной форме;
- строить графики и распределения;
- проверять статистические гипотезы;
- определять взаимосвязи признаков;
- вычислять и анализировать статистические коэффициенты связи.

В результате изучения учебной дисциплины студенты должны ***владеть***:

- процедурами сбора и представления данных количественных социологических исследований;
- методами обработки и анализа статистической информации;
- методологией использования статистической информации в социологическом исследовании.

Структура учебной дисциплины

Дисциплина изучается в первом и втором семестрах очной формы получения высшего образования, в первом, втором и третьем семестрах –заочной формы получения высшего образования. Всего на изучение учебной дисциплины «Статистический анализ социологической информации» отведено:

– для очной формы получения высшего образования – 204 часа, в том числе 102 аудиторных часа, из них:

– в первом семестре: 52 аудиторных часа, из них: лекции – 26 часов, семинарские занятия – 24 часа, управляемая самостоятельная работа – 2 часа (ДО);

– во втором семестре: 50 аудиторных часов, из них: лекции – 26 часов, семинарские занятия – 22 часа, управляемая самостоятельная работа – 2 часа (ДО).

Трудоемкость учебной дисциплины составляет в первом семестре 3 зачетные единицы, во втором семестре 3 зачетные единицы.

Форма текущей аттестации в первом семестре – зачет, во втором семестре – экзамен.

– для заочной формы получения высшего образования – 52 аудиторных часа, из них:

– в первом семестре (установочная сессия): лекции – 10 часов, семинарские занятия – 4 часа;

– во втором семестре: лекции – 6 часов, семинарские занятия – 6 часов;

– в третьем семестре: лекции – 16 часов, семинарские занятия – 10 часов.

Трудоемкость учебной дисциплины составляет во втором семестре 3 зачетные единицы, в третьем семестре 3 зачетные единицы.

Форма текущей аттестации во втором семестре – зачет, в третьем семестре – экзамен.

ЭУМК состоит из теоретического, практического разделов, раздела контроля знаний, информационно-методического раздела. Теоретический раздел включает основополагающие темы учебной дисциплины, конспект текстов лекций. Представленный материал может быть использован для самостоятельной подготовки студентов к лекциям и практическим занятиям, контрольным мероприятиям по изучаемой учебной дисциплине (тесту, самостоятельным работам, управляемым самостоятельным работам, контрольным работам). Практический раздел содержит темы семинарских занятий с перечнем вопросов, списка первоисточников, примерами решения типовых задач. Раздел контроля знаний представлен перечнем практических заданий для выполнения на семинарских занятиях, тестом, вопросами к экзамену. Информационно-методический раздел содержит список рекомендуемой литературы.

1. ТЕОРЕТИЧЕСКИЙ РАЗДЕЛ

1.1. Статистика как наука. Данные социологического исследования, матрица «объект-признак»

Прежде чем стать наукой в ее современном понимании статистика прошла многовековую историю развития.

Числовые данные, относящиеся к тем или иным явлениям, начали применяться уже в глубокой древности. Так, известно, что еще за 5 тыс. лет до н. э. проводился подсчет населения в Китае, велся учет имущества в Древнем Риме, в Средние века проводились переписи населения, домашнего имущества, земель. Эти сведения использовались, в основном, в военных целях и при обложении налогами. В столь отдаленные времена осуществлялся лишь сбор статистических сведений, а их обработку и анализ, т. е. зарождение статистики как науки следует отнести ко второй половине 18 века.

Так, в это время возникает английская *школа политической арифметики*, основателями которой были В. Петти (1623-1687) и Дж. Граунт (1620-1674). Политические арифметики путем обобщения и анализа фактов стремились цифрами охарактеризовать состояние и развитие общества, показать закономерности развития общественных явлений, проявляющихся в массовом материале. Термин «статистика», происходящий от латинского слова status, что в средние века означало политическое состояние государства, введен в науку немецким ученым Готфридом Ахенвалем (1719-1772), в то время означавший *государствоведение*. С 1746 г. профессор физиологии и права Г.Ахенваль начал читать впервые в Марбургском университете новую дисциплину, которую он назвал статистикой. Основным содержание этого курса было описание политического состояния и достопримечательностей государства. Таким образом, статистика изначально выполняла только описательную функцию.

В первой половине XIX века возникает статистико-математическое направление. Родоначальником этого направления является бельгийский статистик А. Кетле (1796-1874гг.) основоположник учения о средних величинах, а также английские ученые Ф. Гальтон (1822-1911гг.), К. Пирсон (1875-1936гг.), В. Госсет (стьюдент) (1876-1937гг.), Р. Фишер (1890-1962гг.), которые ввели в статистику теорию вероятностей, которая позволила шагнуть от описательной статистики к аналитической.

Статистические законы, хотя и не дают однозначных и достоверных предсказаний, тем не менее, являются единственно возможными при исследовании массовых явлений случайного характера.

В настоящее время термин статистика имеет несколько значений. Употребляемый во множественном числе (*статистики*) он относится к области описательной статистики, обозначая собранные статистические данные, а также вычисленные статистические характеристики этих данных, в единственном числе (*статистика*) он обозначает статистическую теорию и методы, при помощи которых анализируются данные. Таким образом, термин применим как к интерпретации набора чисел, так и к самим числам.

Статистические принципы лежат в основе всех количественных методик сбора и анализа первичных данных. А именно то, что массовые явления имеют статистический характер, и если изучить достаточно большое количество проявлений изучаемого явления, то само явление будет познано. Объекты взаимозаменяемы, и их индивидуальные особенности, как таковые, не представляют для исследователя особого интереса. В этом случае достаточно иметь репрезентативную выборку, которая позволяет результаты, полученные по выборке, распространить на генеральную совокупность.

Как правило, статистический подход заключается в том, что исследователя интересует не тот факт, что гражданин Ваня Иванов предпочитает утолять жажду напитком «Фанта», а более общее явление, например, что среди молодых людей в возрасте от 18-25 лет велика доля тех, кто из всех марок безалкогольных газированных напитков предпочитает напиток «Фанта».

Опр. Статистика – это наука, занимающаяся сбором, хранением, и анализом данных о массовых явлениях и процессах.

Из определения следуют три основные задачи, которые выполняет статистика. Это:

- сбор данных;
- хранение данных;
- анализ данных.

Статистическому анализу подвергается не индивид, а статистическая совокупность индивидов (население страны, отдельного населенного пункта, клиенты конкретной организации, учащаяся молодежь и т. д.). Однако изучаемая совокупность часто не может быть обследована полностью, и выводы о всей совокупности делают на основании выводов об отдельной ее части. Поэтому в статистике вводятся два основополагающих понятия, которые лежат в основе всех количественных методик, используемых в социологических исследованиях: генеральная совокупность и выборка.

Опр. Генеральная совокупность (ГС) – это полная совокупность объектов, имеющих отношение к изучаемой проблеме (все население страны, молодежь, клиенты фирмы, сеть магазинов и т. д.).

Опр. Выборочная совокупность (выборка) – это та часть генеральной совокупности, которую мы непосредственно изучаем.

Таким образом, сначала определяются границы генеральной совокупности, затем из ГС отбирается выборочная совокупность, которая непосредственно изучается и описывается, а затем, по итогам выборочного исследования, делаются выводы относительно всей генеральной совокупности.

В связи с тем, что, используя статистические методы, необходимо сначала определить, изучить и описать выборочную совокупность а затем распространить знания о выборочной совокупности на всю генеральную совокупность, статистика делится на три основных направления.

1. *Дескриптивная (описательная) статистика*. Занимается описанием наблюдаемой выборочной совокупности.

2. *Теория статистического вывода.* Занимается обобщением результатов выборочного исследования на генеральную совокупность

3. *Аналитическая статистика*. Занимается анализом взаимосвязей двух или более признаков.

Данные социологического исследования. Данные социологического исследования с точки зрения статистики представляют собой совокупность изучаемых объектов, каждому из которых поставлен в соответствие набор интересующих исследователя признаков.

Опр. Признак – это некоторое общее для всех объектов свойство (качество), конкретные проявления которого могут меняться от объекта к объекту.

Опр. Все возможные проявления признака называются значениями данного признака.

Например, признак «Пол» имеет значения: мужской; женский.

Признак «Место жительства» имеет значения: столица; областной центр; районный центр; малый город; село.

Признак «Семейное положение» имеет значения: женат (замужем); холост (не замужем); разведен (разведена); вдовец (вдова).

Если учесть, что при проведении социологического исследования число объектов (объем выборки) часто может превышать тысячу человек, а число признаков исчисляется десятками, то перед социологами встает вопрос: «Как должны быть представлены данные, чтобы могла быть осуществлена их статистическая обработка?».

Статистические методы могут быть применены к данным только после перевода всех значений признака с вербального представления на язык чисел.

Мы должны определенным образом формализовать исходные данные, чтобы затем их подвергать статистической обработке. Процедура перевода вербальных значений на язык чисел называется измерением.

Опр. Измерение – это определенная процедура приписывания символов (чаще всего чисел) значениям признака в соответствии с определенным правилом.

Цель измерения – получить числовую модель, в которой формируются такие взаимосвязи между переменными, что ее исследование могло бы заменить исследование самого исходного объекта.

Это возможно лишь тогда, когда свойства модели соответствуют свойствам объекта, т. е. отношения между числами, образующими числовую модель, соответствуют отношениям между изучаемыми свойствами объекта.

Исходной числовой моделью данных социологических исследований является матрица «объект-признак» (Таблица 1), каждая строка которой соответствует одному объекту, а каждый столбец – одному признаку (переменной). В матрице «объект-признак» на пересечении i-ой строки и j-ого столбца располагается результат измерения j-ого признака для i-ого объекта.

Таблица 1 – Пример построения матрицы «объект-признак»

№	Пол респондента 1. мужской 2. женский	Марка личного автомобиля 1. Мерседес 2. Ауди 3. Фольксваген	Удовлетворенность своим авто 1. полностью уд. 2. скорее уд. 3. и да, и нет 4. скорее неуд. 5. полностью неуд.
1	1	1	1
2	2	3	3
3	1	2	2
4	1	3	2
5	1	3	2
6	2	1	3

Чтобы реализовать процедуру измерения, нам необходим инструмент, который позволит перевести вербальные значения признака в числовые коды.

1.2. Измерение в социологии, измерительные шкалы

Инструментом формализации значений признаков выступает измерительная шкала, которая набору свойств изучаемого объекта ставит в соответствие набор сопоставляемых им чисел.

Признаки делятся на качественные и количественные:

1. Качественные признаки описывают свойства объекта, которые выражаются с помощью понятий и определений (вербально).
2. Количественные признаки описывают свойства объекта, которые выражаются с помощью чисел.

Измерительные шкалы различают по уровню измерения исходного признака. В статистике применяются четыре классические измерительные шкалы.

Для измерения качественных признаков используют:

- номинальную (категориальную) шкалу;
- порядковую шкалу.

Для измерения количественных признаков используют:

- интервальную шкалу;
- шкалу отношений.

Процедура измерения качественных признаков отличается от процедуры измерения количественных признаков тем, что числа, приписываемые значениям качественного признака, только обозначают определенную категорию (значение признака), к ним не применимы никакие арифметические правила для чисел.

При измерении количественных признаков для чисел, соответствующих значениям признака, могут выполняться арифметические правила (больше-меньше, во сколько раз больше-меньше).

Номинальная шкала. В случае номинальной шкалы имеем дело с самым низким уровнем измерения, потому что в рамках этого уровня моделируются самые простые отношения между объектами измерения, а именно, отношения равенства и неравенства

Опр. Номинальная шкала – это такая измерительная шкала, по которой в процессе измерения мы устанавливаем отношение равенства-неравенства объекта значению признака.

Пример номинальной шкалы:

Где Вы обычно покупаете товар «N»?

1. Гипермаркет
2. Магазин
3. Торговая палатка, лоток
4. Рынок
5. С рук (у знакомых)

Шкала для измерения значений признака «Где Вы обычно покупаете товар X?» строится из 4-х шкальных значений.

Номинальная шкала предназначается для разбиения изучаемой совокупности на непересекающиеся классы.

Качественные шкалы в социологических исследованиях встречаются довольно часто. Их основным недостатком является ограничение применения возможных методов анализа первичных данных. Для того, чтобы избежать ограничений, применяют процедуру преобразования номинальной шкалы в дихотомическую.

Опр. Дихотомическая шкала – это шкала, у которой имеется только два значения признака – 1 и 0, что обозначает наличие или отсутствие определенного значения признака у изучаемого объекта. К дихотомическому виду может быть приведена любая номинальная шкала.

Например, признак Пол: 1. Мужчина 2. Женщина

Может быть представлен в виде двух дихотомических признаков:

Респондент является мужчиной: 1. Да 0. Нет

Респондент является женщиной: 1. Да 0. Нет

Перевод в дихотомический вид любого номинального признака осуществляется, как правило, на этапе статистического анализа данных.

Порядковая шкала. Для построения порядковой шкалы необходимо уметь устанавливать не только отношения равенства-неравенства, но и отношения последовательности или порядка.

Опр. Порядковая шкала – это такая измерительная шкала, для которой в процессе измерения, наряду с отношением равенства-неравенства устанавливается отношение порядка. Отношение порядка – это отношение типа «больше, чем», «лучше, чем».

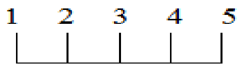
С помощью порядковых шкал в социологических исследованиях чаще всего измеряются показатели удовлетворенности.

Пример порядковой шкалы:

Насколько вы удовлетворены обслуживанием в нашем отеле?

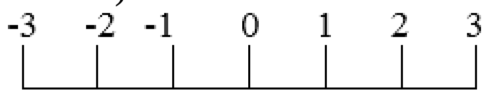
1. Полностью удовлетворен
2. Скорее удовлетворен, чем нет
3. И да, и нет
4. Скорее не удовлетворен
5. Совершенно не удовлетворен

С помощью данного вопроса можно упорядочить клиентов гостиницы по степени их удовлетворенности обслуживанием.

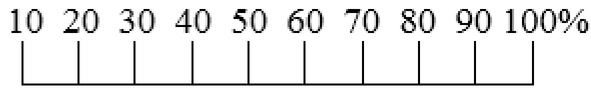
Приведенный пример порядковой шкалы – это пример порядковой шкалы с вербальной интерпретацией шкальных значений. Графически ее можно изобразить следующим образом. 

Переход от вербального представления порядковой шкалы к графическому позволяет рассматривать такую шкалу как псевдоинтервальную, т. е. делать предположение, что расстояние между делениями шкалы равны, что позволяет не только упорядочить значения, но и сравнивать их между собой.

Преобразование вербальной порядковой шкалы в графическую дает возможность увеличения числа делений на шкале. Так, вопрос об удовлетворенности обслуживанием в гостинице, может быть сформулирован: «Отметьте на шкале то положение, которое соответствует вашей удовлетворенности обслуживанием в нашем отеле?» (*где -3 соответствует ответу «Совсем не удовлетворен обслуживанием», а 3 «Полностью удовлетворен обслуживанием»*)



Также в графической интерпретации может быть использована шкала процентов. Тогда вопрос будет звучать: «Отметьте, насколько процентов Вы удовлетворены обслуживанием в нашем отеле?»



Две графические интерпретации теоретически являются порядковой шкалой, однако исследователь может работать с такой шкалой уже как с количественной интервальной шкалой.

Интервальная шкала. В основе построения интервальной шкалы лежит эмпирическая процедура, позволяющая определить равенство дистанций между парами объектов. Отличием интервальной шкалы является то, что в ней произволен выбор точки отсчета 0 и произволен масштаб (цена деления может быть разная).

Опр. Интервальная шкала – это такая измерительная шкала, для которой в процессе измерения, наряду с отношением равенства-неравенства и порядка устанавливается отношение «на сколько меньше», «на сколько больше».

Классическими примерами интервальной шкалы являются все температурные шкалы и шкала летоисчисления. В социологических исследованиях чаще применяют псевдоинтервальные шкалы.

Шкала отношений. Самая сильная измерительная шкала. Используется для измерения количественных признаков, для которых можно точно установить исходную точку (0).

Для шкалы отношений, наряду со всеми вышеперечисленными отношениями, характерными для шкал более низкого измерения, а именно отношений равенства-неравенства, упорядоченности, больше-меньше, справедливо и отношение во сколько раз.

Опр. Шкала отношений – это такая измерительная шкала, для которой в процессе измерения, наряду с отношением равенства-неравенства, порядка и «на сколько меньше», «на сколько больше» добавляется отношение «во сколько раз меньше», «во сколько раз больше».

В социологических исследованиях такие шкалы используются для измерения «физических» величин: времени, возраста, стажа, дохода и т. д.

Количественные признаки подразделяются на дискретные и непрерывные.

Опр. Дискретные – это такие признаки, у которых значения изменяются скачкообразно (количество детей в семье, стаж работы)

Опр. Непрерывные – это такие признаки, значения которых могут принимать промежуточные значения (возраст, доход).

Необходимо отметить, что каждый последующий тип шкалы содержит возможности предыдущей и одновременно добавляет свои, увеличивающие «силу» измерения (Таблица 2).

Таблица 2 – Возможности измерения четырех измерительных шкал

Тип шкалы	Возможности измерения
Номинальная	Идентифицирует изучаемый объект с точки зрения присутствия или отсутствия определенных значений признака.
Порядковая	Как выше + упорядочивание, приписывание изучаемому объекту определенного значения признака в соответствии с выражением его интенсивности. Интенсивность (критерий порядка) не может быть точно измерена.
Интервальная	Как выше + возможность сравнивать полученные значения признака. Предполагается, что интервалы на шкале одинаковые, но ноль не определен.
Шкала отношений	Как выше +возможность сравнивать во сколько раз одно значение признака больше или меньше другого. Абсолютный ноль.

1.3. Одномерное частотное распределение

Первой статистической процедурой анализа данных, представленных в виде матрицы «Объект-признак», является расчет одномерного частотного распределения. Чтобы получить одномерное частотное распределение подсчитывают, сколько объектов обладают данным значением признака.

Если X– некоторый признак, то X_i – одно из значений признака, где i изменяется от 1 до k (k – число значений данного признака), тогда f_i (frequencies/частота) называется распределением по данному признаку.

Опр. Совокупность значений признака и их частот называется одномерным частотным распределением.

Для данных, представленных в виде таблицы «объект-признак» (Таблица 3) рассмотрим одномерное частотное распределение для признака «Пол»

Таблица 3. – Одномерное частотное распределение качественного признака

X_i пол	f_i	f_i (%)	f_i (доли)
X_1 мужской	2	33,3	0,333
X_2 женский	4	66,7	0,667
Всего	n = 6	100	1

В таблице одномерного частотного распределения могут быть представлены два вида частот: абсолютные и относительные.

Опр. Абсолютная частота (f_i) – это количество объектов, имеющих то или иное значение признака.

Опр. Относительные частоты (f_i%) – это отношение абсолютной частоты к общему числу изучаемых единиц, умноженное на 100%.

Относительные частоты могут рассчитываться двумя способами:

1. относительная частота в процентах (f_i%);
2. относительная частота в долях от 1 (f_i доли).

Формула расчета относительных частот в процентах: $\mathrm{f_i}\% = \frac{f_i}{n}100\%$

Формула расчет относительных частот в долях от единицы: $\mathrm{f_i}\ \text{доли} = \frac{f_i}{n}$, где

f_i – абсолютная частота, соответствующая значению признака;

n – объем выборки.

Таблица одномерного частотного распределения количественного признака будет иметь следующий вид (Таблица 4).

Таблица 4 – Одномерное частотное распределение количественного признака

Сколько часов в неделю Вы проводите за рулем своего авто?	f_i	f_i (%)	$\mathrm{F}_i\%$	$\mathrm{F}_i^*\%$
5	2	33,3	33,3	100
9	1	16,7	50,0	66,7
10	2	33,3	83,3	50,0
20	1	16,7	100	16,7
Всего	6	100	-	-

Для количественного признака одномерное частотное распределение также называют вариационным рядом.

Значение признака называют вариантом, а число объектов, обладающих данным значением, частотой. Варианты вместе с частотами образуют вариационный ряд данного признака или одномерное частотное распределение по данному количественному признаку.

Для всех количественных и порядковых признаков можно рассчитать еще один вид частот, который называется Накопленная частота.

Накопленная частота бывает абсолютной и относительной, возрастающей и убывающей:

- абсолютная накопленная частота рассчитывается по абсолютным частотам распределения;
- относительная накопленная частота рассчитывается из относительных частот в процентах;
- возрастающая накопленная частота рассчитывается суммированием от первого значения в вариационном ряду к последнему.
- убывающая накопленная частота рассчитывается суммированием от последнего значения в вариационном ряду к первому.

Относительная возрастающая накопленная частота обозначается $F_i\%$.

Относительная убывающая накопленная частота обозначается $F_i^*\%$.

1.4. Группировки количественных признаков в интервалы

Количественные признаки, имеющие достаточно большие разбросы значений, принято группировать в интервалы. При этом понятие частоты для сгруппированных данных относится к интервалам.

Для сгруппированных данных абсолютная частота – это количество объектов из выборки значения, которых попали в данный интервал.

Проводя группировку в интервалы, необходимо решить ряд задач, чтобы группировка наиболее точно характеризовала распределение изучаемого признака.

Необходимо определить:

1. сколько интервалов будет содержать данная группировка;
2. какова будет длина интервалов;
3. интервалы будут одинаковой или разной длины;
4. каким образом будут определены границы интервалов;
5. все ли интервалы будут закрытыми или нет.

Выполнение этих действий необходимо для того, чтобы все возможные значения признака имели свое место в группировке, а также, чтобы каждое значение признака могло войти только в один интервал.

Наиболее важным из перечисленных условий является определение границ интервалов. Интервалы должны быть представлены в таком виде, чтобы, во-первых, их границы не пересекались, во-вторых, чтобы для каждого из объектов мог быть определен только один интервал.

Данная проблема имеет следующее решение. Если группируем дискретную переменную, то используем группировку с непересекающимися границами интервалов, например, 0-6 и 7-10.

Если группируем непрерывную переменную, указываем интервалы, с пересекающимися границами, но при этом однозначно оговариваем, какая из границ входит в интервал. Например, если есть два интервала 0-6 и 6-10, для того, чтобы определить в какой из двух интервалов входит пересекающееся значение 6, поступают следующим образом:

а) указывают, что интервалы закрыты сверху или справа, в этом случае все значения признака от 0 до 6 включительно относятся к интервалу 0-6, все значения, превышающие 6 и до 10 относятся ко второму интервалу;

б) указывают, что интервалы закрыты снизу или слева, в этом случае значение признака равное 6, попадает во второй интервал.

Существует три типа группировок:

1. типологическая;
2. аналитическая;
3. процентильная.

Типологическая группировка. При построении типологической группировки не применяются какие-либо конкретные методики, а производится деление на интервалы в соответствии с задачами исследования и теоретического представления о том, как изменяется отношение к предмету исследования в зависимости от значения признака. Типологическая группировка, как правило, строится с непересекающимися границами и определенным небольшим числом интервалов, таки образом, чтобы затем каждый интервал мог получить свое наименование.

Примеры типологических группировок:

1. Доход за месяц в бел. рублях

До 500;
501 – 900;
901 – 1300;
1301 – 2500;
более 2500.

2. Возраст респондентов

18 – 24;
25 – 34;
35 – 50;
51 – 64;
65 и старше

Аналитическая группировка. ***Опр.*** Аналитическая группировка – это разбиение на заданное число небольших интервалов равной длины.

Условие деления на заданное число небольших интервалов равной длины решается за счет расчета длины интервала.

1. Определяем размах вариации $d = x_{max} - x_{min}$

2. Вычисляем длину интервала: если число наблюдений меньше 100 по формуле d/7<L<d/6, если больше 100 по формуле d/15<L<d/12

Числа в дроби указывают на число интервалов, которые будут в группировке. Если объем выборочной совокупности небольшой и признак имеет относительно небольшое число значений, то длину интервала рассчитывают исходя из построения 7/6 интервалов. Если объем выборочной совокупности большой и изучаемый признак имеют достаточно большое число значений, то длину интервала рассчитывают исходя из построения 15/12 интервалов.

Необходимо помнить, что длина интервала (L) всегда целое число, лежащее в границах вычисленного соотношения: d/7<L<d/6 или d/15<L<d/12.

Если вычисленные значения границ таковы, что невозможно определить целое значение длины интервала, осуществляют ближайшее округление до целого значения.

Пример построения аналитической группировки, имеются данные о стаже руководителей (Рисунок 1).

стаж работы руководителей

		Частота	Процент	Валидный процент	Кумулятивный процент
Валидные	3	3	7,0	7,0	7,0
	4	4	9,3	9,3	16,3
	6	4	9,3	9,3	25,6
	7	4	9,3	9,3	34,9
	8	7	16,3	16,3	51,2
	9	4	9,3	9,3	60,5
	10	5	11,6	11,6	72,1
	11	2	4,7	4,7	76,7
	12	1	2,3	2,3	79,1
	13	1	2,3	2,3	81,4
	14	2	4,7	4,7	86,0
	15	1	2,3	2,3	88,4
	17	1	2,3	2,3	90,7
	18	2	4,7	4,7	95,3
	20	1	2,3	2,3	97,7
	26	1	2,3	2,3	100,0
	Итого	43	100,0	100,0	

Рисунок 1 – Вариационная таблица, содержащая данные о стаже работы руководителей (в годах)

Чтобы построить аналитическую группировку рассчитаем длину интервала:

1. Вычислим размах вариации: d = 26-3= 23

2. Определим длину интервала:

d/7<L<d/6,

23/7<L<23/6,

3,28<L<3,8 округляем до целого и получаем L=4

3. По исходной вариационной таблице строим группировку прибавляя к левой границе значений длину интервала L=4. Для каждого интервала подсчитываем абсолютную частоту, соответствующую данному интервалу, затем вычисляем все относительные частоты. В итоге получаем группировку (Таблица 5).

Таблица 5 – Аналитическая группировка для данных о стаже руководителей

Стаж работы руководителей (в годах)	Частота	Процент f_i%	Накопленная частота F_i%
3-7	15	34,9	34,9
7-11	18	41,9	76,8
11-15	5	11,6	88,4
15-19	3	7,0	95,4
19-23	1	2,3	97,7
23-27	1	2,3	100
Всего	n= 43	100	-

Процентильная группировка. ***Опр***. Процентильная группировка – это разбиение на заданное число интервалов, содержащих равный процент наблюдений (респондентов).

Для построения процентильной группировки вводится понятие **квантиля**.

Опр. Квантиль (X_p) – это значение признака, которое делит распределение изучаемого признака таким образом, что слева от квантиля находится заданная доля наблюдений (p).

Например, если $X_p = X_{0,25}$ – доля наблюдений слева от квантиля будет равна 25%.

В зависимости от числа квантилей существует несколько видов процентильных группировок.

Опр. Квартильная группировка – это разбиение на четыре интервала по 25% наблюдений в каждом интервале.

Для построения квартильной группировки необходимо использовать значения крайних точек X_{min} ; X_{max} и определить три квантиля (квартиля): $X_{0,25}$; $X_{0,5}$;$X_{0,75}$.

Опр. Квинтильная группировка – это разбиение на пять интервалов по 20% наблюдений в каждом интервале.

Для построения квинтильной группировки необходимо использовать значения крайних точек X_{min} ; X_{max} и определить четыре квантиля : $X_{0,2}$; $X_{0,4}$; $X_{0,6}$; $X_{0,8}$.

Опр. Терцильная группировка – это разбиение на три интервала по 33,3% наблюдений в каждом интервале.

Для построения терцильной группировки необходимо использовать значения крайних точек X_{min} ; X_{max} и определить два квантиля : $X_{0,333}$; $X_{0,666}$.

Опр. Децильная группировка – это разбиение на десять интервалов по 10% наблюдений в каждом интервале.

Для построения децильной группировки необходимо использовать значения крайних точек X_{min} ; X_{max} и определить девять квантилей : $X_{0,1}$; $X_{0,2}$; $X_{0,3}$; $X_{0,4}$; $X_{0,5}$; $X_{0,6}$; $X_{0,7}$; $X_{0,8}$; $X_{0,9}$.

Наиболее часто используемой процентильной группировкой является квартильная группировка.

Чтобы построить процентильную группировку необходимо уметь определять значения квантилей.

Определить квантиль можно двумя способами:

1. Приблизительно по возрастающей накопленной частоте вариационного ряда.

2. Точно, используя формулу и предварительно построенную аналитическую группировку.

Определение квантиля по возрастающей накопленной частоте. Чтобы определить квантиль необходимо найти возрастающую накопленную частоту равную или впервые превысившую значение процента, искомого квантиля. Тогда, значение признака, соответствующее данной накопленной частоте и будет значением квантиля.

Определение квантиля по формуле и аналитической группировке. Чтобы воспользоваться формулой для точного определения квантиля необходимо предварительно построить аналитическую группировку. По возрастающей накопленной частоте определить интервал, содержащий квантиль и далее применить формулу для вычисления квантиля.

Формула расчета квантиля:

$$x_{p=} x_0 + L\frac{p\times 100\% - F_0}{f_p}, \text{ где}$$

x_0 – левая граница интервала, в котором находится искомый квантиль;

L – длина интервала;

F_0 – предшествующая накопленная частота;

f_p – частота, соответствующая интервалу, содержащему квантиль.

Рассмотрим пример построения квартильной группировки с точными границами. В качестве исходных данных воспользуемся аналитической группировкой для стажа руководителей, которую мы построили в предыдущем примере (Таблица 5).

Шаг 1. По возрастающей накопленной частоте определяем интервал, содержащий квантиль $X_{0,25}$. Это будет интервал 3-7 лет.

Шаг 2. Вычисляем значение квантиля $X_{0,25}$.

$$X_{0,25} = 3 + 4\frac{0,25*100\% - 0}{34,9} = 5,9$$

Шаг 3. По возрастающей накопленной частоте определяем интервал, содержащий квантиль $X_{0,5}$. Это будет интервал 7-11 лет.

Шаг 4. Вычисляем возрастающей значение квантиля $X_{0,5}$

$$X_{0,5} = 7 + 4\frac{0,5*100\% - 34,9}{41,9} = 8,4$$

Шаг 5. По возрастающей накопленной частоте определяем интервал, содержащий квантиль $X_{0,75}$. Это будет интервал 11-15 лет.

Шаг 6. Вычисляем значение квантиля $X_{0,75}$

$$X_{0,75} = 7 + 4*\frac{0,75*100\% - 34,9}{41,9} = 10,8$$

Вычисленные значения квантилей будут соответствовать границам интервалов квартильной группировки. Каждый интервал в квартильной группировке будет содержать по 25% (Таблица 6).

Таблица 6 – Квартильная группировка с точными границами для данных о стаже работы руководителей

Стаж работы руководителей	Проценты f_i%)
3-5,9	25
5,9-8,4	25
8,4-10,8	25
10,8-26	25
Всего	100

1.5. Графическое представление социологических данных

Данные социологических исследований часто представляются в графическом виде.

Опр. Графическое представление данных – это метод условных изображений распределения данных при помощи линий, точек, геометрических фигур и других символов.

Существуют десятки видов графиков, которые помогают визуально представить распределение того или иного изучаемого признака или совместное распределение нескольких признаков.

Условно графики можно разделить на:

– Презентационные, предназначенные для улучшения восприятия аудиторией результатов исследования.

– Научные или аналитические, предназначенные для более глубокого анализа распределения изучаемого признака.

При построении графиков необходимо учитывать свойства измерительной шкалы и основные правила построения графиков.

Основные требования к построению графиков.

1. График обязательно должен иметь заголовок, т.е. указание того, что представлено на данном графике.
2. Геометрические фигуры, представляющие распределение, должны сохранять пропорцию, фиксирующуюся в распределении признака, представленного на графике.
3. График должен сопровождаться легендой, т.е. словесным или символьным описанием значений и частот, представленных на графике.

Рассмотрим основные виды графиков и правила их построения.

Диаграммы. Диаграммы – наиболее часто используемый вид графиков, предназначенный для презентации данных.

Круговая диаграмма. Круговая диаграмма применяется для представления качественных признаков, измеренных с помощью номинальной шкалы.

Опр. Круговая диаграмма представляет собой окружность, на которой частоты представлены в виде секторов, началом отсчета является нулевой радиус, выходящий из центра перпендикулярно вверх (Рис 2). Чтобы определить площадь сектора необходимо составить пропорцию:

$$360^0 - 100\%$$
$$x^0 - f_i\%$$

Например, нужно построить круговую диаграмму для распределения признака «Материальное положение семьи». Используем указанную пропорцию (Рисунок 2):

$$x^0 = 360^0\, 31{,}3/100\% = 111{,}9^0$$
$$x^0 = 360^0\, 63/100\% = 226{,}8^0$$
$$x^0 = 360^0\, 5{,}9/100\% = 21{,}2^0$$

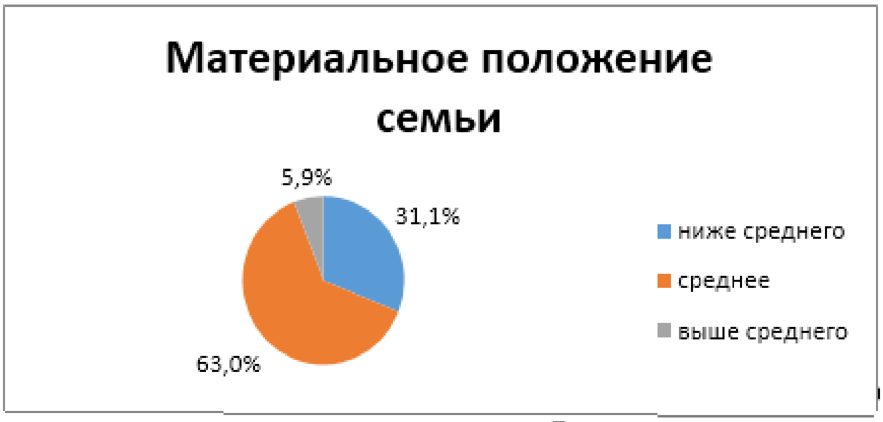


Рисунок 2 – Круговая диаграмма

Диаграмма столбцов (столбиковая диаграмма) и диаграмма полос. Столбиковая диаграмма и диаграмма полос строится для номинальных, порядковых и дискретных количественных шкал.

Опр. Столбиковая диаграмма представляет собой набор прямоугольников с равными основаниями, расположенными на одинаковом расстоянии друг от друга.

Если основания располагаются по оси X, то это столбиковая диаграмма, если основания располагаются по оси У – это диаграмма полос.

Высота столбика и длинна полоски соответствуют частоте данного значения признака. Частота может выражаться как в абсолютных значениях, так и в относительных в % (Рисунок 3).

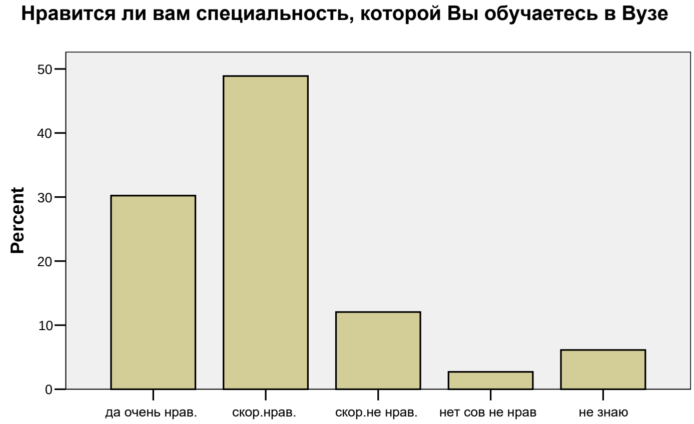


Рисунок 3 – Диаграмма столбцов (столбиковая диаграмма)

Ленточная (штабельная) диаграмма. Применяется для номинальных и порядковых шкал. Предназначена для изображения структуры распределения.

Опр. Ленточная диаграмма представляет собой полоску, длина которой равняется 100%. Внутри полоска разделена на сектора пропорционально частотам каждого значения признака (Рисунок 4).

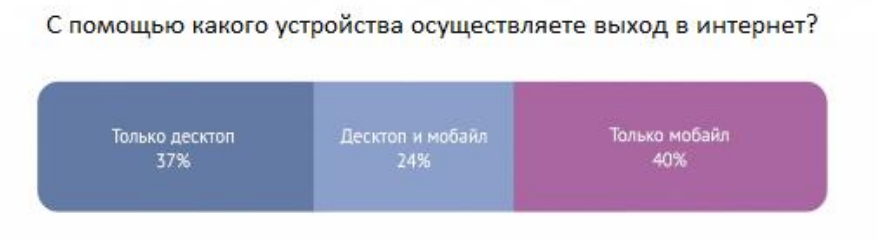


Рисунок 4 – Ленточная диаграмма

Опр. Штабельная диаграмма представляет собой вертикальную полоску, внутри разделенную на сектора пропорционально частотам каждого значения признака. Если штабельная диаграмма построена по относительным частотам в процентах, то сумм частот каждого столбика равна 100% (Рисунок 5).

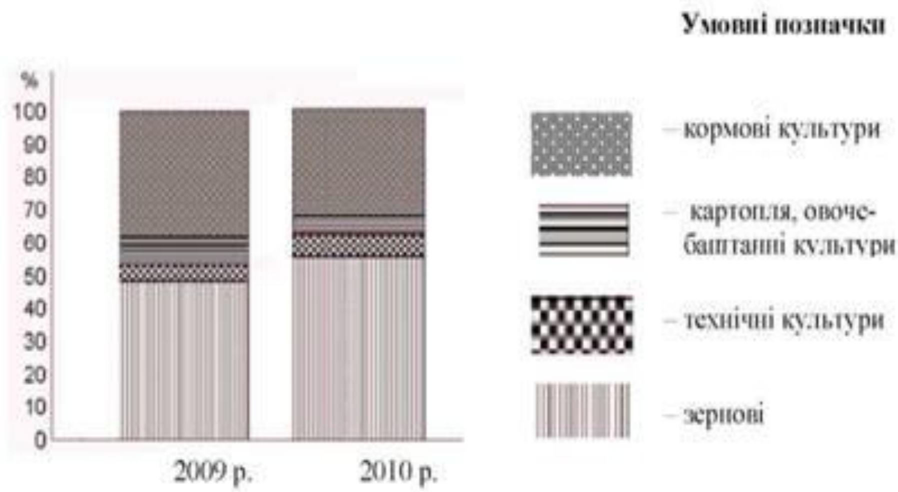


Рисунок 5 – Штабельная диаграмма

Особенно эффективна при сравнительном анализе структур нескольких совокупностей, а также при анализе динамики структуры во времени.

Гистограмма. Применяется для количественных непрерывных признаков, сгруппированных в интервалы. Предназначена для анализа формы распределения изучаемого признака.

Опр. Гистограмма представляет собой последовательность примыкающих друг к другу прямоугольников, основаниями которых служат интервалы группировки, а высотой плотность распределения (Рисунок 6).

Рисунок 6 – Гистограмма

Построение гистограммы. По оси X откладываются значения точных границ интервалов, стрелка оси указывает направление возрастания значений признака. По оси У откладывается плотность распределения. Если гистограмма строится по аналитической группировке по оси У откладывается соответствующая частота.

Плотность распределения показывает значение частоты из расчета на одну единицу измерения признака и вычисляется по формуле:

$\rho_i = n_i/l_i$, где

n_i – частота попадания в интервал

l_i – длина интервала

ρ_i – плотность

Полигон распределения. Полигон распределения используется для представления формы распределения количественного признака. Строится по одномерному частотному распределению или по группировке в интервалы.

Опр. Полигон распределения представляет собой ломаную линию, для построения которой по оси X откладываются значения признака, а по оси У частоты соответствующие этим значениям (Рисунок 7).

Рисунок 7 – Полигон распределения

Если полигон распределения строится для сгруппированных данных, то в качестве координаты по оси X берется середина интервала, а в качестве координаты по оси У значение соответствующей частоты (или плотности, если интервалы не равны между собой).

График интерквартильного диапазона. Используется для представления количественных признаков, сгруппированных с помощью квартильной группировки.

Строится по одной оси, на которой последовательно располагаются значения крайних точек и квантилей X_{min}; $X_{0,25}$; $X_{0,5}$; $X_{0,75}$;X_{max}.

Интерквартильный диапазон представляется в виде прямоугольника, к которому примыкают «усы» до минимального и максимального значения. Внутри интерквартельного диапазона помечается значение медианы.

Опр. Интерквартильным диапазоном называется разность между квантилями $X_{0,75}$ и $X_{0,25}$.

График интерквартильного диапазона часто используется для сравнения распределения в разных группах респондентв (Рис 8).

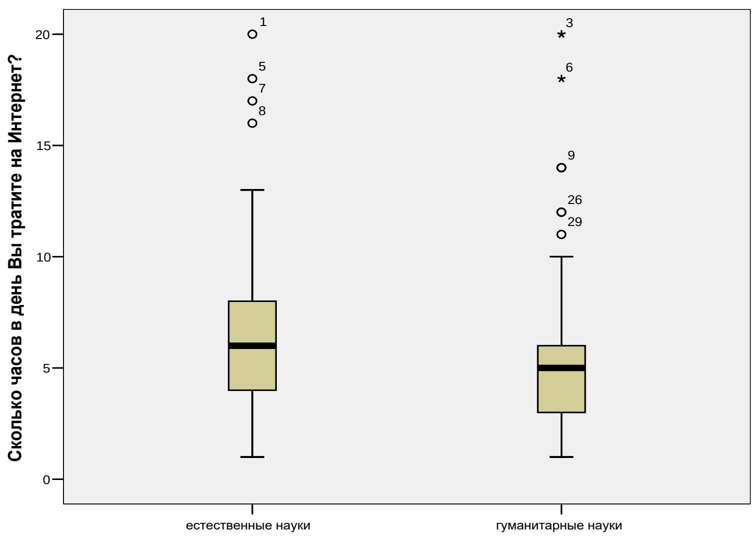


Рисунок 8 – График интерквартельного диапазона

1.6. Характеристики центра распределения признака

Характеристики центра распределения (меры центральной тенденции) – это типичные значения в распределении изучаемого признака.

Измерение центральной тенденции состоит в выборе одного числа, которое наилучшим образом описывает все значения изучаемого признака.

Характеристиками положения центра распределения являются три показателя:

- Мода (M_o)
- Медиана (M_e)
- Среднее арифметическое ($\overline{x}$)

*Мода. **Опр***. Модой называется значение признака, обладающее наибольшей частотой.

Моду можно определить для любой измерительной шкалы. Для номинальных, порядковых и дискретных количественных переменных мода определяется по одномерному частотному распределению. Значение признака, обладающее наибольшей частотой, является модальным значением.

Например, необходимо определить значение моды для распределения ответов студентов-первокурсников (Таблица 7).

Таблица 7 – Распределение ответов студентов-первокурсников на вопрос «Какое среднее учебное заведение Вы Закончили?»

Какое среднее учебное учреждение Вы закончили?	Частота f_i	Процент f_i (%)
Средняя школа	18	72,0
Гимназия	4	16,0
Лицей	3	12,0
Всего	25	100

Большинство респондентов (72%) указали, что закончили СШ (среднюю школу), значит модой в данном распределении будет значение признака – средняя школа (M_o=средняя школа). Полученное значение моды можно проинтерпретировать следующим образом – большинство студентов-первокурсников закончили среднюю школу.

Чтобы определить моду для непрерывного количественного признака необходимо предварительно сгруппировать данные с помощью аналитической группировки. По частотному распределению аналитической группировки определяем интервал, содержащий моду (модальный интервал).

Опр. Модальный интервал, это интервал, которому соответствует наибольшая частота.

Внутри модального интервала вычисляем значение моды по формуле:

$$\mathrm{M_o} = x_0 + L\frac{f_{Mo} - f^-}{2f_{Mo} - (f^- + f^+)}, \text{ где}$$

x_0 – левая граница модального интервала;

L – длина интервала;

f_{M0} – частота в модальном интервале;
f^- – частота в интервале, предшествующем модальному;
f^+– частота в интервале, следующем за модальным.

Медиана. ***Опр.*** Медиана – это значение признака, которое делит упорядоченное множество данных пополам, таким образом, что число единиц совокупности с большим и меньшим чем медиана значением признака, одинаково. Медиана – это 50-й процентиль ($M_e = X_{0,5}$)

Медиану можно определить только для признаков, обладающих свойством порядка (порядковых шкал, интервальных/псевдоинтервальных шкал и шкал отношений).

Нахождение медианы аналогично нахождению квантиля $X_{0,5}$:

1. Приблизительное значение медианы можно найти по возрастающей накопленной частоте вариационного ряда.

2. Точное значение медианы можно найти по формуле для сгруппированных данных.

3. Для сгруппированных данных медиану можно найти по пересечению кривых огивы.

Например, чтобы определить медиану по вариационному ряду признака «Стаж работы руководителей» (Таблица 6) необходимо найти возрастающую накопленную частоту равную или впервые превысившую 50%, соответствующее значение признака и будет искомой медианой. M_e=8 лет.

Чтобы определить точное значение медианы для признака «Стаж работы руководителей» предварительно необходимо построить аналитическую группировку (Таблица 7), по возрастающей накопленной частоте найти медианный интервал (7-11) и применить формулу для вычисления 50-го процентиля. В итоге получаем точное значение медианы M_e=8,4 года.

$$M_e = x_0 + L\frac{50 - F_0}{f_p}, \quad \text{где}$$

x_0 - левая граница медианного интервала;
L - длина интервала;
F_0- предшествующая накопленная частота;
f_p - частота, соответствующая медианному интервалу.

*Среднее арифметическое****. Опр.*** Среднее арифметическое – это сумма всех значений признака деленое на число наблюдений.

Среднее арифметическое рассчитывается только для количественных признаков. Формула расчета среднего арифметического зависимости от того как представлены исходные данные: в «сыром» виде, в виде вариационного ряда, в виде группировки.

1. Данные представлены в «сыром» виде: $\bar{x} = \frac{\sum x_i}{n}$

2. Данные представлены в виде вариационного ряда: $\bar{x} = \frac{\sum x_i f_i}{n}$

3. Данные представлены в сгруппированном виде: $\bar{x} = \frac{\sum \tilde{x}_i f_i}{n}$

Например, чтобы определить среднее арифметическое по вариационному ряду признака «Стаж работы руководителей», представленного на рисунке 1, необходимо произвести следующий расчет:

$$\bar{x} = \frac{\sum x_i f_i}{n} = \frac{3*3+4*4+6*4+\ldots+20*1+26*1}{43} = 9{,}44$$

Средний стаж работы руководителей равен $\bar{x} = 9{,}44$ лет.

1.7. Показатели вариации признака

Опр. Вариация (изменчивость) – это различия в значениях какого-либо признака у разных единиц данной совокупности относительно среднего.

Вариация тесным образом связана со средним арифметическим. Среднее арифметическое дает обобщающую характеристику признака изучаемой совокупности, но среднее не раскрывает строения совокупности, которое очень существенно для ее анализа.

Среднее не показывает, как располагаются около него варианты значений признака, сосредоточены ли они вблизи среднего значения или значительно отклоняются от него. Среднее арифметическое для двух совокупностей может быть одинаковым, но в первом случае все индивидуальные значения отличаются от нее мало, а во втором – эти отличия велики, это имеет весьма важное значение для характеристики надежности средней величины.

Поэтому и вводятся специальные показатели, которые характеризуют отклонение отдельных значений от общего среднего значения.

К показателям вариации относятся:

- размах вариации (R);
- дисперсия (S^2);
- среднеквадратическое отклонение (S);
- коэффициент вариации (V).

Рассмотрим пример: предположим, что обследуются две группы семей по количеству детей.

1 группа. Число детей в семье: 1 5 3 1 5 $\bar{x} = 3$

2 группа. Число детей в семье: 4 2 3 3 3 $\bar{x} = 3$

В среднем каждая группа характеризуется 3 детьми на одну семью, однако, в первой группе колебания по каждой семье гораздо более существенные, чем во второй группе. Поэтому, анализируя средние значения, необходимо обращать внимание и на отклонения значений вокруг среднего, т.е. на вариацию.

Размах вариации. ***Опр.*** Размах вариации – это разность между максимальным и минимальным значением признака $d = x_{max} - x_{min}$.

Размах вариации показывает лишь крайние отклонения признака и не отражает отклонений всех вариантов в ряду, поэтому широко не используется.

Дисперсия. ***Опр.*** Дисперсия – это средний квадрат отклонений вариантов признака от их среднего арифметического.

В зависимости от того как представлены исходные данные существует три формулы для вычисления дисперсии:

1. Данные представлены в «сыром» виде $S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$

2. Данные представлены в виде вариационного ряда $S^2 = \frac{\sum f_i(x_i - \bar{x})^2}{n-1}$

3. Данные представлены в сгруппированном виде $S^2 \frac{\sum f_i(\tilde{x}_i - \bar{x})^2}{n-1}$

В знаменателе формулы дисперсии вычитают единицу, если объем выборки n<50. Вычитание единицы – это поправка на малый объем выборки.

Недостатком дисперсии как показателя вариации является то, что ее нельзя проинтерпретировать, так как дисперсия не соотносится с измерением изучаемого признака.

Среднеквадратическое отклонение. ***Опр.*** Среднеквадратическое отклонение (СКО) – это корень квадратный из дисперсии. Среднеквадратическое отклонение является абсолютной мерой изменчивости признака, т. к. выражается в тех же единицах, что и значение признака, поэтому может интерпретироваться.

Среднеквадратическое отклонение показывает на сколько в среднем отклоняются индивидуальные значения признака от среднего арифметического.

В зависимости от того как представлены исходные данные существует три формулы для вычисления среднеквадратического отклонения (СКО):

1. Данные представлены в «сыром» виде: $S = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$

2. Данные представлены в виде вариационного ряда: $S = \sqrt{\frac{\sum f_i(x_i - \bar{x})^2}{n-1}}$

3. Данные представлены в сгруппированном виде: $S = \sqrt{\frac{\sum f_i(\tilde{x}_i - \bar{x})^2}{n-1}}$

В знаменателе формулы среднеквадратического отклонения вычитают единицу, если объем выборки n<50. Вычитание единицы – это поправка на малый объем выборки.

Пример. Вернемся к задаче о количестве детей в двух группах семей.

1 группа. Число детей в семье: 1 5 3 1 5 $\bar{x}$ 3

2 группа. Число детей в семье: 4 2 3 3 3 $\bar{x}$ =3

$$S_1 = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{(0\text{-}3)^2 + (6\text{-}3)^2 + (3\text{-}3)^2 + (1\text{-}3)^2 + (5\text{-}3)^2}{5\text{-}1}} = 2{,}6 \text{ детей}$$

$$S_2 = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{(4\text{-}3)^2 + (3\text{-}3)^2 + (2\text{-}3)^2 + (3\text{-}3)^2 + (3\text{-}3)^2}{5\text{-}1}} = 0{,}7 \text{ детей}$$

Так, в первой группе при $\bar{x}_1$= 3 S_1 = 2,6, а во второй $\bar{x}_1$= 3 S_2 = 0,7. Можно сделать вывод, что первая группа имеет сильную вариацию по признаку «количество детей в семье», а вторая группа имеет очень незначительную вариацию по аналогичному признаку даже не имея исходных данных перед глазами.

Коэффициент вариации. ***Опр.*** Коэффициент вариации является относительным показателем вариации и представляет собой выраженное в процентах отношение среднеквадратического отношения к среднему арифметическому $V = \frac{S}{\bar{x}}100\%$.

Коэффициент используется для сравнительной оценки вариации разных совокупностей. Коэффициент вариации также можно использовать как показатель однородности изучаемой совокупности по данному признаку.

Опр. Совокупность считается однородной по данному признаку, если значение коэффициента вариации не превышает 33,3%

Пример. Вернемся к задаче о количестве детей в двух группах семей.

Группа 1. $\bar{x}_1$= 3 детей, $S_1 = 2{,}6$

Группа 2. $\bar{x}_2$= 3 детей, $S_2 = 0{,}7$

Вычислим коэффициент вариации для двух групп семей:

$V_1 = 83{,}3\%$

$V_2 = 23{,}3\%$

Получаем, что коэффициент вариации для первой группы семей равен 83,3%, а для второй – 23,3%. Делаем вывод, что первая группа семей является неоднородной по количеству детей в семье, а вторая группа однородной.

1.8. Анализ формы распределения признака

Характеристиками формы распределения признака являются три показателя:

- Модальность
- Симметричность
- Протяженность

Модальность. По числу и характеру Мод эмпирические распределения делятся на одномодальные и полимодальные.

Опр. Одномодальное распределение – это распределение с одним модальным значением переменной.

Одномодальные распределения могут быть колоколообразными, в этом случае $X_{min} < M_o < X_{max}$ и j-образными, тогда $X_{min} = M_o$ или $M_o = X_{max}$.

Опр. Полимодальное распределение – это распределение с несколькими модальным значениями переменной.

Чаще всего в качестве полимодальных распределений рассматривают бимодальные распределения, характеризующиеся наличием двух ярко выраженных Мод.

Наличие полимодальности и в частности бимодальности говорит о том, что изучаемое распределение является неоднородным.

Симметричность. Распределения бывают симметричные и асимметричные.

Опр. Симметричное распределение – это одномодальное колоколообразное распределение, для которого справедливо соотношение $M_e = M_o = \bar{x}$.

Опр. Асимметричное распределение – это такое распределение, для которого фиксируется отклонение от симметрии. Ассиметричные распределения делятся на левосторонние и правосторонние.

Опр. Левосторонняя (отрицательная) асимметрия характеризуется наличием длинного левого хвоста распределения и соотношением мер центра $M_e < \bar{x} < M_o$ или $\bar{x} < M_e < M_o$.

Опр. Правосторонняя (положительная) асимметрия характеризуется наличием длинного правого хвоста распределения и соотношением мер центра $M_o < M_e < \bar{x} <$ или $M_o < \bar{x} < M_e$.

Если в эмпирическом распределении фиксируется сильная правосторонняя или левосторонняя асимметрия – это значит, что в качестве характеристики центра распределения вместо среднего арифметического лучше использовать медиану.

Протяженность. Протяженность эмпирического распределения связана со степенью разброса (вариации) данных. Чем больше дисперсия и соответственно среднеквадратическое отклонение, тем более протяженным будет колокол распределения.

Чем больше дисперсия, тем форма колокола становится более пологой. В этом случае форма распределения носит название платокритическое распределение. Чем меньше дисперсия, тем форма колокола становится более выпуклой. В этом случае форма распределения носит название лептокритическое распределение.

1.9. Стандартизация количественных переменных. Z-оценки

Работая с наборами различных количественных переменных, мы имеем дело с различными масштабами измерения (рубли, доллары, сантиметры, часы, минуты, годы, штуки и т.д.), пока мы изучаем одну переменную масштаб для нас не имеет большого значения.

Когда же нам необходимо работать со множеством различных переменных, сопоставлять их воздействие на другие переменные, мы сталкиваемся с проблемой их несоразмерности из-за различного масштаба измерения. Чтобы нивелировать влияние масштаба измерения прибегают к процедуре стандартизации.

При переводе в стандартный масштаб каждое индивидуальное значение признака должно быть переведено в Z-оценку.

Опр. Процедура перевода исходных значений переменной в Z-оценку называется стандартизацией количественной переменной.

Формула перевода исходного значения переменной в Z-оценку:

$$Z_i = \frac{xi - \bar{x}}{s}, \text{ где}$$

x_i – исходное значение переменной;

$\bar{x}$ – среднее арифметическое;

S – среднеквадратическое отклонение (СКО)

Свойства Z-оценок:

1. Среднее арифметическое Z-оценок всегда равно нулю.
2. Среднеквадратическое отклонение Z-оценок всегда равно 1.
3. Знак Z-оценки показывает, является ли значение признака большим или меньшим по сравнению со средним арифметическим.
4. Значение Z-оценки, превышающее по модулю величину 3, свидетельствует о значительном отклонении индивидуального значения от среднего. Такие сильные отклонения в статистике называются выбросами и часто рассматриваются как нетипичные значения.

Если в распределении появляется выброс, то с ним поступают следующим образом:

- Выбросы исключают из распределения признака при расчете среднего арифметического, т.к. они сильно искажают его реальное значение;
- Игнорируют появление выброса и оставляют его в распределении, подчеркивая, что появление таких значений закономерно. Однако если выбросы остаются, то в качестве меры центральной тенденции рекомендуется использовать не среднее арифметическое, а медиану.

Замечание: Переход от исходной переменной к Z-оценке изменяет масштаб измерения изучаемой количественной переменной, но не меняет его форму распределения.

1.10. Теоретические распределения и их статистические таблицы

Теоретическое нормальное распределение Гаусса. ***Опр.*** Нормальное распределение Гаусса – это распределение, при котором переменная величина изменяется непрерывно, крайние значения появляются редко, но чем ближе значения признака к центру (к средней арифметической), тем они чаще встречаются.

Графически нормальное распределение представляет собой симметричный колокол, центр которого расположен в точке μ (Мю), а ширина зависит от значения σ (Сигмы).

Нормальное распределение обозначается: $N(\mu; \sigma)$**,**

где μ и σ являются параметрами нормального распределения.

μ – это математическое ожидание (или среднее арифметическое для генеральной совокупности)

σ – среднеквадратическое отклонение (среднеквадратическое отклонение для генеральной совокупности).

В зависимости от того, какие значения принимает μ и σ существует бесконечное множество нормальных распределений (Рисунок 9). Чем больше дисперсия и соответственно σ, тем более плоский вид принимает нормальная кривая и наоборот.

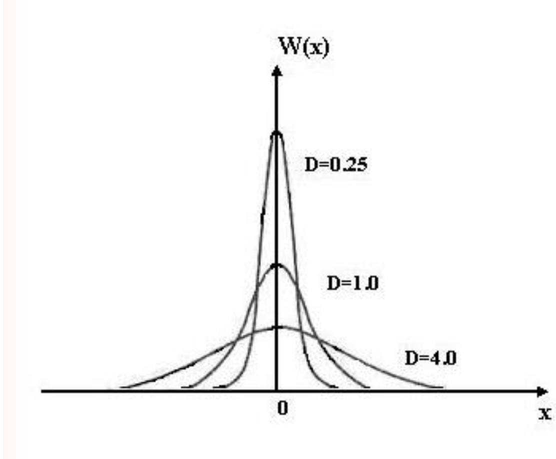


Рисунок 9 – Вид нормальной кривой в зависимости от значения дисперсии

Свойства кривой нормального распределения:

1. Кривая симметрична относительно максимального значения по оси X. Данная точка соответствует значению μ, причем, для нормального распределения характерно $\mu = M_o = M_e$
2. Кривая асимптотически приближается к оси абсцисс, продолжаясь в обе стороны до бесконечности. Следовательно, чем больше значения отклоняются от μ, тем реже они встречаются.
3. Кривая имеет две точки перегиба, находящиеся на расстоянии ± σ от μ.
4. При μ=const, с увеличением σ кривая становится более пологой. При σ=const с изменением μ кривая не меняет своей формы, а лишь сдвигается вправо или влево по оси X.
5. Правило 3σ (трех сигм) (Рис. 10).

В интевале **μ±σ** находится **68,2%** всех наблюдаемых значений.

В интевале **μ±2σ** находится **95,4**% всех наблюдаемых значений.

В интевале **μ±3σ** находится **99,7%** всех наблюдаемых значений.

Так как хвосты распределения асимптотически приближается к оси X, они никогда не пересекутся с осью. Правило трех сигм позволяет ограничить хвосты распределения справа и слева.

Рис. 10 –. Иллюстрация правила трех сигм

Стандартизация нормального распределения. Для того чтобы унифицировать множество нормальных распределений, определяемых значениями μ и σ прибегают к процедуре стандартизации. Переход к стандартному нормальному распределению происходит по формуле Z-оценки. Для теоретического распределения формула Z-оценки будет иметь вид:

$$Z = \frac{x - \mu}{\sigma}, \text{ где}$$

x – значение случайной величины (квантиль);

μ – математическое ожидание;

σ – среднеквадратическое отклонение (СКО).

В соответствии со свойствами Z-оценки, среднее арифметическое Z-оценки всегда равно 0, а среднеквадратическое Z-оценки всегда равно 1. Тогда параметрами стандартного нормального распределения будут μ=0 и σ=1.

Стандартное нормальное распределение обозначается Z(0;1), т.е. параметрами стандартного нормального распределения являются 0 и 1.

Формула Z-оценки позволяет каждой точке произвольного нормального распределения поставить в соответствие точку стандартного нормального распределения, что позволяет рассчитать одну статистическую таблицу для стандартного распределения с возможностью обратного перевода стандартного значения в произвольное.

По таблице стандартного нормального распределения можно решать задачи двух типов:

Первый тип задач. Для заданного числа (квантиля) «а» определить значение функции распределения F(a) и далее использовать это значение для определения любой заданной вероятности.

P(x<a)=F(a) – вероятность попадания случайной величины в область слева от квантиля «а» (Рисунок 11)

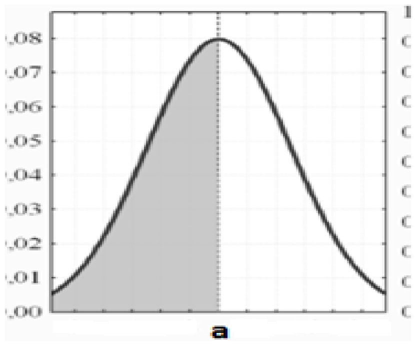


Рисунок 11 – Площадь под нормальной кривой слева от квантиля «а»

P(x>a)=1-F(a) – вероятность попадания случайной величины в область справа от квантиля «а» (Рисунок 12).

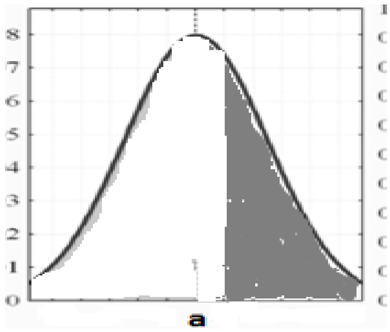


Рисунок 12 – Площадь под нормальной кривой справа от квантиля «**a**»

P(a<x<b)=F(b)-F(a) – вероятность попадания случайной величины в область, ограниченную двумя квантилями «a» и «b» (Рисунок 13).

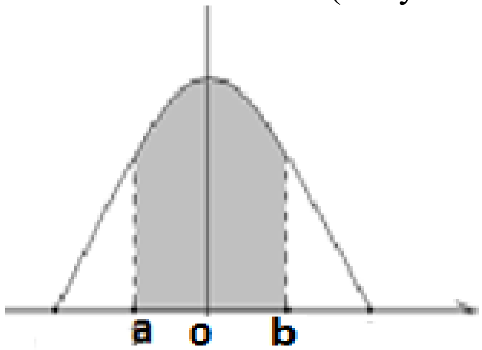


Рисунок 13 – Площадь под нормальной кривой, ограниченная двумя квантилями «a» и «b»

Второй тип задач. Для заданного значения вероятности определить соответствующий ей квантиль «a»: p=F(a). Для решения данной задачи рассматриваем вариант попадания случайной величины только слева от квантиля.

При решении задач с использованием статистической таблицы стандартного нормального распределения используется свойство симметрии нормального распределения, которое заключается в следующем:

Свойство симметрии. Вероятность попадания в область под нормальной кривой слева от квантиля «a» равна вероятности попадания в область под нормальной кривой справа от квантиля «a» (Рисунок 14) P(x<-a)=P(x>a).

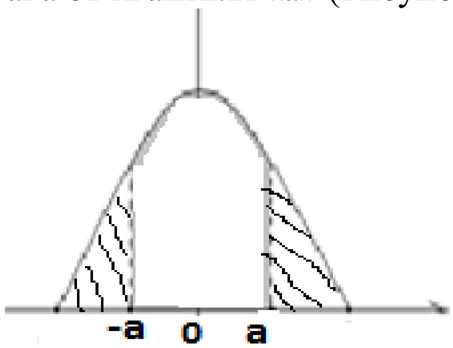


Рисунок 14 – Иллюстрация свойства симметрии нормального распределения. Площадь слева от квантиля «-a» равна площади справа от квантиля «a»

На Рисунке 15 приводится фрагмент статистической таблицы стандартного нормального распределения для положительных значений квантиля «а». Внутри таблицы располагаются вероятности, в верхней строке и крайнем левом столбце значения положительных квантилей.

z	·00	·01	·02	·03	·04	·05	·06	·07	·08	·09
·0	·5000	·5040	·5080	·5120	·5160	·5199	·5239	·5279	·5319	·5359
·1	·5398	·5438	·5478	·5517	·5557	·5596	·5636	·5675	·5714	·5753
·2	·5793	·5832	·5871	·5910	·5948	·5987	·6026	·6064	·6103	·6141
·3	·6179	·6217	·6255	·6293	·6331	·6368	·6406	·6443	·6480	·6517
·4	·6554	·6591	·6628	·6664	·6700	·6736	·6772	·6808	·6844	·6879
·5	·6915	·6950	·6985	·7019	·7054	·7088	·7123	·7157	·7190	·7224
·6	·7257	·7291	·7324	·7357	·7389	·7422	·7454	·7486	·7517	·7549
·7	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	·7764	·7794	·7823	·7852

Рисунок 15 – фрагмент статистической таблицы нормального распределения

1.11. Статистический вывод. Оценка параметров генеральной совокупности

Опр. Выборочное исследование – это такое исследование, при котором отбор единиц исследования осуществляется в случайном порядке, отобранная часть изучается, а результаты распространяются на всю генеральную совокупность.

Основная задача выборочного исследования состоит в том, чтобы на основе характеристик выборочной совокупности получить достоверные суждения о показателях генеральной совокупности.

Опр. Статистический вывод – это некоторое утверждение об изучаемой генеральной совокупности на основании результатов изучения выборочной совокупности.

Математическая статистика рассматривает не любые утверждения о генеральной совокупности, а лишь касающиеся числовых характеристик, таких как средние, меры вариации, доли признака.

Числовые характеристики, описывающие генеральную совокупность, называются параметрами. Те же характеристики, но рассчитанные по выборке, называются статистиками (Таблица 8).

Определение статистического вывода также можно сформулировать, используя параметры и статистики.

Опр. Статистический вывод – это некоторое утверждение об изучаемых параметрах генеральной совокупности на основании изучения выборочных статистик.

Таблица 8 – Параметры генеральной совокупности и соответствующие им выборочные статистики

Параметры генеральной совокупности	Выборочные статистики
N – объем генеральной совокупности	N – объем выборочной совокупности
μ – математическое ожидание (генеральное среднее)	$\overline{x}$ – среднее арифметическое, полученное по выборке
σ^2 – дисперсия генеральной совокупности	S^2 – дисперсия, полученная по выборке
σ – среднеквадратическое отклонение ГС	S – среднеквадратическое отклонение по выборке
p – доля признака в ГС	p – доля признака в выборке

Опр. Выборочное распределение некоторой статистики представляет собой теоретическое частотное распределение этой статистики, которое могло бы быть получено в результате осуществления очень большого (практически бесконечного) числа выборок.

Связь между параметром генеральной совокупности и выборочным распределением статистики можно построить теоретически, основываясь на следствии из центральной предельной теоремы.

Следствие из Центральной предельной теоремы. Если из бесконечной генеральной совокупности методом простого случайного отбора извлекается бесконечное число выборок одного и того же объема n, то выборочные средние имеют нормальное распределение с параметрами μ и $\sigma/\sqrt{n}$, где

μ – это математическое ожидание;

σ – среднеквадратическое отклонение.

Утверждения о параметрах генеральной совокупности на основании изучения выборочных статистик носят вероятностный характер и подразделяются на три вида. Таким образом, рассматривают три вида статистического оценивания:

1. Точечное оценивание;
2. Интервальное оценивание;
3. Проверка статистических гипотез.

Опр. Оценка – это любое число, рассчитанное по выборке и характеризующее неизвестный параметр.

Точечное оценивание. ***Опр.*** Точечное оценивание состоит в том, что параметр генеральной совокупности приравнивается к значению выборочной статистики $\mu = \overline{x}$.

Свойства точечного оценивания:

1. Несмещенность – свойство, состоящее в том, что среднее выборочного распределения оценки равно величине параметра.

Среднее арифметическое является несмещенной оценкой, а дисперсия смещенной оценкой.

Несмещенность нарушается при совершении систематической ошибки отбора элементов генеральной совокупности в выборку.

2. Состоятельность – это свойство оценки при увеличении объема выборки приближаться к значению оцениваемого параметра. Дисперсия является смещенной, но состоятельной оценкой.

3. Эффективность – чем ниже дисперсия, тем меньше отличаются оценки, полученные в разных выборках, тем выше эффективность оценки.

Основное преимущество точечного оценивания – это его простота $\mu=\bar{x}$. (параметр приравнивается к вычисленной статистике).

Основной недостаток точечного оценивания – невозможность оценить вероятность случайной ошибки и, соответственно, ее контролировать.

*Интервальное оценивание***.** Интервальное оценивание представляет собой построение некоторого интервала, в который параметр генеральной совокупности попадает с некоторой вероятностью $P\ (a<\mu<b)=0{,}9$.

Опр. Вероятность, с которой параметр попадает в интервал, называется доверительной вероятностью, а сам интервал называется доверительным интервалом.

Доверительная вероятность (обозначается – β) выбирается исследователем самостоятельно и может принимать значения – 0,9; 0,95; 0,99; 0,999.

Вероятность случайной ошибки при интервальном оценивании обозначается α, и связана с доверительной вероятностью по формуле $\beta=1-\alpha$.

$0{,}9=1-\alpha$, следовательно $\alpha=0{,}1$

$0{,}95=1-\alpha$, следовательно $\alpha=0{,}05$

$0{,}99=1-\alpha$, следовательно $\alpha=0{,}01$

$0{,}999=1-\alpha$, следовательно $\alpha=0{,}001$

При построении доверительного интервала используют следствие из центральной предельной теоремы, а именно, что выборочные средние имеют нормальное распределение с параметрами μ и $\sigma/\sqrt{n}$

Формула доверительного интервала для количественного признака будет иметь вид: $\bar{x}\pm Z_{1-\alpha/2}\ \frac{S}{\sqrt{n}}$, где

$Z_{1-\alpha/2}$ – значение квантиля, зависящее от вероятности совершения случайной ошибки отбора и полученное по таблице стандартного нормального распределения;

$\bar{x}$ – среднее арифметическое, вычисленное по выборке;

S – среднеквадратическое отклонение, вычисленное по выборке%

n – объем выборки.

Формула доверительного интервала для качественного признака будет иметь вид: $p \pm Z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$, где

p – доля положительных значений, полученная по выборке;

$Z_{1-\alpha/2}$ – значение квантиля, зависящее от вероятности совершения случайной ошибки отбора и полученное по таблице стандартного нормального распределения;

n – объем выборки.

1.12. Простая случайная выборка из генеральной совокупности. Ошибка простой случайной репрезентативной выборки и ее объем

Количественные методы сбора данных предполагают изучение выборочной совокупности. Правильный расчет объема и выбор оптимальной процедуры построения выборочной совокупности являются одной из главных задач для получения достоверных результатов при реализации количественных методов.

Если рассматривать выборку как средство изучения генеральной совокупности, то главным требованием к ней является возможность обобщения результатов выборочного исследования на генеральную совокупность с высокой степенью уверенности в том, что эти выводы будут справедливы.

Результаты выборочных исследований всегда отчасти неопределенны. Это происходит потому, что изучается только часть генеральной совокупности, и при реализации процедуры измерения могут допускаться ошибки. Однако при отсутствии грубых просчетов в планировании и реализации выборки эти ошибки можно контролировать, то есть с высокой вероятностью «удерживать» в некоторых допустимых пределах.

Опр. Ошибкой выборки для некоторого признака называется разность между средним арифметическим значением, полученным для данного признака по выборке, и средним арифметическим, соответствующим генеральной совокупности $\Delta=\mu-\bar{x}$.

Так как генеральное среднее как правило неизвестно, ошибку выборки точно рассчитать нельзя, но можно контролировать статистически.

При отборе элементов генеральной совокупности в выборку можно совершить два вида ошибок отбора: систематическую и случайную.

Опр. Систематические ошибки направлены в одну сторону, вследствие чего ошибка выборки не уменьшается при увеличении объема выборочной совокупности. Систематические ошибки обычно связывают с ошибками проектирования выборки.

Опр. Случайные ошибки (непреднамеренные) обычно уравновешивают друг друга, поскольку не имеют преимущественного направления в сторону преувеличения или преуменьшения значения изучаемого признака. Ошибки такого рода неустранимы, но они подчиняются статистическим законам и, соответственно, поддаются контролю. Важнейшее свойство случайных ошибок состоит в том, что они уменьшаются с увеличением объема выборки.

Как правило, при проведении исследования среднее значение признака для генеральной совокупности неизвестно. Напротив, исследование проводится с целью его оценить. Поэтому ошибка выборки не может быть вычислена точно, а только может быть оценена с помощью статистических процедур.

При отсутствии систематических ошибок степень точности для каждого признака (репрезентативность выборки) задается двумя показателями: величиной предельно допустимой случайной ошибки и вероятностью того, что эта ошибка не будет превышена.

Опр. Выборка называется репрезентативной, если мы с определенной вероятностью (доверительной вероятностью) можем контролировать ее ошибку (случайную ошибку).

Например, когда нам сообщают, что продукт X предпочитает покупать 15% опрошенных респондентов и ошибка выборки не превышает 3% с вероятностью 95%, это означает, что с вероятностью 95% продукт X предпочитают покупать от 12% до 18% потребителей, входящих в генеральную совокупность (15%±3%). При этом вероятность того, что ошибка выборки выйдет за пределы ±3% составляет 5%. (т. е. вероятность того, что продукт X выберет менее 12% или более 18% генеральной совокупности, составляет 5%).

Величина предельно допустимой ошибки существенным образом зависит от объема выборки и способа ее извлечения. Стремление повысить точность приводит к быстрому росту необходимого объема выборки и, соответственно, стоимости исследования.

Формула расчета объема выборки для доли признака с учетом поправки на конечность генеральной совокупности.

$$n = \frac{t^2\, p(1-p)N}{\Delta^2 N + t^2\, p(1-p)}, \text{ где}$$

t - коэффициент, соответствующий доверительной вероятности. Так, если доверительная вероятность 95%, то значение t = 1,96, для доверительной вероятности 99% значение коэффициента t = 2,59.

p - доля признака для генеральной совокупности. Как правило, неизвестна, поэтому при расчете используют максимальное значение, которое достигается при p= 0,5, тогда 0,5(1-0,5)=0,25

Δ^2 – величина допустимой ошибки в долях. Если мы устанавливаем допустимую ошибку равную 3%, то допустимая ошибка в долях будет равна 0,03.

N – объем генеральной совокупности.

В Таблице 9 приводится зависимость объема выборки от объема генеральной совокупности при допустимой ошибке 5% и доверительной вероятности 95%.

Таблица 9 – Соотношение объема выборочной совокупности к объему генеральной совокупности для доверительной вероятности 95% и ошибки выборки 5%

Объем ГС	500	1000	3000	5000	10000	100000	Бесконечная
Объем выборки	222	286	350	370	385	398	400

Из таблицы видно, что для обеспечения заданной репрезентативности при исследовании города с населением 100 тыс. жителей надо опросить 398 человек, а при исследовании всей страны практически столько же — 400 человек.

Начиная с некоторого момента, увеличение объема генеральной совокупности не оказывает существенного влияния на увеличение объема выборки, поэтому при больших генеральных совокупностях, начиная с N>5000, поправкой на объем генеральной совокупности можно пренебречь, тогда формула будет иметь вид:

$$n = \frac{t^2\, p(1-p)}{\Delta^2},\ \text{где}$$

t - коэффициент, соответствующий доверительной вероятности. Так, если доверительная вероятность 95%, то значение t = 1,96, для доверительной вероятности 99% значение коэффициента t = 2,59.

р - доля признака для генеральной совокупности. Как правило, неизвестна, поэтому при расчете используют максимальное значение, которое достигается при р= 0,5, тогда 0,5(1-0,5)=0,25

Δ^2 – величина допустимой ошибки в долях. Если мы устанавливаем допустимую ошибку равную 3%, то допустимая ошибка в долях будет равна 0,03.

N – объем генеральной совокупности.

При планировании объема выборки следует иметь в виду следующее. Рассчитанный объем выборки позволяют получить заданную точность при анализе выборки в целом, т. е. если мы не будем расчленять ее на части. Если, например, требуется определить долю потребителей, предпочитающих продукт X, то, опросив 400 отобранных человек, мы с вероятностью 95% определим искомую с ошибкой, не превышающей 5%. Но если мы хотим определить эту долю не для всего массива, а для женщин и мужчин, нам необходимо, чтобы в выборке было 400 женщин и 400 мужчин, т.е. 800 человек. Чем больше будет дробиться массив при анализе информации, тем больший объем выборки понадобится.

Процедуры построения выборочной совокупности. Возможность обобщения результатов выборочного исследования на генеральную совокупность базируется на теоретических выводах математической статистики, которая основывается на приложениях теории вероятностей. Основанием применения этих методов служит предположение, что все элементы выборочной совокупности были отобраны случайным образом. При случайном отборе все элементы генеральной совокупности имеют одинаковую вероятность быть отобранными в выборку. Для больших и неоднородных генеральных совокупностей обеспечить выполнение этого условия достаточно сложно.

Простой случайный отбор предполагает, что имеется полный список элементов генеральной совокупности. Респондентов отбирают из этого списка одним из методов случайного отбора:

- генерация случайных чисел;
- лотерея,
- систематический отбор с определенным шагом.

Если ГС слишком велика и неудобна для обследования целиком, применяют расслоенный случайный отбор, который предполагает предварительное разделение генеральной совокупности на непересекающиеся части с последующим извлечением из каждой части простой случайной выборки.

При расслоенном отборе используются две стратегии размещения выборки по слоям. Равномерное размещение предполагает, что в каждом слое обследуется одинаковое количество объектов. Используется в сравнительных исследованиях. Пропорциональное размещение означает, что объем выборки в каждом слое пропорционален объему слоя в генеральной совокупности. Используется в описательных исследованиях.

В условиях неопределенной генеральной совокупности, списки которой невозможно получить, используется кластерный отбор. При применении кластерного отбора ГС подразделяется на непересекающиеся подсовокупности по некоторому объективному и мало зависящему от наблюдателей основанию. В качестве кластеров могут использоваться административные единицы (области, районы, населенные пункты, кварталы, улицы, отдельные дома и квартиры), предприятия, академические группы.

Многоступенчатый кластерный отбор. В большинстве национальных исследований используются многоступенчатый кластерный отбор. На первой ступени осуществляется кластеризация по областям, на второй по населенным пунктам пропорционально размеру, на третьей ступени в разных населенных пунктах могут применяться различные стратегии отбора респондентов, например, маршрутная выборка, либо дальнейшая кластеризация населения больших городов (с использованием административных районов города, жилищно-эксплуатационных управлений, почтовых отделений).

Маршрутная выборка, наряду с кластерным отбором, представляет собой еще один способ формирования случайной выборки в условиях неопределенной генеральной совокупности. Метод применяется в отдельных населенных пунктах. Единицей отбора при осуществлении маршрутной выборки является жилое помещение, семья или домохозяйство.

Метод заключается в том, что интервьюер следует в населенном пункте предписанному маршруту, отбирая жилые помещения по заданной схеме. Например, интервьюеру дано задание отбирать дома с нечетными номерами, номера квартир с шагом 10, т.е. 1, 10, 100, 200 и т д.. При отборе улиц можно руководствоваться одним из следующих принципов:

- выбирать в каждой зоне типические улицы, с характерными типами застройки;
- составить полный список улиц и извлечь из него случайную выборку.
- Выбирать улицы случайным образом, например, улицы, в названии которых одновременно встречаются две заранее заданные буквы.

После того как жилое помещение выбрано, необходимо определить, кого из членов семьи следует проинтервьюировать. Как правило, это тот, у кого ближайший день рождения.

1.13. Понятие статистической гипотезы. Процедура проверки гипотезы

Во многих случаях исследование проводится с целью проверки некоторых априорных гипотез относительно параметров генеральной совокупности (ГС).

Гипотезы могут касаться самых различных сфер общественной жизни – различий в оплате труда женщин и мужчин, инфляционных ожиданий населения, особенностей оплаты труда различных категорий занятых, изменения ценностных ориентаций потребителей и т.п.

Процедуры проверки гипотез применяются также тогда, когда необходимо убедиться в достоверности различий между ответами различных групп респондентов, наличии и характере связи между двумя или более переменными и при решении многих других прикладных задач.

Статистической гипотезой называется утверждение относительно параметров генеральной совокупности, подлежащее проверке с помощью выборочного исследования и сформулированное по определенным правилам. Любая содержательная гипотеза относительно параметров ГС может быть сформулирована в виде статистической гипотезы.

Статистическая гипотеза состоит из двух утверждений. Первое утверждение, постулирующее отсутствие различий между параметрами ГС или связи между переменными, называется *нулевой гипотезой* и обозначается H_0. Второе утверждение называется *альтернативной гипотезой* и обозначается H_1 формулирует наличие различий или наличие связи с учетом их характера.

Статистическая гипотеза формулируется в терминах параметров генеральной совокупности с использованием отношений равенства (H_0) и неравенства – не равно, больше, меньше (H_1).

Например, содержательная гипотеза о том, что зарплата женщин, в среднем, ниже, чем зарплата мужчин, может быть сформулирована следующим образом.

$H_0 : \mu_{ж} = \mu_{м}$ – средняя зарплата женщин и мужчин одинакова

$H_1 : \mu_{ж} < \mu_{м}$ – средняя зарплата женщин ниже, чем средняя зарплата мужчин

Гипотеза о том, что на должность менеджера наниматели предпочитают брать мужчин, можно сформулировать так:

$H_0 : p_{м} = 0{,}5$ – среди менеджеров мужчины составляют 50%

$H_1 : p_{м} > 0{,}5$ – среди менеджеров мужчины составляют более 50%

Альтернативные гипотезы, в соответствии с используемым знаком неравенства, делят на *ненаправленные или двусторонние* ($\neq$) и на *направленные или односторонние* (<; >). Направленные альтернативные гипотезы, в свою очередь, можно разделить на *левосторонние* (<) и *правосторонние* (>).

Например, гипотеза о средней зарплате женщин и мужчин является левосторонней, а гипотеза о преобладании мужчин среди менеджеров – правосторонней.

Проверка статистической гипотезы. Проверка статистической гипотезы состоит в том, чтобы по данным выборочного исследования сделать вывод, следует ли в отношении ГС принять нулевую гипотезу или отклонить ее в пользу альтернативной. При этом нулевая гипотеза считается справедливой до тех пор, пока не будет найдено убедительное подтверждение того, что она не верна.

Решение о принятии или отклонении нулевой гипотезы принимается в соответствии с *критерием*, который строится на основе специально подобранной для каждой нулевой гипотезы численной функции. Функция вычисляется по выборке и называется *статистикой критерия*.

Поскольку любое решение (принять или отклонить H_0) принимается на основе выборки, оно может быть для правильным, так и ошибочным генеральной совокупности. При этом возможны два типа ошибок.

Ошибка, заключающаяся в том, чтобы по данным выборки отклонить нулевую гипотезу, которая на самом деле верна, называется *ошибкой первого рода*; ее вероятность обозначается буквой α.

Ошибка, состоящая в том, чтобы принять нулевую гипотезу, которая на самом деле не верна, называется *ошибкой второго рода*; ее вероятность обозначается буквой β.

Статистическая задача состоит в том, чтобы найти решающую процедуру (критерий гипотезы), минимизирующий вероятность совершения любой из этих ошибок. Проблема заключается в том, что с уменьшением вероятности ошибки I рода увеличивается вероятность ошибки II рода и наоборот. Эту проблему учитывают при разработке критериев, благодаря чему исследователь, не являющийся профессиональным статистиком, может на практике руководствоваться простым правилом:

При проверке статистической гипотезы фиксируют некоторое малое значение α стандартно равное 0,1; 0,05; 0,01 и предполагают, что β также будет мало. Фиксированное значение α называется уровнем значимости критерия.

Процедура проверки статистической гипотезы:

1. На основании данных, полученных по выборке, формулируем статистическую гипотезу.
2. Вычисляем статистику критерия, используя соответствующий гипотезе статистический критерий.
3. В соответствии с выбранным уровнем значимости, определяем в теоретическом распределении критическую область или области.
4. Если статистика критерия попадает в центр распределения, т. е. в область принятия нулевой гипотезы, принимается H_0, если статистика критерия попадает в критическую область, соответствующую альтернативной гипотезе, то нулевая гипотеза отвергается и принимается альтернативная H_1.

Трем вида альтернативной гипотезы соответствуют три вида критических областей.

Если H_1 сформулирована со знаком больше, то критической областью является правый хвост распределения ограниченный квантилем $Z_{1-\alpha}$

Если H_1 сформулирована со знаком меньше, то критической областью является левый хвост распределения ограниченный квантилем Z_{α}

Если H_1 сформулирована со знаком неравенства, то критическими областями будут два хвоста распределения ограниченные слева квантилем $Z_{\alpha/2}$, а справа квантилем $Z_{1-\alpha/2}$.

1.14. Виды статистических гипотез: гипотезы о долях, гипотезы о средних, гипотезы о дисперсиях

В данной теме будут рассмотрены гипотезы о равенстве параметров генеральной совокупности в зависимости от измерения исходных переменных.

Гипотеза о долях для дихотомического признака. Рассмотрим гипотезу о равенстве доли положительных значений некоторой константе. В данном случае проверяется нулевая гипотеза о равенстве доли положительных значений дихотомической переменной некоторому числу (константе). Значение константы определяем по предыдущим исследованиям или по официальной статистике.

В основу формулы критерия положено теоретическое нормальное распределение. Чтобы проверить гипотезу сравниваем значение статистики критерия со значением критической точки, полученной по статистической таблице стандартного нормального распределения основываясь на величине значимости критерия, назначенного исследователем. Если статистика критерия по модулю больше критической точки, подтверждается альтернативная гипотеза.

Рассмотрим гипотезу о равенстве доли положительных значений в двух совокупностях. В данном случае проверяется нулевая гипотеза о равенстве доли положительных значений дихотомической переменной двух совокупностей, например, совокупности мужчин и совокупности женщин. Знак альтернативной гипотезы определяем по значениям, полученным в выборке.

В основу формулы критерия положено теоретическое нормальное распределение. Чтобы проверить гипотезу сравниваем значение статистики критерия со значением критической точки, полученной по статистической таблице стандартного нормального распределения основываясь на величине значимости критерия, назначенного исследователем. Если статистика критерия по модулю больше критической точки, подтверждается альтернативная гипотеза.

*Гипотеза о средних***.** Рассмотрим гипотезу о равенстве среднего некоторой константе. В данном случае проверяется нулевая гипотеза о равенстве среднего некоторому числу (константе). Значение константы определяем по предыдущим исследованиям или по официальной статистике.

В основу формулы критерия положено теоретическое распределение t-стьюдента. Чтобы проверить гипотезу сравниваем значение статистики критерия со значением критической точки, полученной по статистической таблице распределения t-стьюдента основываясь на величине значимости критерия, назначенного исследователем и параметре «число степеней свободы», который определяется по формуле $df=n-1$ (n – объем выборки). Если статистика критерия по модулю больше критической точки, подтверждается альтернативная гипотеза.

Рассмотрим гипотезу о равенстве средних в двух совокупностях, например, совокупность 1 – поколение Z и совокупность 2 – поколение Y. В данном случае проверяется нулевая гипотеза о равенстве среднего в двух изучаемых совокупностях. Знак альтернативной гипотезы определяем по значениям тестируемой переменной, полученной в выборке.

В основу формулы критерия положено теоретическое распределение t-стьюдента. Чтобы проверить гипотезу сравниваем значение статистики критерия со значением критической точки, полученной по статистической таблице распределения t-стьюдента основываясь на величине значимости критерия, назначенного исследователем и параметре «число степеней свободы», который определяется в зависимости от того равны дисперсии тестируемой переменной или не равны между собой. Равенство дисперсий проверяется с помощью гипотезы о равенстве дисперсий двух независимых выборок. Если статистика критерия по модулю больше критической точки, подтверждается альтернативная гипотеза.

Гипотеза о дисперсиях. Проверяется нулевая гипотеза о равенстве дисперсий двух изучаемых независимых совокупностей, например, совокупность 1 – поколение Z и совокупность 2 – поколение Y. В гипотезе о дисперсиях всегда знаком альтернативной гипотезы будет неравенство.

В основу формулы критерия положено теоретическое распределение F-Фишера. Чтобы проверить гипотезу сравниваем значение статистики критерия со значением критической точки, полученной по статистической таблице распределения F-Фишера основываясь на величине значимости критерия, назначенного исследователем и параметре «число степеней свободы», который рассчитывается для двух совокупностей по формулам $df_1=n_1-1$ (n_1 – объем выборки первой совокупности) и $df_2=n_2-1$ (n_2 – объем выборки второй совокупности). Если статистика критерия больше критической точки, подтверждается альтернативная гипотеза.

1.15. Статистический анализ взаимосвязи признаков

Анализ поведения изучаемых признаков относительно друг друга необходим для поиска ответов на вопросы: существует ли связь между признаками; влияет ли один признак на другой; можно ли, зная значение одного из них, сделать вывод относительно распределения значения другого.

Опр. Статистической связью двух признаков называется такое соотношение между ними, при котором изменение значения одного признака меняет распределение другого.

Опр. Статистической независимостью двух признаков называется такое соотношение между ними, при котором изменение значения одного признака не приводит к изменению распределения другого.

Существуют десятки моделей статистической связи. Многообразие моделей объясняется тем, что существует множество подходов к определению того, как именно меняется распределение признака при изменении значения другого признака.

Основные классы моделей статистической связи:

1. Частотные модели (анализ таблиц сопряженности) – используются для анализа связи двух качественных признаков.
2. Корреляционные модели (корреляционный анализ) – используются для анализа связи количественных признаков, либо качественных признаков, измеренных с помощью порядковых шкал.
3. Функциональные модели (регрессионный анализ) – используются для анализа причинно-следственной связи количественных признаков.
4. Модели с определением дисперсии (дисперсионный анализ) – используются для анализа причинно-следственной связи качественного и количественного признаков.

Выбор той или иной математической модели статистической связи определяется не только способом измерения признаков, но и характером исследуемой связи.

Статистические связи можно классифицировать следующим образом.

1. По направлению:

а) ненаправленные;

б) прямые (положительные);

в) обратные (отрицательные).

Опр. Ненаправленными называются связи, при которых нельзя сказать, что увеличение значения одного признака приводит к увеличению или уменьшению значения другого.

Опр. Прямыми или положительными связями называются такие, при которых увеличение значения одного признака приводит к увеличению значения другого, и наоборот, уменьшение значения одного приводит к уменьшению значения другого.

Опр. Обратной или отрицательной связью называется такая связь, при которой увеличение значения одного признака ведет к уменьшению значения другого, и наоборот.

2. По наличию причинности связи подразделяются на:

а) корреляционные;

б) причинно-следственные.

Опр. Корреляционными называются такие связи, в которых признаки «равноправны», т. е. нельзя сказать, какая из них является причиной, а какая следствием.

Опр. Причинной называется такая связь, при которой один признак может быть описан как причина, а другой как следствие.

3. По измерению связи подразделяются на:

а) функциональные;

б) статистические.

Опр. Функциональной называется связь, которая может быть описана математической формулой $y=f(x)$. Такие связи встречаются только в точных науках.

Опр. Связь называется статистической, если ее можно выразить функционально, но только до некоторого приближения $y=f(x)+\varepsilon$.

4. По форме связи подразделяются на:

а) линейные;

б) нелинейные.

Опр. Статистическая связь является линейной, если ее можно выразить с помощью линейной функции $y=b_1x+b_0$.

Опр. Статистическая связь является нелинейной, если ее можно выразить с помощью нелинейной функции $y= b_1\log(x)+ b_0$, $y= b_1x^2+ b_0$ и т. д.

Понятие меры связи. Для измерения силы связи разработаны специальные коэффициенты, называемые мерами связи. Коэффициентов существует столько же, сколько и моделей связи, т. е. для каждой модели связи существует свой коэффициент

Общие свойства мер связи:

1. Значения коэффициентов изменяются в интервале [0;1] для ненаправленных связей, и [-1;1] – для направленных.
2. Значения коэффициентов, равное нулю, может свидетельствовать:
   а) об отсутствии связи между признаками;
   б) о том, что выбранная модель не соответствует характеру связи.
3. Значение коэффициента связи близкое к 1 свидетельствует о наличии сильной ненаправленной или сильной положительной связи.
4. Значение коэффициента близкое к -1 свидетельствует о наличии сильной отрицательной связи.
5. Значение коэффициента равное 1 или -1 свидетельствует о наличии полной связи в терминах выбранной модели.

1.16. Таблица сопряженности. Проверка гипотезы о наличии связи в таблице сопряженности

Опр. Таблица сопряженности является средством представления совместного распределения частот, соответствующих значениям двух изучаемых признаков.

Таблица сопряженности строится для двух качественных признаков. Если необходимо построить таблицу сопряженности для количественного признака, его группируют в интервалы.

Структурные элементы таблицы сопряженности:

1. Строка и столбец заголовка – содержат описание значений признаков, для которых строится таблица сопряженности
2. Внутренние ячейки – образуются на пересечении строк и столбцов таблицы и содержат информацию о совместных частотных распределениях изучаемых признаков.

3. Маргинальная строка и маргинальный столбец – нижняя строка и правый крайний столбец, содержат суммы частот по столбцам и строкам или одномерные частотные распределения изучаемых признаков.

4. Ячейка, образованная маргинальной строкой и маргинальным столбцом – содержит информацию о суммарном числе респондентов, ответивших на вопросы, представленные в таблице.

Например, необходимо построить таблицу сопряженности для двух качественных признаков «Пол» и «Удовлетворены ли Вы своим уровнем жизни?» (Таблица 10).

Таблица 10 – Пример построения таблицы сопряженности

Пол	Удовлетворены ли Вы своим уровнем жизни			
	Удовлетворен	И да, и нет	Не удовлетворен	Всего
Мужской	10	15	5	30
Женский	15	8	12	35
	25	23	17	65

Чем больше значений у изучаемых признаков, тем больше в таблице сопряженности внутренних ячеек, что в последствии может привести к тому, что в них могут фиксироваться нулевые или близкие к нулевым частоты. Такая ситуация в итоге может привести к искажению вывода в проверяемой статистической гипотезы. Такие ситуации необходимо отслеживать и предотвращать.

Для таблицы сопряженности не существенно, какой признак располагается в строке, а какой в столбце. Главное правило, таблица должна хорошо читаться и не разрываться на несколько таблиц.

Рассмотрим общий вид таблицы сопряженности. Пусть 1-й признак располагается по столбцам, его значения обозначаются X_i где i изменяется от 1 до c (column). Пусть 2-й признак располагается по строкам, его значения обозначаются Y_j, где j изменяется от 1 до r (row). Тогда число объектов (респондентов), удовлетворяющих условию i-ой строки и j-го столбца будет равняться f_{ij} (Таблица 11).

Таблица 11 – Общий вид таблицы сопряженности

II-й признак (У)	I-й признак (Х)				
	X_1	X_2	… X_j …	X_c	
Y_1	f_{11}	f_{12}	… f_{1j} …	f_{1c}	f_{10}
Y_2	f_{21}	f_{22}	… f_{2j} …	f_{2c}	f_{20}
Y_i	f_{i1}	f_{i2}	… f_{ij} …	f_{ic}	f_{i0}
Y_r	f_{r1}	f_{r2}	f_{rj}	f_{rc}	f_{r0}
	f_{01}	f_{02}	f_{0j}	f_{0c}	f_{00}

Частоты, содержащиеся во внутренних ячейках таблицы сопряженности. Каждая внутренняя ячейка таблицы сопряженности, наряду с абсолютной частотой может содержать до трех видов относительных частот:

1. Относительная частота, удовлетворяющая условию строки
$(f_{ij}/f_{i0})100\%$;

2. Относительная частота, удовлетворяющая условию столбца
$(f_{ij}/f_{0j})100\%$;

3. Относительная частота от числа ответивших на два вопроса респондентов
$(f_{ij}/f_{00})100\%$.

Проверка гипотезы о связи в таблице сопряженности. С помощью таблиц сопряженности мы проверяем гипотезу о наличии статистической связи между двумя переменными.

Нулевая гипотеза всегда утверждает, что связь между переменными отсутствует.

Альтернативная всегда гипотеза утверждает, что связь между переменными присутствует.

Чтобы сформулировать гипотезу математически необходимо вычислить значения теоретических частот во внутренних ячейках изучаемой таблицы сопряженности.

Опр. Таблица сопряженности с вычисленными теоретическими частотами называется теоретической таблицей сопряженности. В теоретической таблице сопряженности полностью отсутствует статистическая связь.

Теоретические частоты обозначаются: e_{ij}

Теоретические частоты вычисляются по формуле: $e_{ij} = \frac{f_{i0} \times f_{0j}}{f_{00}}$

Математическая формулировка гипотезы о связи в таблице сопряженности:

H_0: $f_{ij} = e_{ij}$ для всех i и j;

H_1: $f_{ij} \neq e_{ij}$ хотя бы для некоторых i и j

Формула статистического критерия проверки гипотезы о связи в таблице сопряженности (Хи-квадрат) будет иметь вид:

$$\chi_H^2 = \sum\sum \frac{(f_{ij} - e_{ij})^2}{e_{ij}}, \text{ где}$$

f_{ij} – частоты во внутренних ячейках эмпирической таблицы сопряженности;

e_{ij} – частоты во внутренних ячейках теоретической таблицы сопряженности.

В основе формулы статистического критерия лежит теоретическое распределение Хи-квадрат, которое представляет собой асимметричный колокол, выходящий из точки ноль с длинным правым хвостом.

Вид колокола зависит от параметра «число степеней свободы», который для таблиц сопряженности вычисляется по формуле:

$df=(r-1)(c-1)$, где r – число строк, а c – число столбцов таблицы сопряженности.

Чтобы сделать вывод, сравниваем вычисленное значение статистики критерия и значение критической точки. Если статистика критерия больше критической точки нулевая гипотеза отклоняется и принимается альтернативная, что говорит о наличии статистически значимой связи в таблице сопряженности.

1.17. Таблицы сопряженности размером 2x2

Таблица сопряженности 2x2 строится для двух дихотомических переменных и состоит из двух строк и двух столбцов.

Дихотомический переменная (бинарная переменная) – это переменная, у которой фиксируются два значения:

1. Да (присутствие какого-либо интересующего исследователя признака).
2. Нет (отсутствие данного признака).

В Таблице 12 представлен общий вид таблицы сопряженности размером 2x2 с указанием буквенных обозначений внутренних ячеек:

a – совместное присутствие выделенных признаков;

d – совместное отсутствие выделенных признаков;

c – наличие признака X и отсутствие У;

b – наличие признака У и отсутствие X.

Таблица 12 – Общий вид таблицы сопряженности 2x2

Признак У	Признак X		
	1	0	Всего
1	a	b	a+b
0	c	d	c+d
Всего	a+c	b+d	a+b+c+d

В таблицах сопряженности размером 2x2 может фиксироваться направление связи. Поэтому для таблиц сопряженности 2x2 вводится понятие прямой (положительной) и обратной (отрицательной) связи.

Опр. Прямой называется связь, если признаки чаще появляются и не появляются совместно или, другими словами – положительной называется связь, если сумма частот по главной диагонали больше, чем сумма частот по не главной диагонали.

В Таблице 13 приведен пример таблицы сопряженности 2x2, в которой фиксируется прямая (положительная) связь.

Таблица 13 – Таблица сопряженности 2x2, в которой присутствует прямая (положительная) связь

Признак У	Признак X		
	1	0	Всего
1	23	5	28
0	7	19	26
Всего	30	14	44

Опр. Обратной называется связь, если признаки чаще появляются врозь, чем совместно или, другими словами – отрицательной называется связь, если сумма частот по второй диагонали больше, чем сумма частот по главной диагонали.

В Таблице 14 приведен пример таблицы сопряженности 2х2, в которой фиксируется обратная (отрицательная) связь.

Таблица 14 – Таблица сопряженности 2х2, в которой присутствует обратная (отрицательная) связь.

Признак У	Признак Х		
	1	0	Всего
1	3	29	32
0	17	5	22
Всего	20	34	44

Коэффициенты связи для таблиц сопряженности 2х2. Коэффициент связи ϕ (Фи) вычисляется по формуле:

$$\phi = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}},$$

Свойства коэффициента ϕ:

1. коэффициент ϕ изменяется в интервале от -1 до 1;
2. ϕ =1, если все ненулевые частоты расположены по главной диагонали;
3. ϕ =-1, если по не главной диагонали располагаются нулевые частоты;
4. если ϕ =1, говорим о полной прямой статистической связи;
5. если ϕ =-1, говорим о полной обратной статистической связи;
6. если ϕ =0, говорим об отсутствии статистической связи.

Коэффициент ϕ позволяет измерить связь, но не позволяет установить ее наличие. Наличие связи устанавливается с помощью критерия Хи-квадрат.

Для таблиц 2х2 значение коэффициента связи ϕ связано с Хи-квадрат через формулу: $\phi = \sqrt{\frac{\chi_H^2}{n}}$, тогда Хи-квадрат будет равен $\chi_H^2 = \phi^2 n$

Проверка гипотезы о наличии статистической связи в таблице 2х2. Проверяем гипотезу для заданного α=0,01 или 0,05 и числа степеней свободы равного 1.

Если вычисленное значение Хи-квадрат меньше критической точки, статистическая связь отсутствует.

Если вычисленное значение Хи-квадрат больше критической точки, статистическая связь присутствует.

1.18. Теоретико-информационные меры связи

Опр. Теоретико-информационные меры связи показывают насколько точнее станет прогноз распределения зависимой переменной, если имеются сведения о независимой переменной, по сравнению с точностью в случае, когда о независимой ничего не известно.

Общий вид теоретико-информационных мер связи:

$$\frac{U(y)-U(y/x)}{U(y)},$$ где

U(y) - количество ошибок прогноза У в случае игнорирования независимого признака Х.

U(y/x) – количество ошибок прогноза У с учетом знания распределения признака Х.

Коэффициент связи λ (лямбда) - Гутмана. Обозначение: $\lambda_{y/x}$, где

y – зависимая переменная;

x – независимая переменная.

Если зависимая переменная y располагается по столбцам, а независимая x по строкам, то формула $\lambda_{y/x}$ имеет вид:

$$\lambda_{x/y} = \frac{(\sum f_{imax}) - f_{0jmax}}{f_{00} - f_{0jmax}},$$ где

f_{00} – количество ответивших;

$f_{i\,max}$ – максимальная частота в i строке (модальное значение по строкам);

$f_{0j\,max}$ – максимальная частота в маргинальной строке (модальное значение в маргинальной строке).

Свойства меры λ - Гутмана:

1. $\lambda_{y/x}$ изменяется в интервале от 0 до 1.
2. $\lambda_{y/x}$ стремиться к 1, если в каждой строке существует ярко выраженное модальное значение и эти значения не пересекаются по столбцам.
3. $\lambda_{y/x}$ =1 в случае 100% предсказания y по x. Такая ситуация возникает, если все ненулевые частоты встречаются только по главной диагонали.
4. Значение коэффициента $\lambda_{y/x}$ = 0 в нескольких случаях:

   а) все ненулевые частоты содержатся только в одной строке;

   б) отсутствие феномена «модальности».
5. Если все модальные частоты сосредоточены в одном столбце. В этом случае создается ситуация, когда модальные значения присутствуют и теоретически вероятность предсказания y по x должна превышать 0, а коэффициент равен 0. В этом случае говорят, что коэффициент плохо ведет себя в нуле и прогноз на основании модальных значений не эффективен.

Коэффициент связи τ (тау) - Гудмана и Краскала. Мера *τ* - Гудмана и Краскала конструируется на основе предположения о том, что прогноз сводится не к единственному, хоть и модальному значению зависимой переменной, а к распределению значений зависимой переменной с определенными вероятностями.

Обозначение: $\tau_{y/x}$, где

y – зависимая переменная;

x – независимая переменная.

Если зависимая переменная y располагается по столбцам, а независимая x по строкам, то формула $\tau_{y/x}$ имеет вид:

$$\tau_{y/x} = \frac{\sum\sum\left[\frac{f_{ij}}{f_{00}}\left(\frac{f_{ij}}{f_{i0}} - \frac{f_{0j}}{f_{00}}\right)^2\right]}{1 - \sum\left(\frac{f_{0j}}{f_{00}}\right)^2},\text{ где}$$

f_{00} – количество ответивших;

f_{ij}/f_{i0} – частота в долях, удовлетворяющая условию i-ой строки;

f_{i0}/f_{00} – частота в долях в маргинальном столбце;

f_{0j}/f_{00} – частота в долях в маргинальной строке.

Свойства коэффициента τ - Гудмана и Краскала:

1. $\tau_{y/x}$, изменяется в интервале от 0 до 1;
2. $\tau_{y/x}$, = 0, если структура распределения по строкам одинакова и совпадает с распределением частот в маргинальной строке. В этом случае наблюдается статистическая независимость y от x;
3. $\tau_{y/x}$, =1, если ненулевые частоты располагаются по главной диагонали.

1.19. Ранжированные ряды. Меры парной связи ранжированных рядов

Опр. Ранжирование – это процедура упорядочивания любых объектов по возрастанию или убыванию некоторого их свойства при условии, что они этим свойством обладают.

Опр. Объектами ранжирования являются те объекты, которые непосредственно упорядочиваются.

Опр. Основание ранжирования или (ранжирующий признак) – это то свойство, по которому объекты упорядочиваются.

В результате ранжирования получаем ранжированный ряд, в котором каждому объекту приписывается ранг.

Опр. Ранг – это место объекта в ранжированном ряду. Число мест и, соответственно, число рангов равно числу объектов.

Виды ранжированных рядов:

1. Каждый объект имеет значение признака, отличное от значений признака у других объектов, тогда каждому объекту ранжированного ряда соответствует свой собственный, отличный от другого ранг.

Например, в качестве ранжируемых объектов выступают 9 государств. Необходимо проранжировать их по признаку «Качество жизни» (индекс), с учетом того, что каждый объект ранжирования имеет отличное от другого значение ранжирующего признака (Таблица 15). Ранги присваиваются в порядке убывания значения признака, т. е. 1 соответствует н6аибольшему значению, а 9 – наименьшему

Таблица 15 – Пример ранжирования, когда каждый объект имеет свой собственный ранг

Государства	А	Б	В	Г	Д	Е	Ж	З	И
Качество жизни	6,8	7,0	6,5	5,9	4,6	5,7	4,5	5,8	4,0
Ранги	2	1	3	4	7	6	8	5	9

2. Несколько объектов имеют одинаковые значения признака, тогда этим объектам в ранжированном ряду соответствуют одинаковые ранги, рассчитываемые по определенной формуле. В этом случае ранжированный ряд называется ранжированным рядом со связанными рангами.

Например, для семи работников предприятия определялся обобщенный показатель удовлетворенности работой. Необходимо проранжировать работников предприятия по их удовлетворенности работой с учетом того, что несколько объектов ранжирования могут иметь одинаковые значения ранжирующего признака (Таблица 16).

Таблица 16 – Пример ранжирования со связанными рангами

Работники	А	Б	В	Г	Д	Е	Ж
Удовлетворенность работой (индекс)	0,57	0,35	0,35	0,28	0,41	0,41	0,41
Ранги	1	5,5	5,5	7	3	3	3

Работники (Б и В) и (Д, Е, Ж) имеют одинаковые показатели удовлетворенности работой, следовательно, им необходимо присвоить связанные ранги.

Для значения 0,41 (Д, Е, Ж) ранг рассчитывается как среднее значение номеров мест, занимаемых этими объектами: (2+3+4)/3 = 3.

Для значения 0,35 (Б и В) ранг рассчитывается аналогично: (5+6)/2=5,5. Связанные ранги могут принимать дробные значения.

Ранжируем всего семь объектов, значит в ранжированном ряду должно быть семь рангов.

Коэффициенты ранговой корреляции. ***Опр.*** Коэффициентами ранговой корреляции называются меры связи, позволяющие вычислять степень согласованности в ранжировании одних и тех же объектов по двум различным основаниям или по двум различным признакам.

Коэффициент ранговой корреляции ρ (ро) – Спирмена. Допустим, что n объектов могут быть упорядочены как по признаку X, так и по признаку У.

Пусть R_{xi} – ранг i-го объекта по признаку X, R_{yi} – ранг i-го по признаку У, тогда $d_i = R_{xi} – R_{yi}$ – это мера несовпадения рангов.

Коэффициент ранговой корреляции ρ-Спирмена будет иметь вид:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}, \text{ где}$$

d_i– мера несовпадения рангов;

n – число ранжируемых объектоа

Проверка гипотезы о статистической значимости коэффициента ранговой корреляции ρ-Спирмена.

Формулировка гипотезы:

$H_0 : \rho_{гс} = 0$

$H_1 : \rho_{гс} \neq 0$

Коэффициент ранговой корреляции ρ-Спирмена статистически значим, если его значение для генеральной совокупности отлично от нуля.

Формула статистического критерия имеет вид $t_н = \rho\sqrt{\frac{n-2}{1-\rho^2}}$

Если статистика критерия меньше критической точки, подтверждается альтернативная гипотеза, следовательно, можно сделать вывод, что коэффициент ρ-Спирмена статистически значим для генеральной совокупности.

Коэффициент ранговой корреляции τ (тау)- Кендалла. ***Опр.*** Коэффициент τ – это разность между вероятностями правильного и неправильного порядка для двух наблюдений, извлеченных из совокупности случайно, при условии, что связанные ранги отсутствуют.

Возьмем пару объектов, ранги соответствующие первому объекту обозначим (i_1 j_1)

Ранги, соответствующие второму объекту, обозначим (i_2 j_2),

тогда коэффициент τ – Кендалла имеет вид**:** $\tau = \frac{S-D}{S+D}$, где

S – общее число пар объектов с согласующимся порядком по обоим объектам.

$i_1 > i_2$ и $j_1 > j_2$

$i_1 < i_2$ и $j_1 < j_2$

D – общее число пар объектов с несогласующимся порядком по обоим объектам.

$i_1 > i_2$ и $j_1 < j_2$

$i_1 < i_2$ и $j_1 > j_2$

Проверка гипотезы о статистической значимости коэффициента ранговой корреляции τ-Кендалла.

Формулировка гипотезы:

$H_0 : \tau_{гс} = 0$

$H_1 : \tau_{гс} \neq 0$

Коэффициент значим, если его значение для генеральной совокупности отлично от нуля.

Статистический критерий вид $Z_н = \tau\sqrt{\frac{S+D}{n(1-\tau^2)}}$

Если статистика критерия меньше критической точки, подтверждается альтернативная гипотеза, следовательно, можно сделать вывод, что коэффициент τ – Кендалла статистически значим для ГС.

1.20. Корреляционный анализ

Опр. Корреляционный анализ – это статистический метод, позволяющий обнаружить статистическую связь между двумя изучаемыми количественными переменными.

Чтобы охарактеризовать корреляцию, мы должны ответить на следующие вопросы:

1. Присутствует связь или нет.
2. Если связь присутствует, то каково ее направление.
3. Если связь присутствует, то какова ее сила.
4. Если связь присутствует, то является она линейной (монотонной) или нелинейной.

Предварительный анализ корреляции двух количественных признаков можно осуществить по графику под названием «Диаграмма рассеяния».

Опр. Диаграмма рассеяния представляет собой двумерный график у, которого по оси X откладываются значения одной переменной, а по оси Y – значения другой переменной. Объекты на графике изображаются в виде точек с координатами, равными значениям переменных для данного объекта. В итоге получаем облако точек (облако рассеяния), по которому можно сделать предварительные выводы о характере изучаемой связи.

По виду облака рассеяния можно сделать предварительные выводы о характере связи (корреляции) двух количественных признаков.

Виды и анализ диаграммы рассеяния:

1. Вытянутое облако с наклоном вправо (Рисунок 16) говорит о том, что:

- Связь есть.
- Связь средняя (чем плотнее расположены точки, тем связь сильнее).
- Связь прямая (положительная).
- Связь линейная.

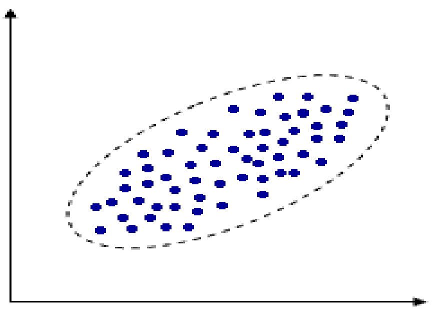


Рисунок 16 –Диаграмма рассеяния: средняя, прямая, линейная связь

2. Вытянутое облако с наклоном влево (Рисунок 17) говорит о том, что:

- Связь есть
- Связь средняя (чем плотнее расположены точки, тем связь сильнее)
- Связь обратная (отрицательная)
- Связь линейная

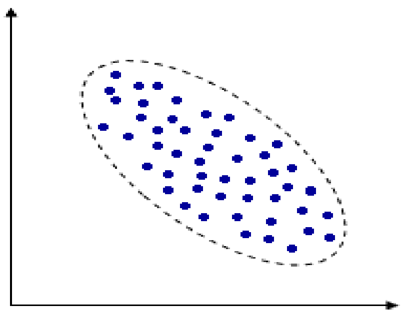


Рисунок 17 –Диаграмма рассеяния: средняя, обратная, линейная связь

3. Облако точек, имеющее перегиб (Рисунок 18), говорит о том, что:
   - Связь есть
   - Связь средняя (чем плотнее расположены точки, тем связь сильнее)
   - Связь нелинейная

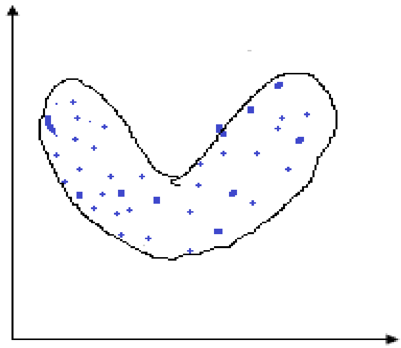


Рисунок 18 –Диаграмма рассеяния, в которой присутствует нелинейная связь (корреляция)

4. Облако точек без наклона к оси Х говорит о том, что связь (корреляция) отсутствует (Рисунок 19).

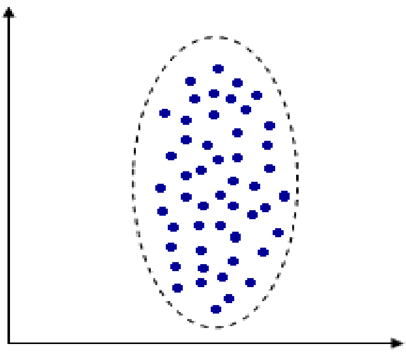


Рисунок 19 –Диаграмма рассеяния, в которой отсутствует связь (корреляция)

Ковариация двух количественных признаков. ***Опр.*** Ковариация – это совместная дисперсия двух изучаемых признаков.

В зависимости от направления связи ковариация может быть положительной или отрицательной.

Недостатком ковариации является то, что ее величина не ограничена сверху, так как ее значение зависит от масштаба измерения изучаемых признаков и объема выборки.

В связи с этим ковариация не удовлетворяет свойству мер связи о том, что меры связи изменяются от -1 до 1 и не может быть проинтерпретирована как мера связи.

Чтобы избавиться от влияния масштаба и объема выборки ковариацию нормируют делением на среднеквадратическое отклонение по X и по У.

Полученный коэффициент называется коэффициентом линейной корреляции Пирсона.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}},$$

Свойства коэффициента линейной корреляции r-Пирсона:

1. Коэффициент равняется 0, когда ковариация равна 0, т.е. когда связь (корреляция) отсутствует.
2. Коэффициент равен 1 при полной линейной прямой связи.
3. Коэффициент равен -1 при полной линейной обратной связи.
4. Если значение коэффициента фиксируются в интервале от 0 до 0,3 – связь слабая.

Если значение коэффициента фиксируются в интервале от 0,3 до 0,6 – связь средняя.

Если значение коэффициента фиксируются в интервале от 0,6 до 1 – связь сильная.

Проверка статистической значимости коэффициентом линейной корреляции Пирсона. Формулировка гипотезы:

$H_0: r_{гс} = 0$

$H_1: r_{гс} \neq 0$

Коэффициент линейной корреляции r-Пирсона статистически значим, если его значение для генеральной совокупности отлично от нуля.

Формула статистического критерия имеет вид $t_н = r\sqrt{\frac{n-2}{1-r^2}}$

Определяем значение критической точки по статистической таблице распределения t-Стьюдента.

Если статистика критерия меньше критической точки, подтверждается альтернативная гипотеза, следовательно, можно сделать вывод, что коэффициент значим для ГС.

1.21. Регрессионный анализ. Парная линейная регрессия

Опр. Регрессионный анализ – это метод изучения статистической причинной связи между одной зависимой количественной переменной и одной или несколькими независимыми количественными переменными.

Процедура регрессионного анализа включает в себя: определение формы связи, построение уравнения регрессии, оценку и анализ полученного уравнения.

Уравнение регрессии приближенно выражает зависимость среднего значения объясняемого признака от одного или нескольких независимых признаков-факторов (предикторов).

Регрессионная модель в общем виде представляет собой функциональную зависимость одной переменной от одной или нескольких независимых переменных $\hat{y}= f\ (x1,\ x2,\ \ldots,\ x3)$

Ограничения регрессионного анализа:

1. Зависимая (результирующая) переменная должна быть непрерывной количественной переменной. Независимая переменная должна быть непрерывной количественной или дихотомической.
2. Изучаемая совокупность должна быть достаточно большой, чтобы показатели связей были статистически надежными (число единиц совокупности должно превосходить число коррелируемых переменных не менее чем в 6-8 раз).
3. Каждое значение зависимой переменной должно быть независимо от других значений.
4. Должно выполняться требование гомоскедактичности, что означает, что ошибки предсказания не становятся меньше, если уменьшается значение **у** и не растут с увеличением значений **у**
5. Ошибка предсказания для каждого значения не должна зависеть от ошибки предсказания других значений, остатки должны быть нормально распределены.
6. Для случая множественной регрессии должно отсутствовать явление мультиколлинеарности, которое возникает, когда независимые переменные сильно коррелируют между собой. Такого рода корреляция может оказать сильное воздействие на зависимый признак и это уже будет иное воздействие, чем независимых переменных по отдельности.
7. Распределение зависимой переменной должно быть нормальным и не иметь явных выбросов

Уравнение регрессии, описывающее зависимость результирующего признака от одного признака-фактора называется уравнением парной линейной регрессии $у=в_0 + вх$, где

$в_0$ и в – параметры уравнения регрессии;

параметр в – коэффициент уравнения регрессии;

параметр $в_0$ – свободный член уравнения регрессии.

Интерпретация коэффициента уравнения регрессии (в). При изменении значения признака-фактора на 1, зависимый признак изменяется на величину коэффициента.

Расчет параметров уравнения парной линейной регрессии. Искомые параметры должны быть средними для всей совокупности данных.

Сумма квадратов отклонений от среднего арифметического меньше суммы квадратов отклонений от любой другой величины. Соответственно, расчет параметров уравнения регрессии осуществляется методом наименьших квадратов.

Оценка полученного уравнения парной линейной регрессии. Вывод о практической значимости полученного уравнения регрессии делается на основе значения коэффициента детерминации.

Опр. Коэффициент детерминации – это коэффициент линейной корреляции Пирсона, возведенный в квадрат.

Коэффициент детерминации показывает процент дисперсии зависимого признака, объясняемый действием признака-фактора. Чем выше этот процент, тем более качественную регрессионную модель Вы получили.

1.22. Дисперсионный анализ. Однофакторная дисперсионная модель

Опр. Дисперсионный анализ – это метод, используемый для изучения влияния различных, одновременно действующих и независимых признаков-факторов на изменчивость одного или нескольких зависимых признаков.

В дисперсионном анализе зависимая переменная всегда количественная, независимые переменные всегда качественные.

Гипотеза однофакторного дисперсионного анализа:

Пусть фактор имеет k значений, тогда

H_0: $\mu_1 = \mu_2 = \ldots = \mu_3 = \mu_k = \mu$ (на всех уровнях фактора средние значения зависимой переменной равны)

H_1: $\mu_i \neq \mu$ для некоторых i, изменяющихся от 1 до k (хотя бы для некоторых групп, образованных уровнями фактора, значения средних не совпадает)

Дисперсия зависимой переменной, как показатель степени разброса или неоднородности данных, включает в себя две составляющие:

Одна составляющая порождается внутри каждого уровня фактора и называется внутригрупповой дисперсией.

Вторая порождается при переходе от одного уровня фактора к другому и называется межгрупповой дисперсией.

В дисперсионном анализе вместо дисперсии используется только числитель формулы дисперсии, который получил название общая сумма квадратов. Общая сумма квадратов, согласно модели разделения дисперсий раскладывается на внутригрупповую и межгрупповую суммы квадратов:

$$SS_{\text{общ}} = SS_{\text{вгр}} + SS_{\text{мгр}} = SS_{\text{ост}} + SS_{\text{A}}$$

Чтобы проверить гипотезу дисперсионного анализа нужно вычислить F-отношение.

F= $\frac{MSS_A}{MSS_{\text{ост}}}$, где

MSS_A – средние квадраты фактора;

$MSS_{\text{ост}}$ – средние квадраты остатков.

Вычисление средних квадратов фактора и средних квадратов остатков осуществляется следующим образом:

$MSS_A = \frac{SS_A}{df_A}$, где

df_A – число степеней свободы фактора,

которое вычисляется по формуле df_A = k-1

$MSS_{\text{ост}} = \frac{SS_{\text{ост}}}{df_{\text{ост}}}$, где

$df_{\text{ост}}$ – число степеней свободы остатков,

которое вычисляется по формуле df_A = n - k

Сумма квадратов фактора вычисляется по формуле:

$SS_A = \sum n_k(\bar{y}_k - \bar{y})^2$, где

n_k – количество объектов на каждом уровне фактора;

$\bar{y}_k$ – среднее арифметическое значение зависимой переменной на каждом уровне фактора;

$\bar{y}$ – среднее арифметическое для всех значений зависимой переменной.

Сумма квадратов остатков вычисляется по формуле:

$SS_{\text{ост}} = \sum\sum(y_i - \bar{y}_k)^2$, где

$\bar{y}_k$ – среднее арифметическое зависимой переменной на каждом уровне фактора;

y_i – индивидуальные значения зависимой переменной

Вычислив F-отношение необходимо сравнить его значение со значением критической точки, которая находится по статистической таблице F-Фишера в соответствии со значением доверительной вероятности, числа степеней свободы фактора и числа степеней свободы остатков.

Если F-отношение больше $F_{\text{кр}}$ подтверждается альтернативная гипотеза и мы можем сделать вывод, что независимая переменная (фактор) статистически значимо влияет на изменчивость зависимой переменной.

2. ПРАКТИЧЕСКИЙ РАЗДЕЛ

Тематика семинарских занятий по учебной дисциплине «Статистический анализ социологической информации»

Семинар № 1. **Тема 2. Измерение в социологии, измерительные шкалы**

Определение понятий «признак» и «значение признака», построение матрицы данных типа "объект-признак". Измерение социологических признаков с помощью основных измерительных шкал: номинальная шкала, порядковая шкала, интервальная шкала, шкала отношений.

(Форма контроля – устный ответ, письменный отчет по аудиторным практическим упражнениям)

Семинар № 2. **Тема 3. Одномерное частотное распределение**

Рассмотрение понятий: «абсолютная частота», «относительная частота в процентах», «относительная частота в долях от единицы». Расчет и интерпретация частот в таблицы одномерных частотных распределений. Расчет накопленных частот.

(Форма контроля – устный ответ, проверка письменных отчетов выполнения домашних упражнений, письменный отчет по аудиторным практическим упражнениям)

Семинар № 3. **Тема 4. Группировки количественных признаков в интервалы**

В каких случаях прибегают к процедурам группировки количественных признаков в интервалы. Задачи, решаемые при осуществлении группировки. Решение задач на построение типологической группировки. Решение задач на построение аналитической группировки. Решение задач на построение процентильной группировки.

(Форма контроля – устный ответ, проверка письменных отчетов выполнения домашних упражнений, письменный отчет по аудиторным практическим упражнениям)

Семинары № 4-5. **Тема 5. Графическое представление социологических данных**

Использование графиков в репрезентации результатов социологических исследований. Основные типы используемых графиков: диаграммы, гистограмма, полигон распределения, графики накопленных частот, график интерквартельного диапазона. Ошибки восприятия информации, возникающие при некорректном построении графиков.

(Форма контроля – письменный контрольный опрос по пройденным темам 1-5. Письменная контрольная работа № 1 по темам 1-5)

Семинар № 6. **Тема 6. Характеристики центра распределения признака**

Характеристики центра распределения: Мода, Медиана, среднее арифметическое. Соотношение понятий «среднее» и «типичное» значение признака. Способы определения характеристик центра.

(Форма контроля – проверка письменных отчетов выполнения домашних упражнений, письменный отчет по аудиторным практическим упражнениям)

Семинар № 7. **Тема 7. Показатели вариации признака**

Понятие вариации признака. Показатели вариации признак: размах вариации, среднее линейное отклонение, дисперсия, среднеквадратическое отклонение, коэффициент вариации. Вычисление показателей вариации.

(Форма контроля – устный ответ, проверка письменных отчетов выполнения домашних упражнений, письменный отчет по аудиторным практическим упражнениям)

Семинар № 8. **Тема 8. Анализ формы распределения признака**

Анализ формы распределения количествееного признака по трем характеристикам: модальность, симметричность и протяженность. Виды модальности распределений: одномодальное и полимодальное; колоколообразное и j-образное. Симметричные и асимметричные распределения, вычисление и интерпретация коэффициента асимметрии. Протяженность распределения: лептокритичечские и платокритические раараспределения.

(Форма контроля – устный ответ, проверка письменных отчетов выполнения домашних упражнений, письменный отчет по аудиторным практическим упражнениям)

Семинар № 9. **Тема 9. Стандартизация количественных переменных. Z-оценки**

Вычисление Z-оценки и использование в качестве безразмерной стандартизованной переменной. Сравнение распределений исходного признака и его Z-оценки. Контрольная работа по пройденным и взаимосвязанным темам: характеристики центра распределения, показатели вариации, анализ формы распределения признака, Z-оценки.

(Форма контроля – письменная контрольная работа № 2 по темам 6-9)

Семинар № 10. **Тема 10. Теоретические распределения и их статистические таблицы**

Нормальное и стандартное нормальное распределение, статистическая таблица стандартного нормального распределения. Задачи, решаемые с помощью статистической таблицы стандартного нормального распределения.

Другие теоретические распределения, используемые в статистическом анализе социологических данных: распределение t - стьюдента, распределение F - Фишера, распределение Хи-квадрат Пирсона.

(Форма контроля – письменный отчет по аудиторным практическим упражнениям)

Семинар № 11. **Тема 11. Статистический вывод. Оценка параметров генеральной совокупности**

Основные параметры генеральной совокупности и соответствующие им статистики, полученные по выборке. Выборочное распределение статистики. Точечные оценки параметров генеральной совокупности. Свойства точечных оценок. Интервальное оценивание параметров генеральной совокупности. Решение задач на построение доверительных интервалов в зависимости от заданного уровня вероятности случайной ошибки и размера ошибки выборки.

(Форма контроля – устный ответ, проверка письменных отчетов выполнения домашних упражнений, письменный отчет по аудиторным практическим упражнениям)

Семинар № 12. **Тема 12. Простая случайная выборка из генеральной совокупности. Ошибка простой случайной репрезентативной выборки и ее объем**

Определение простой случайной выборки, методы формирования простой случайной выборки. Определение репрезентативности выборки. Понятие допустимой ошибки выборки. Расчет объема выборки в зависимости от заданного уровня вероятности случайной ошибки и размера допустимой ошибки выборки.

(Форма контроля – письменный контрольный опрос по темам 10-12)

Семинары № 13-15. **Тема 14. Виды статистических гипотез: гипотезы о доле, гипотезы о средних, гипотезы о дисперсиях**

Проверка статистической гипотезы о долях для дихотомического признака. Проверка статистической гипотезы о средних для одной выборочной совокупности, проверка статистической гипотезы о средних для двух независимых выборочных совокупностей, проверка статистической гипотезы о равенстве дисперсий двух независимых выборочных совокупностей.

(Форма контроля – письменный контрольный опрос по темам 13,14; проверка письменных отчетов выполнения домашних упражнений; письменная контрольная работа № 3 по темам 13,14)

Семинар № 16. **Тема 16. Таблица сопряженности. Проверка гипотезы о наличии связи в таблице сопряженности**

Процедура построения таблицы сопряженности. Формулировка гипотезы о статистической независимости строк и столбцов таблицы сопряженности двух признаков. Теоретическая таблица сопряженности. Проверка гипотезы по критерию Хи-квадрат.

(Форма контроля – устный ответ, проверка письменных отчетов выполнения домашних упражнений, письменный отчет по аудиторным практическим упражнениям)

Семинар № 17. **Тема 17. Таблицы сопряженности 2х2**

Общий вид таблиц сопряженности 2х2, обозначение ячеек. Построение таблицы 2х2 из таблицы с произвольным количеством строк и столбцов. Меры связи для таблицы 2х2: коэффициент Юла, коэффициент Пирсона Сравнительное использование значений коэффициентов Юла и Пирсона.

(Форма контроля – письменный отчет по аудиторным практическим упражнениям)

Семинары № 18-19. **Тема 19. Ранжированные ряды. Меры парной связи ранжированных рядов**

Процедура ранжирования, процедура ранжирования рядов со связанными рангами. Прямая и обратная ранговая связь. Вычисление ранговых коэффициентов связи (мер связи) ρ - Спирмена и τ - Кендала. Интерпретация полученных коэффициентов.

(Форма контроля – устный ответ, проверка письменных отчетов выполнения домашних упражнений, письменный отчет по аудиторным практическим упражнениям)

Семинар № 20. **Тема 20. Корреляционный анализ**

Построение и интерпретация диаграммы рассеяния как предварительный этап анализа линейной корреляции двух количественных признаков. Расчет коэффициента линейный корреляции r – Пирсона. Интерпретация силы связи анализируемых признаков по значению коэффициента r – Пирсона.

(Форма контроля – устный ответ, проверка письменных отчетов выполнения домашних упражнений, письменный отчет по аудиторным практическим упражнениям)

Семинар № 21-22. **Тема 21. Регрессионный анализ. Парная линейная регрессия**

Вычисление параметров уравнения парной линейной регрессии. Интерпретация коэффициента уравнения регрессии. Проверка качества полученной регрессионной модели через вычисление коэффициента детерминации.

Написание контрольной работы по пройденным темам.

(Форма контроля –письменный отчет по аудиторным практическим упражнениям; письменная контрольная работа № 4 по темам 16-21)

Семинар № 23. **Тема 22. Дисперсионный анализ. Однофакторная дисперсионная модель**

Постановка задачи дисперсионного анализа. Общая модель дисперсионного анализа с разделением дисперсий. Понятие суммы квадратов. Разложение общей суммы квадратов на межгрупповую и внутригрупповую. Расчет F- отношения, проверка гипотезы

(Форма контроля – устный ответ, письменный отчет по аудиторным практическим упражнениям)

3. РАЗДЕЛ КОНТРОЛЯ ЗНАНИЙ

3.1. Методические рекомендации и примерный перечень заданий по УСР

При изучении учебной дисциплины рекомендуется использовать следующие формы самостоятельной работы: – поиск (подбор) литературы и электронных источников, условий задач, основанных на данных количественных социологических исследований для выполнения расчетов, построения графиков и диаграмм. Изучение теоретических материалов и выполнение практических заданий, размещенных на образовательном портале (LMS Moodle), тестирование по пройденным темам с использованием возможностей образовательного портала (LMS Moodle).

Примерный перечень заданий для УСР

Задание 1. Для данных, представленных в Таблице 17:

Таблица 17 – Данные для выполнения задания 1

Сколько времени Вы вчера провели в интернете (в часах)	Частота	Проценты
1	3	3,6
2	9	9,7
3	8	10,8
4	15	18,2
4,5	1	1,2
5	18	21,7
5,5	1	1,2
6	7	8,4
7	1	1,2
8	5	6,0
9	2	2,4
10	6	7,2
11	1	1,2
12	6	7,2
Всего	n=83	100

1. По вариационному ряду построить аналитическую группировку, указать абсолютные частоты, относительные частоты в процентах и возрастающую накопленную частоту в процентах.
2. По вариационному ряду рассчитать возрастающую накопленную частоту построить квартильную группировку с приблизительными границами.
3. По аналитической группировке построить квартильную группировку с точными границами.
4. По аналитической группировке построить терцильную, квинтильную и децильную группировки.

5. По аналитической группировке построить гистограмму и полигон распределения.

6. По квартильной группировке построить график интерквартильного диапазона.

Задание 2. Имеются данные о продолжительности подготовки студентов к занятиям (Таблица 18). На вопрос «Сколько времени Вы в среднем затрачиваете на подготовку домашних заданий по предметам в день?» были получены ответы 35 человек, представленные в виде вариационного ряда.

Таблица 18 – Данные для выполнения задания 2

Время на подготовку домашних заданий (в часах)	Частота	Процент
1	4	11,4
1,5	7	20,0
2	10	28,6
3	5	14,3
3,5	7	20,0
4	2	5,7
	n=35	100

Необходимо выполнить следующие задания:

1. Определить или вычислить моду, медиану, среднее арифметическое, проинтерпретировать полученные характеристики центра распределения.
2. Определить или вычислить СКО, коэффициент вариации, интерквартельный диапазон. Проинтерпретировать полученные показатели вариации.
3. Построить полигон распределения
4. По полигону распределения, характеристикам центра и показателям вариации проанализировать форму распределения признака.
5. Перевести исходные значения в стандартный масштаб с помощью формулы Z-оценки, построить полигон распределения для данных в стандартном масштабе, сравнить два графика и сделать вывод о влиянии стандартизации на форму распределения.

Задание 3. В городе N было проведено выборочное исследование и получена информация о возрасте первого вступления в брак у мужчин и женщин. Данные исследования представлены в «сыром виде»:

Мужчины: 20 28 23 27 35 30 18 19 21 38 35 40 26;

Женщины: 23 22 24 18 19 20 30 32 31 17 18 28 20 18.

1. Проверить гипотезу о том, что доля вступивших в брак до 20 лет включительно у мужчин и женщин совпадает. Проверку произвести на уровне значимости α=0,01
2. Проверить гипотезу о том, что доля вступивших в первый брак до 25 лет включительно по городу не зависимо от пола статистически значимо

отличается от данных официальной статистики по стране, которая равна 30%. Проверку произвести на уровне значимости α=0,01

3. Проверить гипотезу о том, что средний возраст вступления в первый брак у мужчин и женщин не совпадает. Проверку произвести на уровне значимости α=0,05

4. Проверить гипотезу о том, что средний возраст вступления в первый брак по городу не зависимо от пола статистически значимо отличается от данных официальной статистики по стране, которая равна 23 годам. Проверку произвести на уровне значимости α=0,05

Задание 4. Проранжировать объекты (Таблица 19) по каждому из двух признаков и вычислить коэффициент ранговой корреляции ρ-Спирмена, проинтерпретировать полученный коэффициент. Проверить гипотезу о статистической значимости коэффициента ρ-Спирмена для α=0,05. Проверить статистическую значимость полученного коэффициента.

Таблица 19 – Данные для выполнения задания 4

Ученики	А	Б	В	Г	Д	Е	Ж	З
Показатель IQ	100	101	120	115	112	108	118	106
Оценка теста по математике	15	18	20	17	13	11	19	12

Задание 5. Проранжировать объекты (Таблица 20) по каждому из двух признаков и вычислить коэффициент ранговой корреляции τ-Кендалла, проинтерпретировать полученный коэффициент. Проверить гипотезу о статистической значимости коэффициента τ-Кендалла для α=0,05. Проверить статистическую значимость полученного коэффициента.

Таблица 20 – Данные для выполнения задания 5

Ученики	А	Б	В	Г	Д	Е	Ж	З
Показатель IQ	105	101	110	115	112	108	118	107
Оценка теста по химии	16	17	20	10	13	11	19	14

Задание 6. Сформулировать и проверить гипотезу о наличии связи в таблице сопряженности (Таблица 21), проинтерпретировать результат. Вычислить коэффициент связи Крамера.

Таблица 21 – Данные для выполнения задания 5

Курс	Важно ли для вас получение хорошего образования?			
	Не важно	Затрудняюсь	Важно	Всего
Второй	24	12	54	90
Третий	16	4	63	83
Четвертый	7	13	43	63
Всего	47	29	160	236

3.2. Примерные варианты тестовых заданий

1. Что в матрице «объект-признак» располагается по столбцам?
   а) объекты   в) проценты
   б) частоты   г) признаки

2. Какая группировка предполагает разбиение на заданное число интервалов равной длины?
   а) типологическая   в) децильная
   б) аналитическая   г) квартильная

3. Признак «Возраст респондента» является?
   а) качественным
   б) непрерывным количественным
   в) дискретным количественным
   г) качественно-количественным

4. Для каких измерительных шкал рассчитывается среднее арифметическое?
   а) номинальных
   б) порядковых
   в) шкал отношений
   г) категориальных

5. Параметрами стандартного нормального распределения являются?
   а) μ и σ б) 0 и 1   в) x и S   г) μ и S

6. Признаком левосторонней асимметрии является?
   а) вытянутый левый хвост распределения
   б) симметрично вытянутые левый и правый хвосты распределения
   в) вытянутый правый хвост распределения
   г) симметричные левый и правый хвосты распределения

7. Признаком правосторонней асимметрии является?
   а) вытянутый левый хвост распределения
   б) совпадение значений трех мер центральной тенденции
   в) вытянутый правый хвост распределения
   г) симметричные левый и правый хвосты распределения

8. Признаком симметрии распределения является?
   а) вытянутый левый хвост распределения
   б) совпадение значений трех мер центральной тенденции
   в) вытянутый правый хвост распределения
   г) платокритическая форма колокола

9. Гистограмма – это график для отображения распределения?
- а) качественного признака
- б) количественного признака
- в) количественного признака, сгруппированного в интервалы
- г) дихотомического признака

10. Какая статистика соответствует параметру μ (математическое ожидание)?
- а) среднеквадратическое отклонение
- б) доля признака
- в) дисперсия
- г) среднее арифметическое

11. С каким знаком всегда формулируется нулевая статистическая гипотеза?
- а) больше
- б) меньше
- в) равенства
- г) неравенства

12 С каким знаком не может быть сформулирована альтернативная статистическая гипотеза?

а) больше	в) равенства
б) меньше	г) неравенства

13. Для проверки гипотезы о наличии связи в таблице сопряженности используется теоретическое распределение?
- а) Хи-квадрат
- б) стандартное нормальное
- в) t-распределение Стьюдента
- г) F-распределение Фишера

14. Для проверки гипотезы о равенстве средних двух совокупностей используется теоретическое распределение?
- а) Хи-квадрат
- б) стандартное нормальное
- в) t-распределение Стьюдента
- г) F-распределение Фишера

15. В однофакторном дисперсионном анализе для проверки статистической гипотезы используется теоретическое распределение:
- а) Хи-квадрат
- б) стандартное нормальное
- в) t-распределение Стьюдента
- г) F-распределение Фишера

16. Таблица сопряженности используется для проверки наличия связи?
а) двух количественных признаков
б) количественного и порядкового признаков
в) дихотомического и количественного признаков
г) двух качественных признаков

17. Почему ковариация не может быть использована как показатель тесноты связи двух количественных признаков?
а) ее значение не ограничено сверху
б) ее значение может быть как со знаком плюс, так и со знаком минус
в) ее значение слишком мало
г) ее значение не равняется нулю

18. В уравнении парной линейной регрессии параметр «в» – это?
а) свободный член уравнения регрессии
б) сомножитель уравнения регрессии
в) коэффициент уравнения регрессии
г) погрешность уравнения регрессии

19. В каком случае мера связи двух признаков может быть распространена на генеральную совокупность?
а) если ее значение для генеральной совокупности отлично от нуля
б) если ее значение для генеральной совокупности равно нулю
в) если можно вычислить ее значение

20. Какое действие не является одним из видов статистического вывода?
а) статистическое оценивание точечное
б) статистическое оценивание интервальное
в) построение случайной выборки
г) проверка статистических гипотез

3.3. Примерный перечень вопросов к зачету и экзамену

Примерный перечень вопросов к зачету:

1. Данные социологического исследования. Матрица «объект-признак».
2. Процедура измерения в социологии.
3. Измерительные шкалы для качественных признаков.
4. Измерительные шкалы для количественных признаков.
5. Одномерный частотный анализ: абсолютные, относительные, накопленные частоты.

6. Общие принципы группировки количественных признаков в интервалы
7. Типологическая группировка.
8. Аналитическая группировка.
9. Процентильные группировки.
10. Графическое представление качественных признаков (диаграммы).
11. Графическое представление количественных признаков (гистограмма, полигон распределения, график интерквартильного диапазона).
12. Характеристики положения центра распределения: мода, медиана, среднее арифметическое.
13. Вариация признака. Показатели вариации: размах вариации, дисперсия, среднеквадратическое отклонение, коэффициент вариации.
14. Анализ формы распределения по его характеристикам.
15. Стандартизация количественных переменных: Z – оценки и их свойства.
16. Нормальное распределение, его параметры и свойства.
17. Стандартное нормальное распределение.
18. Статистическая таблица стандартного нормального распределения.
19. Определение статистического вывода. Статистики и соответствующие им параметры генеральной совокупности.
20. Выборочное распределение. Следствие из центральной предельной теоремы.
21. Точечное оценивание параметров генеральной совокупности.
22. Интервальное оценивание параметров генеральной совокупности.
23. Ошибка выборки.
24. Репрезентативность простой случайной выборки.
25. Объем выборки.

Примерный перечень вопросов к экзамену:
1. Данные социологического исследования. Матрица «объект-признак».
2. Процедура измерения в социологии.
3. Качественные признаки. Измерительные шкалы для качественных признаков.
4. Количественные признаки. Измерительные шкалы для количественных признаков.
5. Одномерный частотный анализ. Расчет, виды и анализ накопленных частот.
6. Основные принципы группировки данных в интервалы.
7. Виды группировок: типологическая, аналитическая, процентильная.
8. Графическое представление номинальных и порядковых шкал
9. Графическое представление количественных признаков
10. Характеристики центра распределения признака: мода, медиана, среднее арифметическое.

11. Характеристики степени разброса данных: размах, дисперсия, среднеквадратическое отклонение, коэффициент вариации.

12. Анализ формы распределения.

13. Нормальное распределение, его параметры и свойства.

14. Стандартизация количественных переменных (Z-оценки и их свойства).

15. Стандартное нормальное распределение, его статистическая таблица.

16. Теоретические распределения (Стьюдента, Хи-квадрат, Фишера) и их статистические таблицы.

17. Понятие статистического вывода. Точечное оценивание. Свойства точечного оценивания.

18. Интервальное оценивание. Построение доверительных интервалов

19. Определение простой случайной выборки. Виды случайного отбора.

20. Ошибка выборки, стандартная ошибка, допустимая ошибка.

21. Понятие репрезентативной выборки.

22. Объем выборки для бесконечной и конечной генеральной совокупности, объем выборки для доли.

23. Понятие статистической гипотезы.

24. Статистические критерии проверки гипотезы.

25. Процедура проверки статистической гипотезы.

26. Виды статистических гипотез.

27. Понятие статистической связи и статистической независимости двух показателей, классификация статистических связей.

28. Основные модели статистических связей.

29. Понятие меры связи. Общие свойства мер связи.

30. Таблица сопряженности двух признаков.

31. Проверка гипотезы о наличии связи в таблице сопряженности. Меры связи, основанные на критерии Хи-квадрат.

32. Таблица сопряженности (2х2) для двух дихотомических признаков.

33. Коэффициент φ (Фи) для таблицы сопряженности (2х2). Проверка гипотезы о статистической значимости коэффициента Фи.

34. Общая конструкция теоретико-информационных мер связи.

35. Мера связи λ (лямбда) - Гутмана. Свойства меры λ - Гутмана.

36. Мера связи τ (тау) - Гудмана и Краскала. Свойства меры τ - Гудмана и Краскала.

37. Процедура ранжирования.

38. Коэффициент ранговой корреляции ρ-Спирмена. Интерпретация коэффициента ρ-Спирмена. Проверка статистической значимости коэффициента ρ-Спирмена.

39. Коэффициент ранговой корреляции ρ-Спирмена с поправкой на связанные ранги.

40. Коэффициент ранговой корреляции τ-Кендала. Интерпретация коэффициента τ- Кендала. Проверка статистической значимости коэффициента τ-Кендала.

41. Корреляционный анализ. Диаграмма рассеяния, виды диаграммы рассеяния.

42. Ковариация двух количественных признаков. Ковариация и дисперсия.

43. Линейный коэффициент корреляции r_{xy}-Пирсона. Проверка значимости коэффициента корреляции Пирсона.

44. Регрессионный анализ как метод статистического анализа. Требования и ограничения регрессионного анализа к исходным данным.

45. Парная линейная регрессия. Уравнение регрессии. Параметры уравнения парной линейной регрессии.

46. Расчет параметров уравнения парной линейной регрессии методом наименьших квадратов.

47. Вычисление и интерпретация коэффициента детерминации. Анализ качества полученного регрессионного уравнения.

48. Дисперсионный анализ. Однофакторный дисперсионный анализ.

4. ВСПОМОГАТЕЛЬНЫЙ РАЗДЕЛ

4.1. Рекомендуемая литература

Основная

1. Дерр, В. Я. Теория вероятностей и математическая статистика: учебное пособие для вузов / В. Я. Дерр. – Санкт-Петербург Лань : 2021. – 596 с.
2. Математические методы исследования социальных систем: учебное пособие / Торопов Б. А. и др. – М.: Академия упр. МВД России, 2020. – 80 с. – https://mvd.ru/upload/site120/folder_page/015/122/996/Toropov_interak_sayt – Дата доступа 09.06.2023.
3. Плескунов, М.А. Методы статистического анализа социологических данных [Электронный ресурс]: учеб. пособие / М.А. Плескунов.– Екатеринбург : Изд-во Урал. ун-та, 2017. – Режим доступа: – http://hdl.handle.net/10995/46996 – Дата доступа 18.05.2023.

Дополнительная

1. Аптон, Г. Анализ таблиц сопряженности. М.: Финансы и статистика, 1982. – 140 с.
2. Бююль, А. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей / А. Бююль, П. Цефель. – М. : DiaSoft , 2005, – 600 с.
3. Бослаф, С Статистика для всех. / Сара Бослаф, пер. с англ. П. А. Волкова, И. М. Флямлер, М. В. Либерман, А. А. Галицына. – М.: ДМК Пресс, 2017. – 586 с.
4. Гласс, Дж. Статистические методы в педагогике и психологии / Дж. Гласс, Дж. Стенли. – М.: Прогресс, 1976, – 495 с.
5. Доннели-мл, Роберт А. Статистика: шаг за шагом / Роберт А. Доннели-мл. – М., АСТ Астрем, 2007, – 306с.
6. Крыштановский, А. О. Анализ социологических данных / А. О. Крыштановский. – М.: Издательский дом ГУ ВШЭ, 2006. – 281 с.
7. Наследов, А. Д. Математические методы психологического исследования: анализ и интерпретация данных: учебное пособие студентов высших учебных заведений, обучающихся по направлению и по специальностям психология / А. Д. Наследов. –. Санкт-Петербург:Речь, 2012. –387 с.
8. Паниотто, В. И., Максименко В. С. Количественные методы в социологических исследованиях / В. И. Паниотто, В. С. Максименко. – Киев: Наукова думка, 1982. – 272 с.
9. Татарова, Г.Г. Методология анализа данных в социологии / Г. Г. Татарова. – М.: NOTA BENE, 1999. – 223 с.
10. Терещенко О.В. Прикладная статистика для социальных наук: Компьютерный практикум для студентов гуманитарных специальностей / О. В. Терещенко. – Мн. БГУ, 2002. – 93с.

11. Толстова, Ю.Н. Измерение в социологии / Ю. Н. Толстова. – М.: Инфра-М, 1998. – 221 с.

12. Толстова, Ю.Н. Математико-статистические модели в социологии (математическая статистика для социологов): учеб. Пособие / Ю. Н. Толстова. – М.: Изд. дом ГУ ВШЭ, 2008. – 243 с.

13. Хили, Дж. Статистика: социологические и маркетинговые исследования / Дж. Хили. – М. : Dia Soft., 2005. – 634 с.

4.2. Электронные ресурсы

1. Образовательный портал БГУ [Электронный ресурс]. – Режим доступа: http://dl.bsu.by. – Дата доступа: 06.03.2023.

2. Электронная библиотека БГУ [Электронный ресурс]. – Режим доступа: http://elib.bsu.by. – Дата доступа: 06.03.2023.

3. Национальный статистический комитет Республики Беларусь [Электронный ресурс]. – Режим доступа: http:// www.belstat.gov.by. – Дата доступа: 25.09.2021.