ЛЕКЦИОННЫЙ КОМПЛЕКС

Дисциплина: Проектная деятельность и биостатистика

Код дисциплины PDB 2303

Название и шифр ОП: 6В10115 «Медицина»

Объем учебных часов/кредитов: 150/5

Курс и семестр изучения: 2/4

Объем лекций: 8 ч.

Шымкент, 2025 г.

Лекционный комплекс разработан в соответствии с рабочей учебной программой дисциплины (силлабусом) «Введение в научные исследования» и обсужден на заседании кафедр:

Протокол № 12 от « 28 » 05 2025 г.

Зав. кафедрой, к.ф.-м.н., асс. проф. _____ М.Б. Иванова

ЛЕКЦИЯ №1

1. Тема: Введение в биостатистику. Этапы статистического исследования.

2. Цель: формирование у обучающихся представления о дисциплине «Биостатистика»: ее предмете, задачах, методах и истории становления, а также знакомство обучающихся с последовательностью проведения статистического исследования.

3. Тезисы лекции:

Введение в биостатистику.

Биостатистика - это наука, которая занимается сбором, обработкой и анализом данных в медицине и биологии.

Биостатистика играет важную роль в медицине, так как позволяет проводить объективные исследования эффективности лечения, анализировать влияние различных факторов на здоровье, прогнозировать вероятность возникновения заболеваний и т.д. Она помогает врачам и исследователям принимать информированные решения на основе данных.

Задачи биостатистики:

- количественное представление фактов (измерение) – это выражение свойства отдельного биологического объекта в виде числа, варианты или значения переменной;
- обобщенное описание множества фактов (статистическое оценивание) – это расчет показателей и параметров, которые полноценно характеризуют свойства множества однотипных объектов или выборки;
- поиск закономерностей (проверка статистических гипотез) – это доказательство неслучайности отличий между сравниваемыми совокупностями, объектами, зависимости их характеристик от внешних или внутренних причин.

Методы биостатистики:

- сбор данных, который может быть пассивным (наблюдение) и активным (эксперимент);
- описательная статистика, которая занимается описанием и представлением данных;
- сравнительная статистика, которая позволяет проводить анализ данных в исследуемых группах и сравнение групп между собой с целью получения определенных выводов. Эти выводы могут быть сформулированы в виде гипотез или прогнозов;
- методы, позволяющие оценить взаимосвязи между признаками;
- методы, ориентированные на прогнозирование данных, основанные на изучении связей между явлениями или оценки их динамики.

История биостатистики началась в конце XIX века, когда статистические методы начали применяться в медицинских и биологических исследованиях.

Основоположником биостатистики является Френсис Гальтон, он впервые ввел в научный оборот термин «биометрия», разработал основы корреляционного анализа.

Последователем Ф. Гальтона стал английский статистик Карл Пирсон, который разработал многочисленные статистические тесты и методы.

В XX веке широкое применение нашли идеи и методы статистика Рональда Фишера, которые легли в основу биостатистики. В настоящее время биостатистика является неотъемлемой частью медицинских и биологических исследований и продолжает развиваться с развитием новых методов сбора и анализа данных.

Проведение статистического исследования начинается с определения *проблемы*, в соответствии с которой ставятся *цель* и *задачи* исследования, изучается *литература* по данной проблеме и разрабатывается *рабочая гипотеза*. Этот этап исследования называется *подготовительным*.

Проблемой в общественном здоровье и здравоохранении могут быть, например, низкий уровень здоровья населения или его групп, предположение о причине и факторах, влияющих

на здоровье населения или его групп, обнаружение недостатков в организации труда медицинских работников и т.д.

Рабочая гипотеза - основная идея относительно того, как решить стоящую перед исследователем проблему. Гипотезу исследователь предполагает проверить по полученным в процессе исследования эмпирическим данным.

Цель исследования – конечный результат, на достижение которого направлено исследование.

Задачи исследования – пошаговое достижение цели исследования. Задачи исследования отражают частные вопросы, которые необходимо последовательно решить, чтобы достигнуть конечной цели исследования.

Этапы статистического исследования представлены на рис. 1.1.



Рис. 1.1. Этапы статистического исследования

I этап - составление программы и плана статистического исследования.

II этап – организация и проведение сбора необходимых данных, предусмотренных программой исследования. Данные могут быть собраны путем проведения опросов, изучения медицинских записей или проведения наблюдений. Важно гарантировать достаточность и качество данных для проведения анализа.

III этап – осуществление обработки собранных данных (контроль, группировка, шифровка, вычисление статистических показателей, сводка в статистические таблицы). Для обработки данных могут быть использованы статистические программы.

IV этап – анализ и интерпретация результатов исследования. Для анализа данных могут быть использованы проверка гипотез, корреляционный анализ, регрессионный анализ и др. На основе анализа полученных результатов исследования формулируются выводы и предложения.

I этап статистического исследования - составление программы и плана статистического исследования.

Программа статистического исследования предусматривает решение следующих вопросов:

1) определение единицы наблюдения и составление программы сбора данных;
2) составление программы разработки данных;
3) составление программы анализа собранных данных.

Единица наблюдения – каждый первичный элемент статистической совокупности. Например, каждый студент, каждый родившийся, каждый пациент.

Единица наблюдения наделена признаками сходства и различия.

Признаки сходства – общие учетные признаки, свидетельствующие о принадлежности конкретной единицы наблюдения к этой совокупности.

Признаки различия – представляют собой индивидуальные особенности (характеристики) каждой единицы наблюдения, являются конечным объектом статистического исследования. Признаки различия подлежат изучению и регистрации, поэтому называются *учетными признаками*.

Учетные признаки – признаки, по которым различаются элементы единицы наблюдения в статистической совокупности.

Учетные признаки классифицируют по *характеру* и *роли* в совокупности.

По характеру признаки можно разделить на:

- *качественные (атрибутивные, описательные) признаки* – описываются словами. Среди качественных признаков выделяют номинальные и порядковые.

Номинальные – признаки, не поддающиеся непосредственному измерению. Состоят из взаимоисключающих категорий. Например, диагноз, пол, профессия, семейное положение. Номинальные данные, которые могут быть отнесены только к двум противоположным категориям «да» - «нет», принимающие одно из двух значений (выжил – умер, курит – не курит), называются *дихотомическими (бинарными).*

Порядковые – признаки, которые можно расположить в естественном порядке (ранжировать). Например, оценка тяжести состояния пациента, стадия болезни, самооценка состояния здоровья.

- *количественные признаки* – признаки, которые выражены числами. Среди количественных признаков выделяют:

Непрерывные – принимающие любое значение на непрерывной шкале. Например, масса тела, температура, биохимические показатели крови.

Дискретные – принимающие значения лишь из некоторого списка определенных чисел, обычно целых. Например, число рецидивов, число детей в семье, число заболеваний у одного больного, число выкуриваемых сигарет, число вызовов скорой помощи, число больных, поступающих в больницу.

Для учетных признаков используются различные виды шкал.

Шкала – необходимый, обязательный элемент измерительной процедуры. В медицинских исследованиях применяются следующие виды шкал:

- *номинальная* или *шкала наименований (nominal or scale of names)* используется для классификации свойств объекта, присвоения им числовых, буквенных и иных символьных характеристик (пол, национальность, цвет глаз, цвет волос, диагноз и т.д.);

- *порядковая* или *ранговая (ordinal or rank)*– упорядочивает значения признака (шкала стадий гипертонической болезни по Мясникову, шкала степеней сердечной недостаточности по Стражеско-Василенко-Лангу, шкала степени выраженности коронарной недостаточности по Фогельсону и др.);
- *интервальная (interval)* – показывает «размах» отдельных измерений признака (время, шкала температур, тестовые баллы);
- *шкала отношений (ratio scale)* – выявляет соотношение измеренных значений признака (рост, вес, время реакции, количество выполненных заданий теста).

По роли в совокупности выделяют:

- *факторные (независимые)* признаки, влияющие на изменение зависимых признаков;
- *результативные (зависимые)* признаки, изменяющие свое значение под влиянием

факторных признаков.

Например, количество выкуренных сигарет – факторный признак, вероятность возникновения заболевания легких и сердца – результативный признак.

Программа сбора данных представляет собой последовательное изложение учитываемых признаков – вопросов, на которые нужно получить ответы при проведении данного исследования. Программу сбора данных оформляют в виде *регистрационного документа* (анкеты, бланки, карты и др.), включающего признаки, которые исследователь хочет изучить в ходе эксперимента, и который заполняют на каждую единицу наблюдения.

Программа разработки полученных данных предусматривает составление макетов статистических таблиц.

Программа анализа предусматривает перечень статистических методик, необходимых для выявления закономерностей изучаемого явления.

План исследования включает в себя:

1. Выбор объекта исследования.
2. Определение объема статистической совокупности.
3. Сроки и место проведения исследования, виды и способы наблюдения и сбора данных.
4. Характеристика исполнителей (кадры).
5. Характеристика технического оснащения и требуемых материальных средств.

Объект статистического исследования – это совокупность, с которой будут собираться необходимые сведения. Это может быть население, студенты, пациенты, госпитализированные в стационар и т.п.

Статистическая совокупность – это группа, однородных по какому-либо признаку объектов, ограниченных пространством и временем (число жителей одного города; число больных онкозаболеваниями в данной стране и т.д.). Статистическая совокупность состоит из единиц наблюдения (рис. 1.2).



Рис. 1.2. Виды и структура статистической совокупности

Различают два вида совокупностей – *генеральная* и *выборочная*.

Генеральная совокупность – это группа, которая состоит из бесконечно большого числа объектов (все больные с данной патологией; все жители данной территории и т.д.).

Выборочная совокупность (выборка) – отобранная для исследования часть генеральной совокупности и предназначенная для характеристики всей генеральной совокупности. Выборка должна быть *репрезентативна* (представительна) по количеству и по качеству по отношению к генеральной совокупности.

Репрезентативность количественная основана на законе больших чисел и означает

достаточную численность элементов выборочной совокупности, рассчитываемую по специальным формулам.

Репрезентативность качественная означает соответствие признаков, характеризующих элементы выборочной совокупности по отношению к генеральной. Другими словами, внутренняя структура выборки по основным признакам (пол, возраст и др.) должна соответствовать генеральной совокупности.

Объем статистической совокупности (N, n) – это численность элементов совокупности, взятых для исследования.

Сроки и место проведения исследования – это составление календарного плана выполнения данного исследования на конкретной территории.

По времени регистрации различают два вида наблюдения – *текущее (или постоянное)* и *единовременное (или одномоментное)*.

Текущее наблюдение – вид наблюдения, при котором регистрация проводится постоянно по мере возникновения единиц наблюдения. Например, каждый случай рождения, смерти, обращения в медицинскую организацию.

Единовременное наблюдение – изучаемые явления фиксируются на какой-либо определенный момент (час, день недели, дату). Например, перепись населения, состав коечного фонда стационара.

Для исследователя важно определить способ проведения исследования: *сплошное наблюдение* или *несплошное (выборочное).*

Сплошное наблюдение – это регистрация всех единиц наблюдения, составляющих генеральную совокупность.

Несплошное (выборочное) наблюдение – изучение лишь части совокупности для характеристики целого.

Для проведения *выборочного статистического наблюдения* используют различные методы отбора единиц: *случайный, механический, гнездовой, направленный, типологический.*

- *случайный отбор* – отбор, проводимый по жребию (по начальной букве фамилии, по дню рождения и т.п.);
- *механический отбор* – отбор, когда у всей совокупности берется для изучения механически отобранная каждая пятая (20%) или десятая (10%) единица наблюдения;
- *гнездовой (серийный) отбор* – отбор, при котором из генеральной совокупности выбираются не отдельные единицы, а гнезда (серии), которые отбираются путем случайной или механической выборки. Например, из всех кафедр медицинской академии случайным образом было выбрано 10 кафедр для участия в опросе и затем опрошены все преподаватели этих кафедр. Кафедра в данном случае является «гнездом».
- *направленный отбор* - это отбор, при котором из генеральной совокупности отбираются только те единицы наблюдения, которые позволят установить влияние неизвестных факторов при установлении влияния известных. Например, при изучении влияния стажа работы на травматизм отбираются рабочие одной профессии, одного возраста, одного цеха, одного образовательного уровня.
- *типологический отбор* – это отбор единиц из заранее сгруппированных однотипных качественных групп. Например, при изучении закономерностей заболеваемости среди городского населения следует вначале разделить изучаемое население по возрастной структуре. После чего производится случайный отбор в каждой возрастной группе.

Характеристика исполнителей (кадры) - сколько человек и какой квалификации проводят исследование.

Характеристика технического оснащения и требуемых материальных средств – лабораторное оборудование и приборы, соответствующие цели исследования, канцелярские товары (бумага, бланки) и т.д.

II этап статистического исследования - организация и проведение сбора необходимых данных.

Сбор данных – это процесс регистрации, заполнения официально существующих или специально разработанных учетных документов (талоны, карты и т.п.).

Способы сбора данных статистического исследования:

- *непосредственный сбор данных* осуществляется исследователями, которые сами производят регистрацию признаков и фактов путем подсчета, измерения, взвешивания и т.п., а затем вносят данные в формуляры статистического наблюдения;
- *документальный способ сбора данных* предполагает получение статистических сведений из документов, отображающих изучаемые объекты, из отчетно-учетной документации (история болезни, история развития ребенка, больничный лист и т.п.);
- *опрос* - это способ сбора данных, при котором исследователь получает необходимые сведения о каждой единице наблюдения со слов спрашиваемого (устный опрос, анкета).

III этап статистического исследования - обработка собранных данных.

Этот этап статистического исследования включает в себя следующие действия:

1) контроль собранных данных;
2) шифровка;
3) группировка;
4) сводка данных;
5) вычисление статистических показателей и статистическая обработка данных.

1. *Контроль* – это проверка собранного материала с целью отбора учетных документов, имеющих дефекты, для их последующего исправления, дополнения или исключения из исследования. Например, в анкете не указан пол, возраст или нет ответов на другие поставленные вопросы.

2. *Шифровка* – это применение условных обозначений выделяемых признаков. Например, пол: муж. – М, жен. – Ж; курит – 1, не курит – 0.

3. *Группировка данных* – это распределение собранного материала по качественным и количественным признакам. Например, группировка обучающихся по курсам обучения, по полу, по факультетам.

4. *Сводка данных* – занесение полученных после подсчета цифровых данных в таблицы. В таблице различают *подлежащее* и *сказуемое*.

Подлежащее – это то, о чем говорится в таблице (признаки, являющиеся предметом исследования), обычно размещается в левой части таблицы по вертикали.

Сказуемое - это то, что характеризует подлежащее, размещается по горизонтали.

Статистические таблицы подразделяются на *простые* (табл. 1.1), *групповые* (табл. 1.2), *комбинационные* (табл. 1.3).

Таблица 1.1.

Распределение курящих обучающихся по факультетам

Название факультета	Всего обучающихся	
	Абсолютное число обучающихся	%
Медицина		
Фармация		
Итого		100

Таблица 1.2.

Распределение курящих обучающихся различных факультетов по полу и по возрасту, в котором они выкурили первую сигарету

Название факультета	Пол	Возраст, в котором выкурили первую сигарету	Всего

	м	ж	до 15 лет	15-18	старше 18 лет	
Медицина						
Фармация						
Итого						

Таблица 1.3.

Распределение курящих обучающихся различных факультетов по полу и среднему количеству сигарет, выкуриваемых за день

Название факультета	Среднее количество сигарет, выкуриваемых обучающимися в день									Всего		
	10 и менее до 15 лет			11-20			более 20					
	м	ж	оба пола	м	ж	оба пола	м	ж	оба пола	м	ж	оба пола
Медицина												
Фармация												
Итого												

4. Иллюстративный материал: презентация, слайды.

5. Литература:

- Основная:

1. Койчубеков Б. К. Биостатистика. уч. пособие / Б.К. Койчубеков. - Алматы: Эверо, 2016.
2. Койчубеков Б.К. Биостатистика: учебное пособие. - Эверо, 2014.
3. Чудиновских В.Р. Абдикадыр Ж.Н. Применение программ EXCEL и SPSS statistics для статистического анализа медико-биологических данных. Уч. пособие.- ИП "АҚНҰР", 2021

- Дополнительная:

1. Койчубеков Б.К. Биостатистика. Монография.- Алматы: ТОО Эверо, 2024.- 152с.
2. Бухарбаев М. А. Медицинская статистика: учебное пособие / М. А. Бухарбаев, В. Н. Казагачев. - 2-е изд. - Алматы: Эпиграф, 2022. - 268 с.

- Электронные ресурсы:

1. Койчубеков Б.К., Сорокина М.А., Букеева А.С., Такуадина А.И. БИОСТАТИСТИКА в примерах и задачах: Учебно-метод. пособие/– Алматы ТОО «Эверо», 2020. – 80 с. https://elib.kz/ru/search/read_book/870/
2. Койчубеков Б.К. Биостатистика: Учебное пособие: Издательство «Эверо», Алматы, 2020, 154с. https://elib.kz/ru/search/read_book/867/
3. Биостатистика [Электронный ресурс]: учебник /К.Ж. Кудабаев [и др.].- Электрон. текстовые дан. (85,7Мб).- Шымкент: ЮКГФА, 2015. – 187с. эл. опт. диск (CD-ROM)
4. В.Р. Чудиновских, Ж.Н. Абдикадыр. Применение компьютерных программ для проверки статистических гипотез в медико-биологических исследованиях: учебное пособие. – Караганда: ИП «Издательство АҚНҰР».-2016, 100 с. https://aknurpress.kz/reader/web/1344
5. В.Р. Чудиновских, Ж.Н. Абдикадыр, А.Ш. Каипова, А.У. Алтаева. Применение программ EXCEL и SPSS Statistics для статистического анализа медико-биологических данных: уч.пос.– Караганда: ИП «Издательство АҚНҰР».– 2016, 128с. https://aknurpress.kz/reader/web/1342

6. Контрольные вопросы:

1. Перечислите задачи биостатистики и ее методы.
2. Проведите краткий исторический обзор становления биостатистики как науки.
3. Укажите последовательность (этапы) проведения статистического исследования.
4. Перечислите составные элементы программы статистического исследования.
5. Укажите, что включает в себя план статистического исследования.

6. Дайте определение статистической совокупности.
7. Дайте определение единицы наблюдения и приведите классификацию ее учетных признаков.
8. Перечислите виды шкал, применяемых при проведении медицинских исследований. Приведите примеры.
9. Какие требования предъявляются к выборочной совокупности?
10. В чем заключается процесс сбора данных?
11. Какие действия включает в себя этап обработки полученных данных?
12. Что такое группировка данных?

ЛЕКЦИЯ №2

1. Тема: Описательная статистика.

2. Цель: формирование у обучающихся представления о методах описательной статистики для оценки и анализа статистической совокупности при изучении общественного здоровья и деятельности медицинских организаций.

3. Тезисы лекции:

Описательная статистика или *дескриптивная статистика* (от англ. *Descriptive statistics*) – раздел статистики, который занимается обработкой эмпирических данных, их систематизацией, наглядным представлением в форме графиков и таблиц, а также их количественным описанием посредством основных статистических показателей.

Первым шагом систематизации материалов статистического наблюдения является определение *статистического распределения выборки*.

Статистическое распределение выборки (или *вариационный ряд*, или *частотное распределение*) – ряд числовых измерений какого-либо признака, отличающихся друг от друга по своей величине и расположенных в определенном порядке (возрастания или убывания).

Вариантой (x) называется каждое числовое значение в вариационном ряду.

Частота варианты (*v*) – это количество элементов совокупности, имеющих одинаковое числовое значение. Общее число вариант в вариационном ряду обозначают *n*.

Виды вариационных рядов (рис. 2.1):

1. В зависимости от вида величины:

• *дискретный* – содержит варианты, представленные только целыми значениями чисел (например, число рецидивов, число детей в семье, число вызовов бригад скорой помощи);

• *непрерывный* – может содержать любое значение на непрерывной шкале измерения признака (например, масса тела, рост, температура, биохимические показатели крови).

2. В зависимости от частоты, с которой каждая варианта встречается в вариационном ряду:

• *простой* – это ряд, в котором каждая варианта встречается по одному разу (все числа разные);

• *взвешенный* – это ряд, в котором каждая варианта встречается неоднократно (с разной частотой).

3. В зависимости от группировки вариант:

• *несгруппированный* – содержит все значения отдельных вариант;

• *сгруппированный (интервальный)* - представлен интервалами значений вариант и частотой вариант, входящих в каждый из них. Как правило, интервальный ряд применяется при большом числе наблюдений.



Рис. 2.1. Виды вариационных рядов

Пример 2.1. Были зафиксированы значения частоты сердечных сокращений (ЧСС) у 20 пациентов с тахикардией: 100, 100, 100, 112, 112, 112, 112, 112, 120, 120, 120, 120, 120, 120, 124, 124, 124, 124, 128, 128.

Это дискретный, взвешенный, не сгруппированный вариационный ряд.

Этот ряд можно представить в виде таблицы (таблица 2.1) и изобразить графически с помощью *многоугольника распределения* или *полигона частот* (рис. 2.2).

Таблица 2.1.

Варианты (x_i)	Частоты (v_i)
100	3
112	5
120	6
124	4
128	2
Всего:	n=20

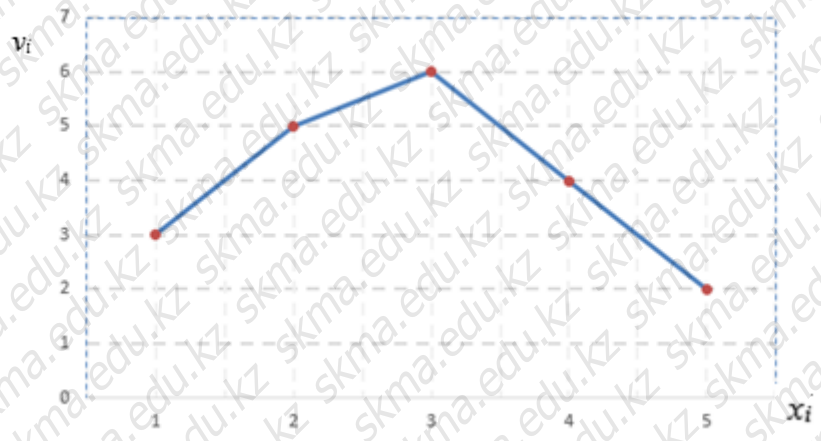


Рис. 2.2. Полигон

Пример 2.2. Для выборочной совокупности, состоящей из 100 мужчин в возрасте 20-25 лет, был определен рост каждой единицы наблюдения.

Согласно результатам измерений построен непрерывный, взвешенный, сгруппированный вариационный ряд (таблица 2.2):

Таблица 2.2.

Варианты (x_i), см	Частоты (v_i)

150-155	1
155-160	11
160-165	14
165-170	26
170-175	26
175-180	13
180-185	8
185-190	1
Всего:	100

Число интервалов для сгруппированных вариационных рядов определяется по формуле Стерджеса: $k=1+3{,}322\cdot\lg n$, (2.1)
где n – объем выборки.

Для вычисления ширины интервала используется формула:

$$h=\frac{x_{\max}-x_{\min}}{1+3{,}322\lg n}, \quad (2.2)$$

где x_{max}, x_{min} – наибольшее и наименьшее значения вариант соответственно.

За начало первого интервала берётся величина: $x_{нач}=x_{min}-0{,}5\cdot h$. (2.3)

Сгруппированный вариационный ряд представляется графически в виде ступенчатой фигуры, называемой *гистограммой* (рис. 2.3).

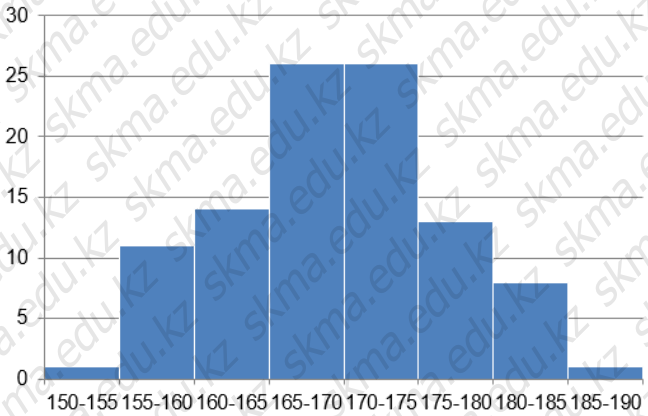


Рис. 2.3. Гистограмма

Визуально оценить форму и размах распределения данных можно также с помощью графика *«стебель с листьями» (Stem and Leaf Plot).*

Пример 2.3. Имеются данные о возрасте пациентов врача-кардиолога, ведущего прием в больнице г. Шымкент, за 2023 год. С помощью графика «стебель с листьями» определить пациенты какого возраста наиболее часто и наиболее редко к нему обращаются.

30	34	35	37	37	38	38	38	38	39	39	40	40	42	42
43	43	43	43	43	43	44	44	44	44	44	44	44	45	45
45	46	46	46	46	46	46	47	47	47	47	47	47	48	48
48	48	48	48	48	49	49	49	49	49	49	49	50	50	50
50	50	50	50	50	51	51	51	51	52	52	52	52	52	52
53	53	53	53	53	53	53	53	53	53	53	53	53	53	53
53	53	54	54	54	54	54	54	54	54	54	54	54	55	55
55	56	56	56	56	56	56	57	57	57	57	57	57	57	58
58	59	59	59	59	59	59	60	60	60	60	61	61	61	61

61	61	61	61	61	61	61	62	62	62	62	62	62	62	63
63	64	64	64	64	64	64	65	65	66	66	66	66	66	66
67	68	68	68	69	69	69	70	71	71	71	71	71	71	71
72	73	75	76	77	78	78	78	82						

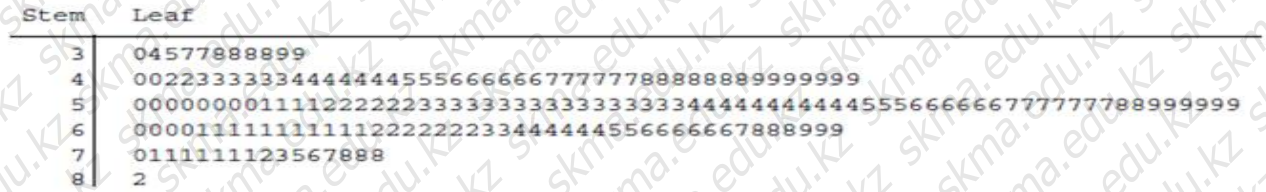


Рис. 2.4. «Стебель с листьями» (Stem and Leaf Plot)

Из графика видно, что наиболее часто обращаются пациенты в возрасте от 50 до 59 лет. Наиболее редко обращаются пациенты в возрасте старше 80 лет. Количество обращений пациентов в возрасте от 40 до 49 и от 60 до 69 почти одинаково.

После того как вариационный ряд построен, приступают к его обработке. Она заключается в нахождении *показателей центральной тенденции* и *показателей изменчивости (разнообразия).*

К показателям центральной тенденции относятся *средние* и *структурные* величины.

Средние величины в медицине и здравоохранении могут быть использованы:

1) для оценки состояния здоровья – например, параметров физического развития (средний рост, средний вес, средний объем жизненной емкости легких и др.), соматических показателей (средний уровень свертываемости сахара в крови, средний пульс, средняя скорость оседания эритроцитов (СОЭ) и др.);

2) для оценки организации работы медицинских организаций, а также деятельности отдельных врачей и средних медицинских работников (средняя длительность пребывания больного на койке, среднее число посещений за 1 час приема в поликлинике и др.).

В зависимости от характера задачи пользуются тем или иным видом средних величин. В данном курсе биостатистики будет рассматриваться только *среднее арифметическое* (*Average, Mean*), т.к. степенные средние (среднее гармоническое, среднее квадратическое, среднее геометрическое) в медицинских исследованиях применяются редко.

• *Среднее арифметическое простое*

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \qquad (2.4)$$

где *n* - общее число членов ряда (объем выборки);

• *Среднее арифметическое взвешенное*

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i v_i}{\sum_{i=1}^{n} v_i}, \qquad (2.5)$$

где v_i – частоты;

Среднее арифметическое взвешенное применяется при расчетах в сгруппированных вариационных рядах, когда ряд разбит на отдельные интервалы и есть данные о частоте каждого из них, но не представлены значения отдельных вариант. В этом случае за варианту принимается середина каждого интервала.

Пример 2.4. Определить среднюю массу тела мужчин, принимающих участие в медицинском исследовании. Данные представлены в виде сгруппированного вариационного

ряда (таблица 2.3).

Таблица 2.3.

Масса тела обследованных мужчин (кг)	Количество обследованных (v_i)	Середина интервала (x_i)	x_i*v_i
66-70,9	11	68,5	753,5
71-75,9	18	73,5	1323
76-80,9	24	78,5	1884
81-85,9	14	83,5	1169
Всего	67		5129,5

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i * v_i}{\sum_{i=1}^{n} v_i} = \frac{5129,5}{67} = 76,6 \text{ кг.}$$

Структурные величины.

• *Мода (Mode, Mo)* – это такое значение признака, которое встречается наиболее часто. Если все значения в вариационном ряду встречаются одинаковое количество раз, то такой ряд *не имеет моды*. Если два значения вариационного ряда имеют одинаковую частоту и она больше частоты любого другого значения, то такой вариационный ряд имеет *две моды* (*бимодальный*).

• *Медиана (Median, Me)* – варианта, находящаяся в середине упорядоченного вариационного ряда. При нахождении медианы следует различать два случая:

1. если объем совокупности *n* – *нечетное число* и варианты упорядочены (записаны от наименьшего к наибольшему), то *медианой* будет варианта, занимающая в ряду центральное положение. Ее порядковый номер может быть найден по формуле (*n*+1)/2, где *n* – объем выборки;

2. если объем совокупности *n* – *четное число*, то медиана равна полусумме вариант, находящихся в середине упорядоченного вариационного ряда:

$$Me = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}. \quad (2.6)$$

В медицинских исследованиях из средних величин наиболее часто используется среднее арифметическое. *Однако, если значения признаков имеют распределение отличное от нормального, то для характеристики центральной тенденции более обоснованным является использование медианы, а не среднего арифметического!*

• *Квантили* - отдельные равные части, на которые разбивается вариационный ряд (рис. 2.5):

- *квартили (Quartilie)*– величины, делящие вариационный ряд на четыре равные части;
- *квинтили* (quintile) - величины, делящие вариационный ряд на пять равных частей;
- *децили (Decile)* - величины, делящие вариационный ряд на десять равных частей;
- *процентили (Percentile)*- величины, делящие вариационный ряд на сто равных частей

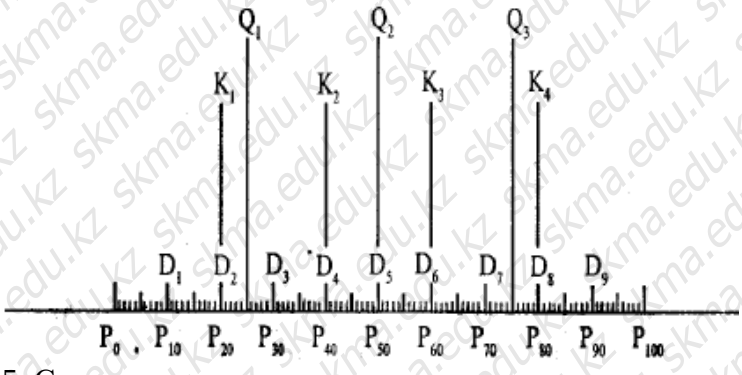


Рис. 2.5. Структурные характеристики вариационного ряда

Наиболее часто в статистике используются квартили.

Первый или *нижний квартиль* (Q_1) или *25-ый процентиль* (P_{25}) - это значение случайной величины, ниже которого находится 25% выборки.

Второй квартиль (Q_2) или *50-ый процентиль* (P_{50}) всегда равен медиане.

Третий или *верхний квартиль* (Q_3) или *75-ый процентиль* (P_{75}) - это значение случайной величины, выше которого находится 25% выборки (рис. 2.6).

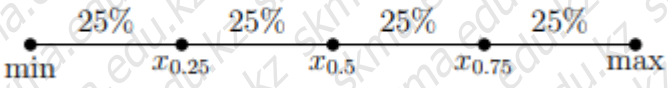


Рис. 2.6. Квартили

Для расчета квартилей надо поделить вариационный ряд медианой на две равные части. Если число вариант четное, то ряд делим пополам. Если нечетное, то делим ряд на две части причем медиана входит в каждую часть. Затем необходимо найти середину ряда для каждой половины. Полученные числа будут являться соответственно верхним и нижним квартилем.

Пример 2.5. Имеются данные о длительности заболевания в годах у отдельных пациентов: 1, 7, 5, 19, 6, 12, 10, 20, 15. Определить квартили.

Упорядочим вариационный ряд: 1, 5, 6, 7, 10, 12, 15, 19, 20.

Количество наблюдений - 9 – нечетное число. Медиана ряда – число с порядковым номером (n+1)/2=10/2=5. *Me*=10.

Для определения квартилей делим ряд на две половины: (1, 5, 6, 7, 10) и (10, 12, 15, 19, 20). Медиана 10 вошла в каждую часть. Далее находим медианы для каждой половины:

(1, 5, 6, 7, 10) – число вариант нечетное, медиана – число с порядковым номером (n+1)/2=6/2=3. *Me*=6. Таким образом, верхний квартиль Q_1=6.

(10, 12, 15, 19, 20) - число вариант нечетное, медиана – число с порядковым номером (n+1)/2=6/2=3. *Me*=15. Таким образом, нижний квартиль Q_3=15.

Показатели изменчивости (разнообразия)

К показателям изменчивости относятся: *размах, межквартильный размах, дисперсия, среднее квадратическое отклонение, коэффициент вариации,.*

• *Размах вариационного ряда (Range, R)* – разница между наибольшим и наименьшим значением вариант в выборке:

$$R = x_{max} - x_{min}, \tag{2.7}$$

• *Межквартильный размах (Interquartile range, IQR)* – разница между третим и первым квартилем:

$$IQR=Q_3-Q_1, \quad (2.8)$$

- *Дисперсия* (*Variance, S^2*) – это средний квадрат отклонений индивидуальных значений признака от его средней величины (безразмерная величина).

- *простая дисперсия:*

$$S^2 = \frac{\sum_{n=1}^{n}(x_i-\bar{x})^2}{n-1}. \quad (2.9)$$

- *взвешенная дисперсия*:

$$S^2 = \frac{\sum_{n=1}^{n}(x_i-\bar{x})^2 \cdot v_i}{n-1}. \quad (2.10)$$

В том случае если дисперсия рассчитывается для генеральной совокупности в знаменателе дроби вместо (n-1) нужно ставить n!

- *Среднее квадратическое отклонение* (*Standart Deviation, S или σ*) - мера разброса случайной величины от ее среднего значения, выраженная в тех же единицах, что и варианты, определяется как квадратный корень из дисперсии: $S=\sqrt{S^2}$; (2.11)

Для характеристики любой совокупности, имеющей *нормальное распределение* достаточно знать два параметра: *среднее арифметическое* и *среднее квадратическое отклонение*. Записывается такая характеристика следующим образом $\bar{x} \pm s$ (или $M \pm \sigma$).

- *Коэффициент вариации (Coefficient of Variation, V)* - мера разброса случайной величины, выраженная в процентах, определяется как отношение среднего квадратического отклонения к средней величине признака:

$$V = \frac{S}{\bar{x}} \cdot 100\%. \quad (2.12)$$

Чем ближе коэффициент вариации к нулю, тем меньше вариация значений признака. Чем больше коэффициент вариации, тем более изменчив признак.

Совокупность считается однородной, если коэффициент вариации не превышает 33%.

При *V<10%* разнообразие ряда считается слабым, при *10%≤V≤20%* – средним, при *V>20%* - сильным.

Пример 2.6. В городе Шымкент в 2023 г. было проведено измерение массы тела 7-летних мальчиков (данные представлены в таблице 2.4). По данным аналогичного исследования, выполненного в этом же городе, но в 2013 г., средняя масса тела 7-летних мальчиков составляла 23,8 кг, s ±3,6 кг.

1) Вычислить среднее арифметическое и показатели разнообразия вариационного ряда (дисперсию, среднее квадратическое отклонение, коэффициент вариации).

2) Оценить полученные результаты, сравнить их вариабельность с данными предыдущего исследования, сделать соответствующие выводы.

Таблица 2.4.

Результаты измерения массы тела 7-летних мальчиков города Шымкент в 2023 г.

Масса тела	Середина интервала (x_i)	Кол-во мальчиков (v_i)	$x_i \cdot v_i$	$x_i-\bar{x}$	$(x_i-\bar{x})^2$	$(x_i-\bar{x})^2 \cdot v_i$
15-18,9	17	16	272	-7	19	784
19-22,9	21	27	567	-3	9	243
23-26,9	25	32	800	1	1	32
27-30,9	29	16	464	5	25	400
31-34,9	33	9	297	9	81	729

Всего		100	2400			2188

Решение.

1) $\bar{x} = \frac{\sum_{i=1}^{n} x_i * v_i}{\sum_{i=1}^{n} v_i} = \frac{2400}{100} = 24$ кг.

2) $S^2 = \frac{\sum_{n=1}^{n} (x_i - \bar{x})^2 \cdot v_i}{n-1} = \frac{2188}{100-1} = 22{,}1.$

3) $s = \sqrt{S^2} = \sqrt{22{,}1} = \pm 4{,}7$ кг.

4) $V = \frac{s}{\bar{x}} \cdot 100\% = \frac{4{,}7}{24} = 19{,}6\%$.

Выводы:

1. Средняя масса тела 7-летних мальчиков в городе Шымкент составляет 24 кг.

2. Дисперсия равна 22,1, среднее квадратическое отклонение ±4,7 кг, коэффициент вариации 19,6%.

3. Величина коэффициента вариации, равная 19,6% свидетельствует о среднем разнообразии признака.

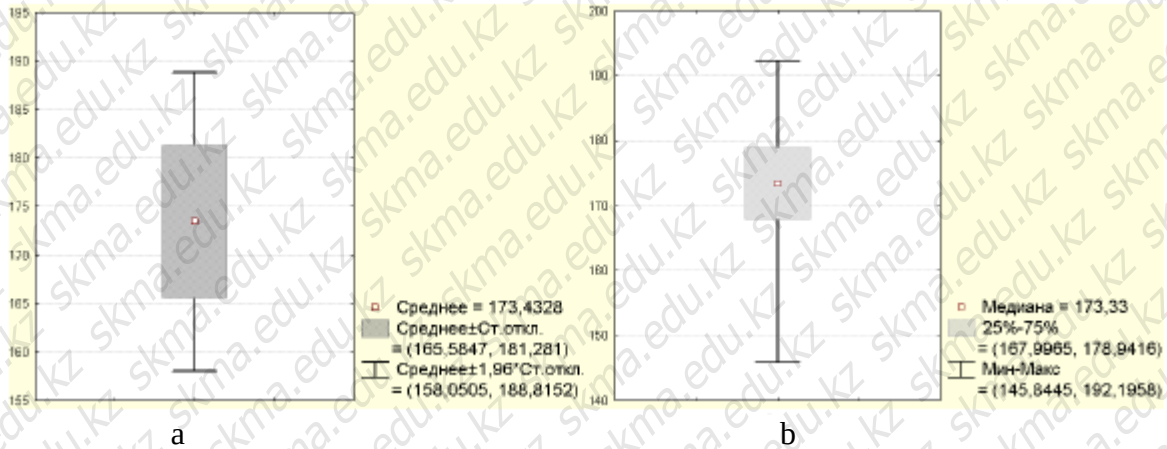
Таким образом, можно считать, что полученная средняя величина массы тела является достаточно типичной. По сравнению с 2013 г. в 2023 г. отмечается более значительная вариабельность массы тела у мальчиков 7 лет (±4,7 кг против ±3,6 кг). Аналогичный вывод вытекает их сопоставления коэффициентов вариации (*V* в 2013 г. (3,6/23,8·100%)=15,1%).

Для иллюстрации основных показателей описательной статистики часто используется график *«ящик с усами» (Box and Whisker plot).*

Например, для выборочных данных о росте 30-летних женщин были построены графики «ящики с усами» (рис. 2.7 a, b).

При анализе таких графиков обязательно нужно обращать внимание на «легенду», т.е. условные обозначения, которые приводятся в нижней части графика.

На первом графике (рис. 2.7, a) приведены среднее, минимальное и максимальное значения, а также стандартное отклонение. На втором графике (рис. 2.7, b) приведены значения медианы, 25-го и 75-го процентилей.



a b

Рис. 2.7. Отображение статистических показателей на графике «ящик с усами»

4. Иллюстративный материал: презентация, слайды.

5. Литература:

● Основная:

1. Койчубеков Б. К. Биостатистика. уч. пособие / Б.К. Койчубеков. - Алматы: Эверо, 2016.
2. Койчубеков Б.К. Биостатистика: учебное пособие. - Эверо, 2014.

3. Чудиновских В.Р. Абдикадыр Ж.Н. Применение программ EXCEL и SPSS statistics для статистического анализа медико-биологических данных. Уч. пособие.- ИП "АҚНҰР", 2021

• Дополнительная:

1. Койчубеков Б.К. Биостатистика. Монография.- Алматы: ТОО Эверо, 2024.- 152с.
2. Бухарбаев М. А. Медицинская статистика: учебное пособие / М. А. Бухарбаев, В. Н. Казагачев. - 2-е изд. - Алматы: Эпиграф, 2022. - 268 с.

• Электронные ресурсы:

1. Койчубеков Б.К., Сорокина М.А., Букеева А.С., Такуадина А.И. БИОСТАТИСТИКА в примерах и задачах: Учебно-метод. пособие/– Алматы ТОО «Эверо», 2020. – 80 с. https://elib.kz/ru/search/read_book/870/
2. Койчубеков Б.К. Биостатистика: Учебное пособие: Издательство «Эверо», Алматы, 2020, 154с. https://elib.kz/ru/search/read_book/867/
3. Биостатистика [Электронный ресурс]: учебник /К.Ж. Кудабаев [и др.].- Электрон. текстовые дан. (85,7Мб).- Шымкент: ЮКГФА, 2015. – 187с. эл. опт. диск (CD-ROM)
4. В.Р. Чудиновских, Ж.Н. Абдикадыр. Применение компьютерных программ для проверки статистических гипотез в медико-биологических исследованиях: учебное пособие. – Караганда: ИП «Издательство АҚНҰР».-2016, 100 с. https://aknurpress.kz/reader/web/1344
5. В.Р. Чудиновских, Ж.Н. Абдикадыр, А.Ш. Каипова, А.У. Алтаева. Применение программ EXCEL и SPSS Statistics для статистического анализа медико-биологических данных: уч.пос.– Караганда: ИП «Издательство АҚНҰР».– 2016, 128с. https://aknurpress.kz/reader/web/1342

6. Контрольные вопросы:

1. Для чего используется описательная статистика?
2. Что такое вариационный ряд?
3. Какие виды вариационных рядов вы знаете?
4. С помощью каких графических инструментов можно представить вариационный ряд?
5. Для чего используются средние величины?
6. Какие виды средних величин вы знаете?
7. В чем различие между простыми и взвешенными величинами?
8. Как определить моду, медиану и квартили вариационного ряда?
9. По каким критериям можно оценить разнообразие признака?
10. Каково назначение дисперсии?
11. Каково назначение среднего квадратического отклонения?
12. Как интерпретируется значение коэффициента вариации?
13. Какие статистические показатели могут быть использованы при построении графика «ящик с усами»?

ЛЕКЦИЯ №3

1. Тема: Нормальное распределение. Основы теории проверки статистических гипотез. Критерии согласия.

2. Цель: формирование у обучающихся представления о нормальном распределении и знакомство с основными понятиями теории проверки статистических гипотез.

3. Тезисы лекции:

Нормальное распределение.

Для того, чтобы правильно выбрать статистический метод для анализа изучаемого признака (переменной) необходимо знать его *закон распределения*.

Закон распределения случайной величины – это соответствие, устанавливаемое между всеми возможными числовыми значениями случайной величины и вероятностями (частотами) их появления в совокупности.

Наиболее часто применяемыми на практике являются следующие виды законов распределения: биномиальный, Пуассона (*для дискретных случайных величин*); равномерный, экспоненциальный, нормальный (*для непрерывных случайных величин*).

При статистическом анализе медицинских данных наибольший интерес представляет *нормальное распределение*, так как многие биологические и медицинские показатели (рост, вес, уровень холестерина, артериальное давление, температура, показатели крови и др.), имеют законы распределения, близкие к *нормальному*.

Нормальное (или Гауссово, или колоколообразное (bell-shaped)) распределение (рис. 3.1) характеризуется тем, что наибольшее число наблюдений имеет значение, близкое к среднему, и чем больше значения отличаются от среднего, тем меньше таких наблюдений.

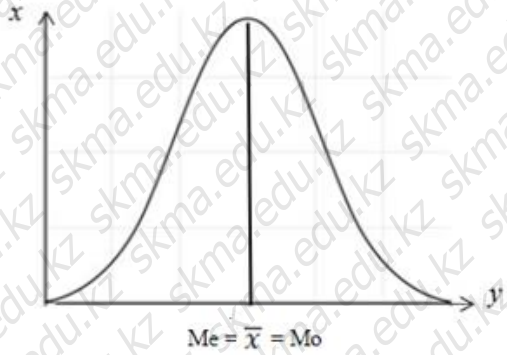


Рис. 3.1. Нормальное распределение переменной

На рисунке 3.1 по оси *y* обозначены значения, которые принимает признак, по оси *x* – частота встречаемости значений признака. Чем чаще встречаются данные значения, тем выше кривая. При нормальном распределении наибольшая частота встречаемости приходится на область средних значений признака.

Случайная величина (признак, переменная) *x*, подчиняющаяся нормальному распределению имеет функцию плотности вероятности вида:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad (3.1)$$

где μ – математическое ожидание ($\overline{x}$ - выборочное среднее является оценкой математического ожидания);

σ – среднее квадратическое отклонение (*s* – выборочное среднее квадратическое отклонение является оценкой σ).

Т.е. колоколообразная кривая описывается функцией (3.1).

Кривая нормального распределения обладает следующими свойствами:

- имеет колоколообразную форму, а значит симметрична относительно среднего значения;
- выборочное среднее, мода и медиана равны и соответствуют вершине распределения;
- асимптотически (бесконечно близко) приближается к оси абсцисс;
- площадь под кривой нормального распределения принимается равной 1 (или 100%);
- форма кривой зависит от двух параметров $\overline{x}$ и s;
- правило «трех сигм» (рис. 3.2):

  68,2% всех значений нормально распределенной случайной величины лежат в интервале $(\overline{x} - \sigma; \overline{x} + \sigma)$;

  95,4% всех значений нормально распределенной случайной величины лежат в интервале $(\overline{x} - 2\sigma; \overline{x} + 2\sigma)$;

99,6% % всех значений нормально распределенной случайной величины лежат в интервале $(\bar{x} - 3\sigma;\ \bar{x} + 3\sigma)$.

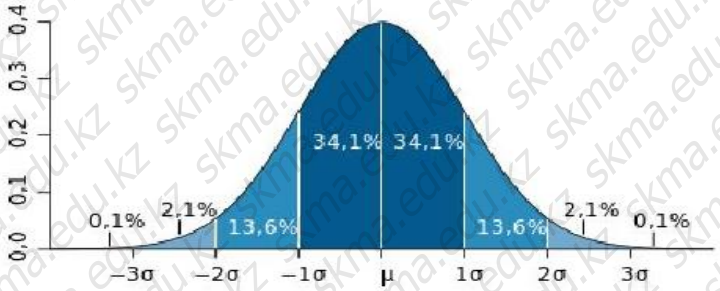


Рис. 3.2. Правило «трех сигм»

В биологии и медицине также нередко встречается *асимметричное распределение*.

В отличие от нормального этот вид распределения не имеет симметричности. Распределение может быть «растянуто» как влево, так и вправо.

Коэффициент асимметрии (Skewness, As) – показатель, который характеризует скошенность распределения:

$$As = \frac{\mu_3}{s^3}, \qquad (3.2)$$

Где $\mu_3 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3 \cdot v_i}{n}$.

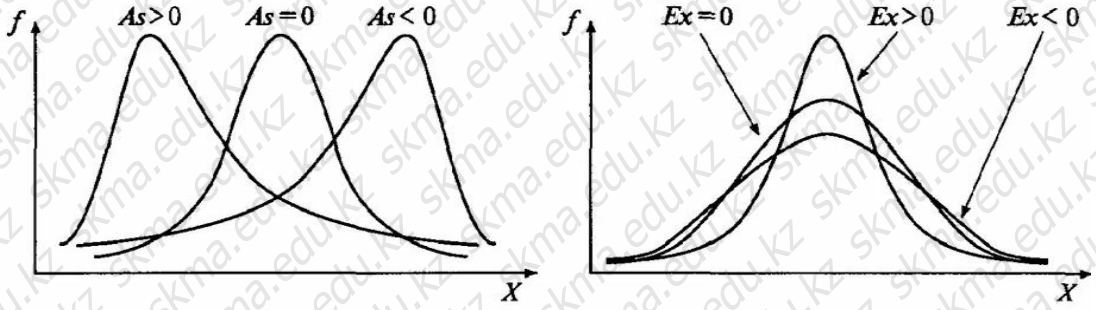
Если *As*>0, то распределение «растянуто» влево, а если *As*<0 – то вправо. У выборки, имеющей нормальное распределение, коэффициент асимметрии равен или очень близок к нулю (рис. 3.3. a).

Коэффициент эксцесса (Kurtosis, Ex) – показатель, который характеризует остроту пика распределения:

$$Ex = \frac{\mu_4}{\sigma^4} - 3, \qquad (3.3)$$

где $\mu_4 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4 \cdot v_i}{n}$.

Если *Ex*>0, то распределение имеет более острый пик, чем у нормального распределения, если *Ex*<0, то распределение имеет более «пологий» пик, чем у нормального распределения. У выборки, имеющей нормальное распределение, коэффициент эксцесса близок к нулю (рис. 3.3. b).



a b

Рис. 3.3. Графическая интерпретация значений коэффициента асимметрии и эксцесса

Основные понятия и определения теории проверки статистических гипотез.

В при проведении научных исследований в области медицины, здравоохранения и фармации часто требуется по наблюдениям выборки высказать некоторое суждение (*гипотезу*) относительно интересующих экспериментатора характеристик генеральной

совокупности, из которой эта выборка извлечена. То есть, речь идет о *проверке статистических гипотез*.

Теория проверки статистических гипотез является основным инструментом доказательной медицины.

Статистическая гипотеза – это некоторое предположение о числовых параметрах известного распределения (*среднем, дисперсии, среднем квадратическом отклонении*) или о виде закона распределения случайной величины.

Всегда выдвигаются две гипотезы: *нулевая (H_0) и альтернативная (H_1).*

Нулевая гипотеза H_0 (основная) - гипотеза об отсутствии различий между группами, либо об определенных значениях параметров, либо о соответствии распределения случайной величины определенному закону (например, нормальному).

Альтернативная гипотеза H_1 (конкурирующая) - гипотеза о существовании различий между группами, либо об отличающихся от заданных значений параметров, либо о несоответствии распределения определенному закону.

Нулевая гипотеза формулируется таким образом, чтобы она была противоположной той исследовательской гипотезе, которая послужила поводом для проведения исследования. Например, нулевая гипотеза при сравнении признаков в двух группах всегда будет утверждать, что различий между ними нет, а альтернативная, что различия есть.

В результате проверки нулевая гипотеза либо принимается, либо отвергается в пользу альтернативной. При этом есть риск допустить ошибки двух типов (рис. 3.4).

	H_0 принимается	**H_0 отвергается**
H_0 верна	Верное решение	**Ошибка I рода (α)**
H_0 ложна	**Ошибка II рода (β)**	Верное решение

Рис. 3.4. Ошибки первого и второго рода

Ошибка первого рода состоит в том, что гипотеза H_0 будет отвергнута, хотя на самом деле она правильная. Вероятность допустить такую ошибку называется *уровнем значимости* (α).

Для медицинских и фармацевтических исследований принимают уровень значимости α=0,05.

Ошибка второго рода состоит в том, что гипотеза H_0 будет принята, но на самом деле она неправильная (ложная). Вероятность допустить такую ошибку называется *уровнем доверительной вероятности* (β).

Значение 1-β называют *мощность юкритерия* – это вероятность отвержения неправильной гипотезы.

Например, пациент болен, но анализ крови этого не показал (*ложноотрицательный* результат). Врач не назначил лечение пациенту и допустил ошибку I рода.

Например, пациент здоров, но анализ крови показал наличие заболевания (*ложноположительный* результат). Врач назначил пациенту лечение и допустил ошибку II рода.

Для проверки нулевой гипотезы применяют статистические методы (тесты или критерии).

Статистическим критерием (тестом) называется математическое правило, согласно которому выясняется, соответствует или нет интересующая исследователя гипотеза опытным (эмпирическим) данным.

Статистика – это функция от выборочных наблюдений, на основе которой принимается или отвергается нулевая гипотеза.

Наблюдаемое (эмпирическое, расчетное) значение критерия – это значение, которое рассчитано по выборочной совокупности, подчиняющейся определённому закону распределения.

Область принятия - множество возможных значений статистического критерия, при которых нулевая гипотеза принимается.

Критическая область - множество возможных значений статистического критерия, при которых нулевая гипотеза отвергается.

Критические точки - точки, разграничивающие критическую область и область принятия гипотезы (рис. 3.5).

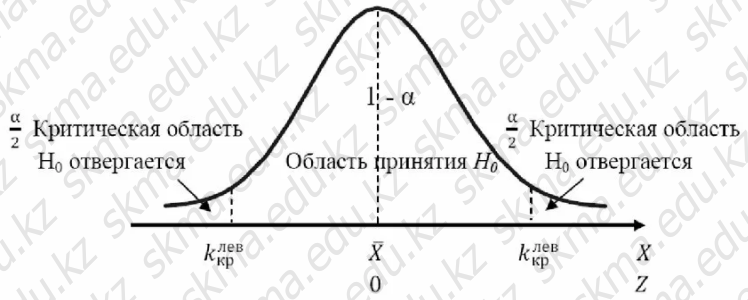


Рис. 3.5. Критическая область и область принятия гипотезы

Вид области принятия нулевой гипотезы зависит от вида альтернативной гипотезы.

Если альтернативная гипотеза содержит знак больше (например, $H_1: \bar{x} > 5$), то область принятия будет правосторонней.

Если альтернативная гипотеза содержит знак меньше (например, $H_1: \bar{x} < 5$), то область принятия будет левосторонней.

Если альтернативная гипотеза содержит знак не равно (например, $H_1: \bar{x} \neq 5$), то область принятия будет двусторонней.

Схема проверки статистических гипотез:

1. Выдвигаются две гипотезы: основная (нулевая) «H_0» и альтернативная «H_1».
2. Задается уровень значимости α.
3. По исходным данным, т.е. по выборке, вычисляется наблюдаемое значение критерия.
4. По специальным статистическим таблицам определяется табличное значение критерия.
5. Путем сравнения наблюдаемого и табличного значений делается вывод о правильности той или иной гипотезы.

Критерии согласия

Проверка выборочной совокупности на соответствие нормальному распределению является важнейшим этапом научного исследования, т.к. от результатов проверки гипотезы о нормальности распределения эмпирической совокупности зависит выбор метода статистического анализа данных. Т.е. без ответа на вопрос «Является ли выборка, подлежащая изучению, нормально распределенной?» нельзя применять, а тем более правильно интерпретировать результаты статистического анализа.

Для проверки гипотезы о нормальном распределении выборки применяют *критерии согласия*.

Критерии согласия дают возможность определить, когда расхождения между теоретическими и эмпирическими частотами следует признать несущественными, т.е. случайными, а когда – существенными, т.е. неслучайными (рис. 3.6).

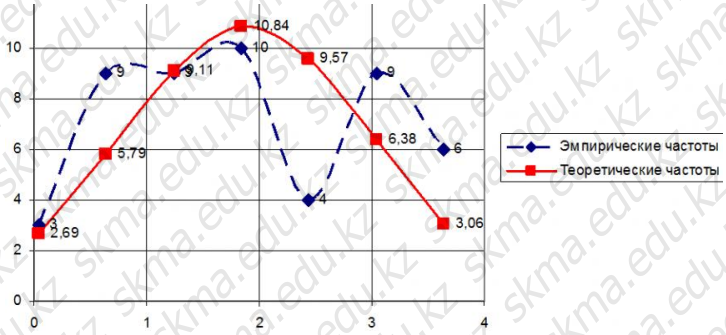


Рис. 3.6. Сопоставление эмпирических и теоретических частот

Наиболее распространенными критериями согласия являются критерии χ^2-Пирсона и Колмогорова-Смирнова.

Схема применения критерия согласия χ^2-Пирсона (chi-square goodness of fit test):

1) H_0: случайная величина «X» нормально распределена.
   H_1: случайная величина «X» не является нормально распределенной.
2) α=0,05 - уровень значимости.
3) $$\chi^2_{расч} = \sum_{i=1}^{k} \frac{(\nu_i - \nu_i^*)^2}{\nu_i^*}, \quad (3.4)$$

где k - число интервалов, на которое разбито эмпирическое распределение, ν_i - наблюдаемая частота признака в i-й группе, ν_i^* - теоретическая частота.

Теоретическая частота ν_i^* вычисляется по формуле:

$$\nu_i^* = n \cdot p_i, \quad (3.5)$$

где p_i - вероятности попадания случайной величины в интервал $[x_i, x_{i+1}]$, которые находятся по формуле:

$$p_i(x_i \le X \le x_{i+1}) = \Phi\left(\frac{x_{i+1} - \bar{x}}{s}\right) - \Phi\left(\frac{x_i - \bar{x}}{s}\right), \quad (3.6)$$

где $\bar{x}$ - выборочное среднее, s - среднее квадратическое отклонение, Ф(х) – функция распределения нормированного нормального распределения.

Для удобства вычислений заполняют таблицу вида 3.1. Сумма, подсчитанная в последнем столбце и будет расчетным значением $\chi^2_{расч}$.

Таблица 3.1.

Интервал $[x_i, x_{i+1}]$	Эмпирические частоты ν_i	Вероятности p_i	Теоретические частоты ν_i^*	$(\nu_i - \nu_i^*)^2$	$\frac{(\nu_i - \nu_i^*)^2}{\nu_i^*}$

4) $\chi^2_{табл}(\alpha; f)$,

где $f=k-3$ - число степеней свободы для нормального распределения (табличное значение), k - число интервалов, на которое разбита выборка.

5) Если $\chi^2_{расч} \le \chi^2_{табл}$, то «$H_0$» принимается.

   Если $\chi^2_{расч} > \chi^2_{табл}$, то «$H_0$» отвергается.

Критерий согласия Пирсона применяется при большом числе наблюдений ($n>30$), при этом частота каждой группы должна быть не менее пяти.

Схема применения критерия согласия Колмогорова – Смирнова (Kolmogorov – Smirnov goodness of fit test):

1) H_0: случайная величина «X» нормально распределена.
   H_1: случайная величина «X» не является нормально распределенной.
2) α=0,05 - уровень значимости.
3) $\lambda_{расч} = d_{\max}\sqrt{n}$, (3.7)

где $d_{\max} = \max|F_n(x) - F(x)|$ - максимальное значение абсолютной величины разности между наблюдаемой функцией распределения $F_n(x)$ и соответствующей теоретической функцией распределения $F(x)$, n - число наблюдений в статистическом ряду.

Значения теоретической функции распределения $F(x)$ для нормального распределения вычисляются по формуле:

$$F(x) = \frac{1}{2} + \Phi\left(\frac{x_{i+1} - \bar{x}}{s}\right), \quad (3.8)$$

где $\bar{x}$ - выборочное среднее, s - среднее квадратическое отклонение, Ф(x) – функция Лапласа.

Для удобства вычислений заполняют таблицу вида 3.2. Наибольшее значение в последнем столбце и будет расчетным значением $\lambda_{расч}$.

Таблица 3.2.

Интервал $[x_i, x_{i+1}]$	Эмпирические частоты ν_i	Накопленные частоты $\nu^*_{i,\ накопл}$	Наблюдаемая функция распределения $F_n(x) = \frac{\nu_{i,накопл}}{n}$	Теоретическая функция распределения $F(x)$	$\lvert F_n(x) - F(x)\rvert$

4) $\lambda_{табл}$=1,36 (табличное значение при α=0,05).
5) Если $\lambda_{расч} \le \lambda_{табл}$, то «$H_0$» принимается.
   Если $\lambda_{расч} > \lambda_{табл}$, то «$H_0$» отвергается.

Критерий Колмогорова-Смирнова применяется при большом числе наблюдений (n>30).

4. Иллюстративный материал: презентация, слайды.

5. Литература:

• Основная:

1. Койчубеков Б. К. Биостатистика. уч. пособие / Б.К. Койчубеков. - Алматы: Эверо, 2016.
2. Койчубеков Б.К. Биостатистика: учебное пособие. - Эверо, 2014.
3. Чудиновских В.Р. Абдикадыр Ж.Н. Применение программ EXCEL и SPSS statistics для статистического анализа медико-биологических данных. Уч. пособие.- ИП "АҚНҰР", 2021

• Дополнительная:

1. Койчубеков Б.К. Биостатистика. Монография.- Алматы: ТОО Эверо, 2024.- 152с.
2. Бухарбаев М. А. Медицинская статистика: учебное пособие / М. А. Бухарбаев, В. Н. Казагачев. - 2-е изд. - Алматы: Эпиграф, 2022. - 268 с.

• Электронные ресурсы:

1. Койчубеков Б.К., Сорокина М.А., Букеева А.С., Такуадина А.И. БИОСТАТИСТИКА в примерах и задачах: Учебно-метод. пособие/– Алматы ТОО «Эверо», 2020. – 80 с. https://elib.kz/ru/search/read_book/870/

2. Койчубеков Б.К. Биостатистика: Учебное пособие: Издательство «Эверо», Алматы, 2020, 154с. https://elib.kz/ru/search/read_book/867/
3. Биостатистика [Электронный ресурс]: учебник /К.Ж. Кудабаев [и др.].- Электрон. текстовые дан. (85,7Мб).- Шымкент: ЮКГФА, 2015. – 187с. эл. опт. диск (CD-ROM)
4. В.Р. Чудиновских, Ж.Н. Абдикадыр. Применение компьютерных программ для проверки статистических гипотез в медико-биологических исследованиях: учебное пособие. – Караганда: ИП «Издательство АҚНҰР».-2016, 100 с. https://aknurpress.kz/reader/web/1344
5. В.Р. Чудиновских, Ж.Н. Абдикадыр, А.Ш. Каипова, А.У. Алтаева. Применение программ EXCEL и SPSS Statistics для статистического анализа медико-биологических данных: уч.пос.– Караганда: ИП «Издательство АҚНҰР».– 2016, 128с. https://aknurpress.kz/reader/web/1342

6. Контрольные вопросы:

1. Что такое закон распределения случайной величины?
2. Какие виды распределений наиболее часто применяются на практике?
3. Почему при статистическом анализе медицинских данных наибольший интерес представляет нормальное распределение?
4. Какими свойствами обладает кривая нормального распределения?
5. От каких параметров зависит форма и расположение кривой нормального распределения?
6. Какие показатели характеризуют асимметричное распределение?
7. Что называется статистической гипотезой?
8. В чем различие между нулевой и альтернативной гипотезами?
9. Что называется ошибкой первого и второго рода?
10. Что называется доверительной вероятностью и уровнем значимости?
11. Что такое статистический критерий?
12. Какова общая схема проверки статистических гипотез?
13. Для чего используются критерии согласия?

ЛЕКЦИЯ №4-5

1. Тема: Параметрические методы сравнительной статистики

2. Цель: формирование у обучающихся представления о параметрических методах сравнительной статистики, их практическом применении и интерпретации результатов в контексте медицинских исследований.

3. Тезисы лекции:

Под *сравнительной статистикой* понимают проведение сравнительного анализа данных в двух и более группах. Это один из основных методов, применяемых в медицине и науке в целом для оценки эффективности различных подходов, стратегий, технологий.

Наиболее часто встречающейся задачей при проведении научных исследований в области медицины является сравнение данных, полученных в процессе наблюдений или экспериментов, в разных выборочных совокупностях. Например, одна выборка – *опытная* (исследователи оказали влияние на изучаемый объект или явление), а вторая выборка – *контрольная* (воздействия на объект наблюдения оказано не было).

Если исследователю удается заметить какие-либо численные различия в характеристиках сравниваемых выборок, то возникает вопрос: «Какова вероятность того, что эти различия неслучайны и будут систематически повторяться в дальнейшем при воспроизведении условий эксперимента?», или др. словами «Являются ли выявленные различия *статистически значимыми*?».

Выбор подходящего метода сравнения выборок определяется несколькими факторами:

- *типом сравниваемых показателей (количественные или качественные);*
- *видом распределения;*

- *числом сравниваемых групп;*
- *зависимостью или независимостью выборок.*

В зависимости от *вида распределения* рассматриваемых выборок к ним могут быть применены *параметрические* и *непараметрические* методы (или статистические критерии).

Параметрические критерии предполагают наличие *нормального распределения* в сравниваемых выборках и используют в процессе расчета параметры распределения (средние, дисперсии, среднее квадратическое отклонение). Например, *t*-критерий Стьюдента, *F*-критерий Фишера и др.

Непараметрические критерии не предполагают нормального распределения в сравниваемых выборках и используют в процессе расчета *ранги* (порядковые номера) значений признака. Например, критерий Манна-Уитни, критерий Уилкоксона, критерий знаков и др.

Непараметрические критерии дают немного более грубые оценки, чем параметрические, но являются более универсальными. Параметрические методы более точны, но могут применяться только для нормально распределенных выборок.

Различают *независимые* и *зависимые* выборочные совокупности.

Независимые (несвязанные) выборки – это разные группы объектов, характеризуются тем, что вероятность отбора любого объекта одной выборки не зависит от отбора любого объекта из другой выборки.

Зависимые (связанные) выборки – это одна и та же группа объектов, но изучаемая в разные моменты времени.

Например, фармацевтическая компания хочет проверить эффективность нового препарата для снижении артериального давления. Данные могут собираться двумя способами:

1 - *одной* группе людей назначается новый препарат, а *другой* - плацебо, а затем сравнивается артериальное давление между группами. Выборки *независимые*;

2 - артериальное давление измеряется у одних и тех же людей *до* и *после* приема препарата. Выборки *зависимые*.

Рассмотрим некоторые параметрические методы сравнительной статистики.

1. F-критерий Фишера (F-test)

Для проверки гипотезы о равенстве дисперсий двух выборок используют *F-критерий Фишера*. Он способен корректно оценить дисперсии только в том случае, если обе выборки *независимы* и имеют *нормальное распределение*.

Схема применения F-критерия Фишера:

1) H_0: $s_1^2 = s_2^2$
   H_1: $s_1^2 \neq s_2^2$.
2) α=0,05 - уровень значимости.
3) $$F_{расч} = \frac{s_1^2}{s_2^2}.$$ (4.1)
4) $F_{табл}(\alpha; f_1; f_2)$,
   где $f_1=n_1-1$, $f_2=n_2-1$ – число степеней свободы.
5) Если $F_{расч} \leq F_{табл}$, то «H_0» принимается.
   Если $F_{расч} > F_{табл}$, то «H_0» отвергается.

2. Двухвыборочный t-критерий Стьюдента (two sample t-test)

Для проверки гипотезы о равенстве двух выборочных средних для *независимых* выборок используют *двухвыборочный t-критерий Стьюдента*.

Правила применения t-критерия Стьюдента:

1) обе сравниваемые выборки должны иметь *нормальное распределение*;
2) возможно сравнивать только *две группы*;
3) данный критерий желательно применять для малых выборок (n<30), т.к. увеличение

объема выборки повышает чувствительность критерия, но при значительном увеличении числа наблюдений возможно выявить изменения, не имеющие существенного значения;

4) необходимо учитывать *наличие/отсутствие однородности дисперсий* (равенство/неравенство дисперсий) в выборках. Для определения однородности дисперсий необходимо применить *F*-критерий Фишера.

Схема применения двухвыборочного t-критерия Стьюдента, в случае если дисперсии равны:

1) H_0: $\bar{x}_1 = \bar{x}_2$
   H_1: $\bar{x}_1 \neq \bar{x}_2$.
2) α=0,05 - уровень значимости.
3) $$t_{расч} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}} \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2} \cdot (n_1 + n_2 - 2)}, \quad (4.2)$$

где n_1, n_2 - объемы рассматриваемых выборок, s_1^2, s_2^2 - дисперсии рассматриваемых выборок, $\bar{x}_1$, $\bar{x}_2$ - сравниваемые средние значения выборок.

4) $t_{табл}(\alpha; f)$, где $f = n_1 + n_2 - 2$- число степеней свободы.
5) Если $t_{расч} \leq t_{табл}$, то «H_0» принимается.
   Если $t_{расч} > t_{табл}$, то «H_0» отвергается.

Схема применения двухвыборочного t-критерия Стьюдента, в случае если дисперсии неравны:

1) H_0: $\bar{x}_1 = \bar{x}_2$
   H_1: $\bar{x}_1 \neq \bar{x}_2$.
2) α=0,05 - уровень значимости.
3) $$t_{\text{расч}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \quad (4.3)$$

где n_1, n_2 - объемы рассматриваемых выборок, s_1^2, s_2^2 - дисперсии рассматриваемых выборок, $\bar{x}_1$, $\bar{x}_2$ - сравниваемые средние значения выборок.

4) $t_{\text{табл}}(\alpha; f)$, где $f = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1 - 1)} + \frac{s_2^4}{n_2^2(n_2 - 1)}}$ - число степеней свободы.
5) Если $t_{расч} \leq t_{табл}$, то «H_0» принимается.
   Если $t_{расч} > t_{табл}$, то «H_0» отвергается.

3. Парный t-критерий Стьюдента (paired t-test)

Для проверки гипотезы о равенстве двух выборочных средних для *зависимых* выборок используют *парный t-критерий Стьюдента*.

Схема применения парного t-критерия Стьюдента:

1) H_0: $\overline{x_1} = \overline{x_2}$,
   H_1: $\overline{x_1} \neq \overline{x_2}$
2) α=0,05- уровень значимости
3) $$t_{расч} = \frac{\bar{d}}{S_d / \sqrt{n}}, \quad (4.4)$$

где $d = x_{i1} - x_{i2}$ - разности между соответствующими значениями пар переменных, n - объем

выборки, $\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$, $S_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}}$

4) $t_{табл}(\alpha; f)$, где $f = n - 1$ - число степеней свободы.

5) Если $t_{расч} \le t_{табл}$, то «H_0» принимается.
Если $t_{расч} > t_{табл}$, то «H_0» отвергается.

4. Одновыборочный t-критерий (one sample t-test)

Данный критерий предназначен для проверки гипотезы о равенстве выборочного среднего какому-либо значению.

Схема применения одновыборочного t-критерия:

1) $H_0: \bar{x} = a$,
$H_1: \bar{x} \neq a$

2) α=0,05- уровень значимости

3) $t_{\text{расч}} = \frac{\bar{x} - a}{s/\sqrt{n}}$, (4.5)

4) $t_{табл}(\alpha; f)$, где $f = n - 1$ - число степеней свободы.

5) Если $t_{расч} \le t_{табл}$, то «H_0» принимается.
Если $t_{расч} > t_{табл}$, то «H_0» отвергается.

5. Однофакторный дисперсионный анализ (One-way Analysis of Variance, ANOVA)

Дисперсионный анализ применяется для проверки гипотезы о равенстве выборочных средних в случае, когда рассматривается более двух независимых выборок (k>2), имеющих нормальное распределение.

Схема применения однофакторного дисперсионного анализа, в случае когда дисперсии равны:

1) $H_0: \bar{x}_1 = \bar{x}_2 = \ldots = \bar{x}_k$.
$H_1: \bar{x}_1 \neq \bar{x}_2 \neq \ldots \neq \bar{x}_k$.

2) α=0,05- уровень значимости.

3)

3.1) общая средняя $\bar{x} = \frac{\bar{x}_1 + \bar{x}_2 + \cdots + \bar{x}_k}{k}$.

3.2) факторная сумма квадратов отклонений $SS_{\text{факт}} = r\sum(\bar{x}_{\text{гр } j} - \bar{x})^2$, (4.6)

где r - количество значений в каждой выборке.

3.3) остаточная сумма квадратов отклонений

$$SS_{\text{ост}} = \sum_{i=1}^{r}(x_{i1} - \bar{x}_{\text{гр1}})^2 + \sum_{i=1}^{r}(x_{i2} - \bar{x}_{\text{гр2}})^2 + \cdots + \sum_{i=1}^{r}(x_{ik} - \bar{x}_{\text{гр}k})^2 \quad (4.7)$$

где k – количество выборок

3.4) факторная дисперсия $S^2_{\text{факт}} = \frac{SS_{\text{факт}}}{k-1}$. (4.8)

3.5) остаточная дисперсия $S^2_{\text{ост}} = \frac{SS_{\text{ост}}}{k(r-1)}$.

(4.9)

3.6) $F_{\text{расч}} = \frac{S^2_{\text{факт}}}{S^2_{\text{ост}}}$. (4.10)

4) $F_{табл}(\alpha; f_1; f_2)$,

где $f_1 = k-1$, $f_2 = k(r-1)$ – число степеней свободы (k - количество выборок, r - количество значений в каждой выборке).

5) Если $F_{расч} \leq F_{табл}$, то «H_0» принимается.
Если $F_{расч} > F_{табл}$, то «H_0» отвергается.

4. Иллюстративный материал: презентация, слайды.

5. Литература:

• Основная:

1. Койчубеков Б. К. Биостатистика. уч. пособие / Б.К. Койчубеков. - Алматы: Эверо, 2016.
2. Койчубеков Б.К. Биостатистика: учебное пособие. - Эверо, 2014.
3. Чудиновских В.Р. Абдикадыр Ж.Н. Применение программ EXCEL и SPSS statistics для статистического анализа медико-биологических данных. Уч. пособие.- ИП "АҚНҰР", 2021

• Дополнительная:

1. Койчубеков Б.К. Биостатистика. Монография.- Алматы: ТОО Эверо, 2024.- 152с.
2. Бухарбаев М. А. Медицинская статистика: учебное пособие / М. А. Бухарбаев, В. Н. Казагачев. - 2-е изд. - Алматы: Эпиграф, 2022. - 268 с.

• Электронные ресурсы:

1. Койчубеков Б.К., Сорокина М.А., Букеева А.С., Такуадина А.И. БИОСТАТИСТИКА в примерах и задачах: Учебно-метод. пособие/– Алматы ТОО «Эверо», 2020. – 80 с. https://elib.kz/ru/search/read_book/870/
2. Койчубеков Б.К. Биостатистика: Учебное пособие: Издательство «Эверо», Алматы, 2020, 154с. https://elib.kz/ru/search/read_book/867/
3. Биостатистика [Электронный ресурс]: учебник /К.Ж. Кудабаев [и др.].- Электрон. текстовые дан. (85,7Мб).- Шымкент: ЮКГФА, 2015. – 187с. эл. опт. диск (CD-ROM)
4. В.Р. Чудиновских, Ж.Н. Абдикадыр. Применение компьютерных программ для проверки статистических гипотез в медико-биологических исследованиях: учебное пособие. – Караганда: ИП «Издательство АҚНҰР».-2016, 100 с. https://aknurpress.kz/reader/web/1344
5. В.Р. Чудиновских, Ж.Н. Абдикадыр, А.Ш. Каипова, А.У. Алтаева. Применение программ EXCEL и SPSS Statistics для статистического анализа медико-биологических данных: уч.пос.– Караганда: ИП «Издательство АҚНҰР».– 2016, 128с. https://aknurpress.kz/reader/web/1342

6. Контрольные вопросы:

1. Какими факторами определяется выбор подходящего метода для сравнения выборок?
2. В чем различие между параметрическими и непараметрическими статистическими методами?
3. В чем различие между зависимыми и независимыми выборками?
4. Для проверки какой гипотезы используется *F*-критерий Фишера? Какова его схема?
5. Для проверки какой гипотезы используется *t*-критерий Стьюдента? Каковы условия его применения?
6. В чем различие между двухвыборочным и парным критерием Стьюдента?
7. Какова схема двухвыборочного критерия Стьюдента?
8. Какова схема парного критерия Стьюдента?
9. Для проверки какой гипотезы используется одновыборочный *t*-критерий? Какова его схема?
10. Для проверки какой гипотезы используется однофакторный дисперсионный анализ? Какова его схема?

ЛЕКЦИЯ №6

1. Тема: Непараметрические методы сравнительной статистики

2. Цель: формирование у обучающихся представления о непараметрических методах сравнительной статистики, их практическом применении и интерпретации результатов в контексте медицинских исследований.

3. Тезисы лекции:

В зависимости от *вида распределения* рассматриваемых выборок к ним могут быть применены *параметрические* и *непараметрические* методы (или статистические критерии).

Непараметрические методы используют в двух случаях: *когда выборочные данные не*

имеют нормального распределения или *при малом количестве наблюдений.*

Для расчета непараметрических критериев применяется процедура *ранжирования* значений сравниваемого признака, т.е. расстановка значений в порядке возрастания.

Ранг - порядковый номер значения признака.

Если числа не повторяются, то их ранги соответствуют их порядковым номерам. Если же некое число повторяется несколько раз, то всем им приписывается *средний ранг.*

Для каждого параметрического критерия имеется, по крайней мере, один непараметрический аналог.

Рассмотрим некоторые непараметрические методы сравнительной статистики.

1. U-критерий Манна - Уитни (Mann – Whitney U-test)

Этот ранговый критерий используется для проверки гипотезы о равенстве выборочных средних в случае двух *независимых* выборок, которые не имеют нормального распределения.

Данный критерий является непараметрическим аналогом двухвыборочного *t*-критерия Стьюдента.

U-критерий подходит для сравнения малых выборок. В каждой из выборок должно быть не менее 3 значений признака. Допускается, чтобы в одной выборке было 2 значения, но во второй тогда должно быть не менее пяти (n_1, $n_2 \geq 3$ или $n_1=2$, $n_2 \geq 5$).

Условием для применения *U*-критерия Манна-Уитни является отсутствие в сравниваемых группах совпадающих значений признака (все числа разные) или очень малое число таких совпадений.

Схема применения U-критерия Манна - Уитни:

1) H_0: $\bar{x}_1 = \bar{x}_2$
   H_1: $\bar{x}_1 \neq \bar{x}_2$.
2) α=0,05 - уровень значимости.
3) Значения двух выборок ранжируют как одну совокупность. Затем подсчитывают отдельно суммы рангов для первой и второй выборок.

$$U_{расч} = n_1 \cdot n_2 + \frac{n_x \cdot (n_x + 1)}{2} - T_x, \qquad (5.1)$$

где T_x - большая из двух ранговых сумм, n_x - объем выборки, соответствующий T_x, n_1, n_2 - объемы рассматриваемых выборок.

4) $U_{табл}(\alpha; n_1; n_2)$
5) Если $U_{расч} > U_{табл}$, то «H_0» принимается.
   Если $U_{расч} \leq U_{табл}$, то «H_0» отвергается.

2. T-критерий Уилкоксона (Wilcoxon T-test)

Этот ранговый критерий используется для проверки гипотезы о равенстве выборочных средних в случае двух *зависимых* выборок.

Данный критерий является непараметрическим аналогом парного *t*-критерия Стьюдента.

Число исследуемых объектов в выборке при использовании *T*-критерия Уилкоксона должно быть не менее 5.

Схема применения T-критерия Уилкоксона:

1) H_0: $\bar{x}_1 = \bar{x}_2$
   H_1: $\bar{x}_1 \neq \bar{x}_2$.
2) α=0,05 - уровень значимости.
3) Вычисляют разности для каждого случая (ДО-ПОСЛЕ). Ранжируют абсолютные значения разностей (т.е. ранжируют по модулю без учета знака), не учитывая нули. Присваивают знаки («+» или «–») разностей каждому рангу. Получают знаковые ранги. $T_{расч}$ определяется как наименьшее из значений $T+$ и $T-$, которые являются суммами положительных и отрицательных рангов соответственно.

4) $T_{табл}$ ($\alpha; n$).

5) Если $T_{расч} > T_{табл}$, то «H_0» принимается.

Если $T_{расч} \leq T_{табл}$, то «H_0» отвергается.

3. H-критерий Краскела-Уоллиса (Kruskal - Wallis H-test)

Данный критерий является непараметрическим аналогом однофакторного дисперсионного анализа и используется для сравнения трех и более независимых групп, не имеющих нормального распределения.

При сопоставлении трех выборок допускается, чтобы в каждой из них было не менее 3 наблюдений, или в одной из них 4 наблюдения, а в двух других – по 2; при этом неважно, в какой именно выборке сколько испытуемых, а важно соотношение 4:2:2.

Таблица критических значений *H*-критерия предусмотрена только для случая, когда число выборок $k \leq 5$, а число испытуемых в каждой группе $n_i \leq 8$. При большом количестве выборок и испытуемых в каждой выборке необходимо пользоваться таблицей критических значений χ^2-критерия, т.к. критерий Краскела-Уоллиса асимптотически приближается к распределению «χ^2».

Схема применения H-критерия Краскела – Уоллиса:

1) H_0: $\bar{x}_1 = \bar{x}_2 = \ldots = \bar{x}_k$.

H_1: $\bar{x}_1 \neq \bar{x}_2 \neq \ldots \neq \bar{x}_k$.

2) α=0,05- уровень значимости.

3)
$$H_{\text{расч}} = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) \quad (5.2)$$

где $n = \sum_{i=1}^{k} n_i$ - общее число наблюдений по всем группам, R_i - сумма рангов i-ой выборки.

В случае, когда число выборок $k \leq 5$ $H_{табл}$ (α;n_1;n_2;…;n_5), где n_1, n_2, …, n_5 – объемы рассматриваемых выборок.

В случае, когда число выборок $k > 5$ $H_{табл} = \chi^2_{табл}$ (α ; f), где $f = k-1$ – число степеней свободы (табличное значение).

5) Если $H_{расч} < H_{табл}$, то «H_0» принимается.

Если $H_{расч} \geq H_{табл}$, то «H_0» отвергается.

4. Иллюстративный материал: презентация, слайды.

5. Литература:

• Основная:

1. Койчубеков Б. К. Биостатистика. уч. пособие / Б.К. Койчубеков. - Алматы: Эверо, 2016.
2. Койчубеков Б.К. Биостатистика: учебное пособие. - Эверо, 2014.
3. Чудиновских В.Р. Абдикадыр Ж.Н. Применение программ EXCEL и SPSS statistics для статистического анализа медико-биологических данных. Уч. пособие.- ИП "АҚНҰР", 2021

• Дополнительная:

1. Койчубеков Б.К. Биостатистика. Монография.- Алматы: ТОО Эверо, 2024.- 152с.
2. Бухарбаев М. А. Медицинская статистика: учебное пособие / М. А. Бухарбаев, В. Н. Казагачев. - 2-е изд. - Алматы: Эпиграф, 2022. - 268 с.

• Электронные ресурсы:

1. Койчубеков Б.К., Сорокина М.А., Букеева А.С., Такуадина А.И. БИОСТАТИСТИКА в примерах и задачах: Учебно-метод. пособие/– Алматы ТОО «Эверо», 2020. – 80 с. https://elib.kz/ru/search/read_book/870/
2. Койчубеков Б.К. Биостатистика: Учебное пособие: Издательство «Эверо», Алматы, 2020, 154с. https://elib.kz/ru/search/read_book/867/

3. Биостатистика [Электронный ресурс]: учебник /К.Ж. Кудабаев [и др.].- Электрон. текстовые дан. (85,7Мб).- Шымкент: ЮКГФА, 2015. – 187с. эл. опт. диск (CD-ROM)
4. В.Р. Чудиновских, Ж.Н. Абдикадыр. Применение компьютерных программ для проверки статистических гипотез в медико-биологических исследованиях: учебное пособие. – Караганда: ИП «Издательство АҚНҰР».-2016, 100 с. https://aknurpress.kz/reader/web/1344
5. В.Р. Чудиновских, Ж.Н. Абдикадыр, А.Ш. Каипова, А.У. Алтаева. Применение программ EXCEL и SPSS Statistics для статистического анализа медико-биологических данных: уч.пос.– Караганда: ИП «Издательство АҚНҰР».– 2016, 128с. https://aknurpress.kz/reader/web/1342

6. Контрольные вопросы:

1. В каких случаях используют непараметрические методы сравнения?
2. Что означают понятия «ранг» и «ранжирование»?
3. Для проверки какой гипотезы используется *U*-критерий Манна - Уитни? Каковы условия его применения?
4. Какова схема *U*-критерия Манна - Уитни?
5. Для проверки какой гипотезы используется *T*-критерий Уилкоксона? Каковы условия его применения?
6. Какова схема *T*-критерия Уилкоксона?
7. Для проверки какой гипотезы используется *H*-критерий Краскела-Уоллиса? Каковы условия его применения?
8. Какова схема *H*-критерия Краскела-Уоллиса?

ЛЕКЦИЯ №7

1. Тема: Анализ качественных признаков

2. Цель: ознакомить обучающихся с некоторыми методами анализа качественных признаков в контексте медицинских исследований.

3. Тезисы лекции:

Качественные признаки - это характеристики, которые не измеряются числовыми значениями, а описывают качество объекта (например, пол, группа крови, цвет глаз, цвет волос, семейное положение, наличие/отсутствие заболевания, место проживания (город/село) и др.).

Анализ качественных признаков позволяет систематизировать и классифицировать информацию, не сводя ее к численным значениям.

Анализ качественных данных важен для выявления взаимосвязей между различными факторами при проведении медицинских исследований. Позволяет выделить группы риска, уточнить диагнозы и принимать обоснованные решения в плане лечения и профилактики.

Существует несколько типов качественных признаков: *бинарные, номинальные* и *порядковые*.

Бинарные признаки - признаки, которые могут принимать только два возможных значения. Например, пол (мужской/женский), наличие/отсутствие заболевания, артериальное давление в норме/повышено, послеоперационные осложнения есть/нет.

Номинальные признаки - признаки, которые разделяют объекты на категории, но не имеют упорядоченного отношения между этими категориями. Например, цвет глаз, тип крови, страна проживания. Нет явного порядка между категориями.

Порядковые признаки - признаки, которые разделяют объекты на категории, которые имеют определенный порядок, но различия между значениями не являются равными. Например, образование (нет, среднее, высшее), степень тяжести болезни (легкая, средняя, тяжелая). Существует порядок, но интервалы между значениями могут быть неодинаковыми.

Построение таблиц сопряженности является важным методом анализа качественных

данных, особенно в медицинском исследовании. Такие таблицы позволяют исследовать связи между двумя и более качественными переменными и определять степень их взаимосвязи.

Рассмотрим таблицу сопряженности размера 2x2 (табл. 6.1). Имеется два бинарных признака: A - с исходами A_1, A_2; B - с исходами B_1, B_2. В центральной части таблицы записаны частоты встречаемости комбинаций этих признаков - *a, b, c, d*.

Таблица 6.1.

Таблица сопряженности размера 2x2

A \ B	B_1	B_2	Сумма
A_1	*a*	*b*	*a+b*
A_2	*c*	*d*	*c+d*
Сумма	*a+c*	*b+d*	*n=a+b+c+d*

Пример 6.1. В таблице 6.2 представлены данные о приеме пациентами с одинаковым диагнозом двух видов лекарственных препаратов (A и B) и их состоянии (болен/здоров) после 5 дней приема. Всего обследовано 39 пациентов, из них здоровых 25, больных – 14. Препарат A принимали 18 человек, препарат B – 21.

Таблица 6.2.

Таблица сопряженности размера 2x2

Вид лекарства \ Состояние	Здоров	Болен	Сумма
A	10	8	18
B	15	6	21
Сумма	25	14	39

Рассмотрим таблицу сопряженности размера rxs (табл. 6.3). Имеется два признака: A – с исходами A_1, A_2, …, A_r и B – с исходами B_1, B_2, …, B_s. В центральной части таблицы записаны частоты встречаемости различных комбинаций признаков A и B - ν_{ij}.

Таблица 6.3.

Таблица сопряженности размера r x s

A \ B	B_1	B_2	…	B_s	Сумма
A_1	ν_{11}	ν_{12}	…	ν_{1s}	$\nu_{1.}$
A_2	ν_{21}	ν_{22}	…	ν_{2s}	$\nu_{2.}$
…	…	…	…	…	…
A_r	ν_{r1}	ν_{r2}	…	ν_{rs}	$\nu_{r.}$
Сумма	$\nu_{.1}$	$\nu_{.2}$	…	$\nu_{.s}$	$\nu_{..}$

Пример 6.2. В таблице 6.4. представлены данные о количестве наблюдений и случаев летальности для четырех форм острых гнойных деструкций легких.

Таблица 6.4.

Таблица сопряженности размера 4x2

Форма заболевания \ Исход	Летальный исход	Выздоровление	Сумма
Гнойный абсцесс	5	136	141

Гангренозный абсцесс	11	37	48
Гангрена доли	7	8	15
Тотальная гангрена	6	5	11
Сумма	29	186	215

Таблицы сопряженности могут быть использованы при проверке гипотезы о статистической значимости связи между качественными переменными.

1. χ^2-критерий Пирсона (Pearson's chi-squared test)

Данный критерий применяется для анализа качественных признаков в независимых выборках, если в ячейках таблицы сопряженности частоты больше или равны 5.

Схема применения χ^2-критерия Пирсона (таблица размера rxs):

1) *H_0:* связи между качественными признаками нет.
H_1: связь между качественными признаками есть.
2) α=0,05 - уровень значимости.
3) $$\chi^2_{расч} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(\nu_{ij} - \nu^*_{ij}\right)^2}{\nu^*_{ij}} \quad (6.1)$$

где ν_{ij} – наблюдаемые частоты, ν^*_{ij} – теоретические (ожидаемые) частоты.

Теоретические (ожидаемые) частоты вычисляются по формуле:

$$\nu^*_{ij} = \nu_{.i} \cdot \frac{\nu_{j.}}{\nu_{..}}, \quad (6.2)$$

где $\nu_{.i}$ – сумма по столбцу i, $\nu_{j.}$ – сумма по строке j, $\nu_{..}$ – общее число наблюдений.

4) $\chi^2_{табл}(\alpha; f)$, где $f=(r-1)(s-1)$ - число степеней свободы
5) Если $\chi^2_{расч} \le \chi^2_{табл}$, то «$H_0$» принимается.
Если $\chi^2_{расч} > \chi^2_{табл}$, то «$H_0$» отвергается.

Схема применения χ^2-критерия Пирсона (таблица размера 2x2):

1) *H_0:* связи между качественными признаками нет.
H_1: связь между качественными признаками есть.
2) α=0,05 - уровень значимости.
3) $$\chi^2_{расч} = \frac{(ad-bc)^2 n}{(a+b)(c+d)(a+c)(b+d)}, \quad (6.3)$$

где *a, b, c, d* – наблюдаемые частоты, *n* – общее число наблюдений (см. таблицу 6.1).

4) $\square^2_{табл}$=3,8 (для таблиц размера 2x2).
5) Если $\square^2_{расч} \square \square_{табл}$, то «$H_0$» принимается.
Если $\square^2_{расч} \square \square_{табл}$, то «$H_0$» отвергается.

Поправка Йетса (Yates's correction).

Формула (6.3) дает завышенные значения. На практике это приводит к тому, что нулевая гипотеза отвергается слишком часто. Чтобы компенсировать этот эффект, в формулу вводят поправку Йетса:

$$\chi^2_{расч} = \frac{\left(ad - bc - \frac{n}{2}\right)^2 n}{(a+b)(c+d)(a+c)(b+d)}. \quad (6.4)$$

При решении задач нужно производить вычисления по формулам (6.3) и (6.4)

2. Точный критерий Фишера (Fisher's exact test)

Данный критерий применяется для анализа качественных признаков в независимых выборках. Он подходит для сравнения очень малых выборок (если наблюдаемые частоты

меньше 5). Используется только для таблиц сопряженности размера 2x2.

Схема применения точного критерия Фишера:

1) H_0: связи между качественными признаками нет.
   H_1: связь между качественными признаками есть.
2) α=0,05 - уровень значимости.
3) $$P_{расч} = \frac{(a+b)!\cdot(c+d)!\cdot(a+c)!\cdot(b+d)!}{a!\cdot b!\cdot c!\cdot d!\cdot n!} \quad (6.5)$$

где a, b, c, d – наблюдаемые частоты, n – общее число наблюдений, ! – факториал, представляет собой произведение числа на последовательность чисел, каждое из которых меньше предыдущего на 1 (например, 4!=4·3·2·1).

4) -
5) Расчетное значение критерия сравнивается с уровнем значимости 0,05.
   Если $P_{расч} \geq 0{,}05$, то «H_0» принимается.
   Если $P_{расч} < 0{,}05$, то «H_0» отвергается.

3. χ^2-критерий Макнемара (McNemar's test)

Данный критерий применяется для сравнения бинарных качественных признаков, взятых из *зависимых выборок (до и после).* Используется только для таблиц сопряженности размера 2x2.

Схема применения χ^2-критерия Макнемара:

1) H_0: между парами категорий нет статистически значимых различий
   (др. словами эффект лечения/вмешательства *одинаков* в двух разных моментах времени, например, до и после лечения).
   H_1: между парами категорий существуют статистически значимые различия
   (др. словами эффект лечения/вмешательства *различается* в двух разных моментах времени, например, до и после лечения).
3) α=0,05 - уровень значимости.
4) $$\chi^2_{расч} = \frac{(b-c)^2}{b+c}. \quad (6.6)$$
5) $\square^2_{табл}$ =3,8 (для таблиц размера 2x2).
6) Если $\square^2_{расч} \square \square_{табл}$, то «H_0» принимается.
   Если $\square^2_{расч} \square \square_{табл}$, то «H_0» отвергается.

4. Иллюстративный материал: презентация, слайды.

5. Литература:

- Основная:

1. Койчубеков Б. К. Биостатистика. уч. пособие / Б.К. Койчубеков. - Алматы: Эверо, 2016.
2. Койчубеков Б.К. Биостатистика: учебное пособие. - Эверо, 2014.
3. Чудиновских В.Р. Абдикадыр Ж.Н. Применение программ EXCEL и SPSS statistics для статистического анализа медико-биологических данных. Уч. пособие.- ИП "АҚНҰР", 2021

- Дополнительная:

1. Койчубеков Б.К. Биостатистика. Монография.- Алматы: ТОО Эверо, 2024.- 152с.
2. Бухарбаев М. А. Медицинская статистика: учебное пособие / М. А. Бухарбаев, В. Н. Казагачев. - 2-е изд. - Алматы: Эпиграф, 2022. - 268 с.

- Электронные ресурсы:

1. Койчубеков Б.К., Сорокина М.А., Букеева А.С., Такуадина А.И. БИОСТАТИСТИКА в примерах и задачах: Учебно-метод. пособие/– Алматы ТОО «Эверо», 2020. – 80 с. https://elib.kz/ru/search/read_book/870/
2. Койчубеков Б.К. Биостатистика: Учебное пособие: Издательство «Эверо», Алматы, 2020, 154с. https://elib.kz/ru/search/read_book/867/

3. Биостатистика [Электронный ресурс]: учебник /К.Ж. Кудабаев [и др.].- Электрон. текстовые дан. (85,7Мб).- Шымкент: ЮКГФА, 2015. – 187с. эл. опт. диск (CD-ROM)
4. В.Р. Чудиновских, Ж.Н. Абдикадыр. Применение компьютерных программ для проверки статистических гипотез в медико-биологических исследованиях: учебное пособие. – Караганда: ИП «Издательство АҚНҰР».-2016, 100 с. https://aknurpress.kz/reader/web/1344
5. В.Р. Чудиновских, Ж.Н. Абдикадыр, А.Ш. Каипова, А.У. Алтаева. Применение программ EXCEL и SPSS Statistics для статистического анализа медико-биологических данных: уч.пос.– Караганда: ИП «Издательство АҚНҰР».– 2016, 128с. https://aknurpress.kz/reader/web/1342

6. Контрольные вопросы:

1. В чем особенность анализа качественных признаков?
2. Почему анализ качественных признаков важен в медицинских исследованиях?
3. Что из себя представляет таблица сопряженности 2x2 и размера rxs?
4. Какие условия должны выполняться при применении χ^2-критерия Пирсона?
5. Какова схема χ^2-критерия Пирсона?
6. Для чего используется поправка Йетса?
7. В каких случаях применяется точный критерий Фишера?
8. Какова схема точного критерия Фишера?
9. В каких случаях применяется χ^2-критерий Макнемара?
10. Какова схема χ^2-критерия Макнемара?

ЛЕКЦИЯ №8

1. **Тема:** Корреляционный анализ.
2. **Цель:** Ознакомить студентов с основами корреляционного анализа.
3. **Тезисы лекции:**

Одной из важных задач эпидемиологии является анализ заболеваемости по факторам риска.

Фактор риска в медицине - это фактор, способствующий возникновению заболевания (например, курение - фактор риска по отношению к инфаркту миокарда или раку, число аварий в сети водопровода - фактор риска по отношению к дизентерии).

Для количественной оценки факторов риска развития заболевания используется корреляционный анализ.

Корелляционный анализ - это количественный метод определения тесноты и направления связи между двумя и более случайными величинами.

Впервые в научный оборот термин «корреляция» ввел французский палеонтолог Ж. Кювье (XVIII в.), а в статистике его первым стал использовать Ф. Гальтон (XIX в.).

Для того чтобы охарактеризовать связь между переменными численно, вводится понятие коэффициента корреляции.

Коэффициент корреляции - показатель, характеризующий силу связи и ее направление, принимает значения в промежутке [-1, 1].

Для оценки силы связи в теории корреляции применяется шкала английского статистика Чеддока (таблица 7.1).

Таблица 7.1.

Количественная мера тесноты связи	Качественная характеристика силы связи
0,1 - 0,3	Слабая
0,3 - 0,5	Умеренная
0,5 - 0,7	Заметная
0,7 - 0,9	Высокая

0,9 - 1	Сильная

По направлению различают *прямую* и *обратную корреляционную связь*. *Прямая корреляционная связь* - связь, при которой увеличение одной переменной связано с увеличением другой переменной (рост заболеваемости дизентерией при увеличении в воде водопровода доли нестандартных проб вод).

Обратная корреляционная связь - связь, при которой увеличение одной переменной связано с уменьшением другой переменной (снижение заболеваемости гепатитом «В» по мере увеличения охвата населения вакцинацией против этой инфекции).

При прямой связи коэффициент корреляции принимает значения от «0» до «+1».

При обратной связи коэффициент корреляции принимает значения от «-1» до «0».

Если коэффициент корреляции равен «0», то связь между явлениями отсутствует.

Если коэффициент корреляции равен «+1» или «–1», то связь между явлениями функциональная.

При анализе зависимости между двумя переменными применяют диаграммы рассеяния.

Диаграмма рассеяния - наглядный способ представления корреляционной зависимости между двумя переменными (рисунок 7.1).

Диаграмма рассеяния - это точечная диаграмма в виде графика, получаемого путем нанесения в определенном масштабе экспериментальных, полученных в результате наблюдений точек. Координаты точек на графике соответствуют значениям рассматриваемой величины и влияющего на него фактора. Расположение точек показывает наличие и характер связи между двумя переменными.

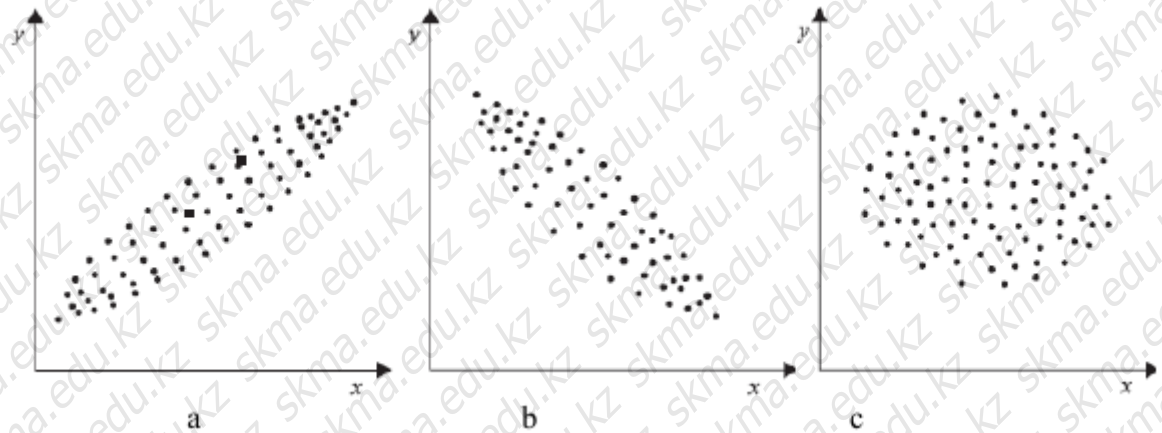


Рис 7.1. Диаграммы рассеяния: a - прямая связь; b - обратная связь; c - связь отсутствует

Линейный (парный) коэффициент корреляции (Пирсона) характеризующий силу связи и ее направление: $r_{xy} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$, где r_{xy} – коэффициент корреляции; x и y – коррелируемые ряды; $\bar{x}, \bar{y}$ - средние значения.

Парный коэффициент корреляции является параметрическим коэффициентом.

Применение парного коэффициента корреляции Пирсона возможно, если выполняются следующие условия:

- сравниваемые переменные должны быть получены в интервальной шкале или шкале отношений;
- распределения переменных должны быть близки к нормальному;
- число значений рассматриваемых переменных должно быть одинаковым.

Достоверность коэффициента корреляции определяется сравнением его с вычисляемой средней ошибкой.

Средняя ошибка коэффициента корреляции: $m_r = \pm\frac{1-r_{xy}^2}{\sqrt{n}}$, где где r_{xy} – коэффициент корреляции; n - число наблюдений.

Коэффициент корреляции считается *достоверным*, если в 3 раза превышает свою среднюю ошибку. Иначе необходимо увеличить число наблюдений.

Достоверность коэффициента корреляции определяется по специальным таблицам.

Пример 7.1. Для следующих данных рассчитать линейный коэффициент корреляции Пирсона:

Заболеваемость населения ОРЗ на1000 населения, x	352	228	340	300	196	258	237
Заболеваемость пневмонией на 1000 населения, y	64	60	52	48	46	41	32

Решение:

1) Составить расчетную таблицу:

№	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})\cdot(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
1	352	64	79	15	1185	6241	225
2	228	60	-45	11	-495	2025	121
3	340	52	67	3	201	4489	9
4	300	48	27	-1	-27	729	1
5	196	46	-77	-3	231	5929	9
6	258	41	-15	-8	120	225	64
7	237	32	-36	-17	612	1296	289
Сумма	1911	343	0	0	1827	20934	718
Среднее	273	49					

2) Вычислить коэффициент корреляции: $r_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}} = \frac{1827}{\sqrt{20934\cdot 718}} = 0{,}47$.

3) Проанализировать полученный результат: связь между рассматриваемыми признаками прямая умеренная.

4) Вычислить среднюю ошибку коэффициента корреляции: $m_r = \pm\frac{1-r_{xy}^2}{\sqrt{n}} = \pm\frac{1-0{,}47^2}{\sqrt{7}} = 0{,}3$,

коэффициент корреляции не является достоверным, т.к. не превышает свою среднюю ошибку в три раза.

При анализе клинических и фармацевтических явлений часто используются следующие *непараметрические* коэффициенты связи:

- ранговой корреляции Спирмена;
- «τ» (тау) Кендалла;
- ассоциации Юла;
- контингенции Пирсона;
- сопряженности Чупрова;

- «γ» (гамма) и др.

Рассмотрим *коэффициент ранговой корреляции,* который был разработан и предложен для проведения корреляционного анализа в 1904 г. Ч.Э. Спирменом, английским психологом, профессором Лондонского и Честерфилдского университета.

Коэффициент ранговой корреляции - это коэффициент, который измеряет связь между рангами данной варианты по разным признакам.

Коэффициент ранговой корреляции Спирмена используется для определения тесноты связей между количественными, так и между качественными признаками при условии, если их значения упорядочить по степени убывания или возрастания признака.

Коэффициент ранговой корреляции Спирмена: $\rho = 1 - \frac{6}{n^3 - n}\sum_{i=1}^{n}\left(r_{x_i} - r_{y_i}\right)^2$, где n - объем совокупности, $r_{xi} - r_{yi}$ - разность между рангами i-го объекта.

Качественную характеристику тесноты связи коэффициента ранговой корреляции, как и других коэффициентов корреляции, можно оценить по шкале Чеддока.

Коэффициент ранговой корреляции Спирмена применяется в случае, если объем выборки «n» удовлетворяет неравенству $5 \le n \le 40$.

Пример 7.2. В одном населенном пункте зарегистрировано наличие хронической эпидемии дизентерии Флекснера. Предварительный анализ и лабораторные исследования показали, что в питьевой воде водопроводной сети наблюдаются частые «проскоки» нестандартных проб по бактериологическим показателям (фактор риска). Необходимо проверить гипотезу о наличии связи между этими двумя признаками.

Месяц	Число больных дизентерией (x)	Доля нестандартных проб воды (y)
Январь	10	0
Февраль	9	0,5
Март	2	1,1
Апрель	7	2,0
Май	6	1,8
Июнь	11	2,9
Июль	26	6,7
Август	32	4,5
Сентябрь	46	8,7
Октябрь	38	7,1
Ноябрь	8	3,2
Декабрь	5	0

Решение:

Составить расчетную таблицу:

№	x	y	r_x	r_y	$r_x - r_y$	$r_x - r_y$□^{2}
1	2	0	7	1,5	5,5	30,25
2	9	0,5	6	3	3	9
3	2	1,1	1	4	-3	9
4	7	2,0	4	6	-2	4
5	6	1,8	3	5	-2	4
6	11	2,9	8	7	1	1
7	26	6,7	9	10	-1	1

8	32	4,5	10	9	1	1
9	46	8,7	11	12	-1	1
10	38	7,1	12	11	1	1
11	8	3,2	5	8	-3	9
12	5	0	2	1,5	0,5	0,25
Сумма						70,5

$$\rho = 1 - \frac{6}{n^3 - n}\sum_{i=1}^{n}\left(r_{x_i} - r_{y_i}\right)^2 = 1 - \frac{6}{12^3 - 12}\cdot 70.5 \approx 0,75.$$

Вычислить коэффициент корреляции:

Проанализировать полученный результат: связь между рассматриваемыми признаками прямая высокая.

$$m_r = \pm\frac{1 - r_{xy}^2}{\sqrt{n}} = \pm\frac{1 - 0,75^2}{\sqrt{12}} \approx 0,12,$$

Вычислить среднюю ошибку коэффициента корреляции:

коэффициент корреляции является достоверным, т.к. превышает свою среднюю ошибку более чем в три раза.

4. Иллюстративные материалы: презентация, слайды.

5. Литература:

- Основная:

1. Койчубеков Б. К. Биостатистика. уч. пособие/ Б.К. Койчубеков.- Алматы: Эверо, 2016.
2. Койчубеков Б.К. Биостатистика: учебное пособие: Алматы.- Эверо, 2014

- Дополнительная:

1. Биостатистика в примерах и задачах: учеб.-методическое пособие/ Б.К. Койчубеков [и др.].- Алматы: Эверо, 2012.
2. Боровиков В.П. STATISTICA. Искусство анализа данных на компьютере: для профессионалов / В. Боровиков. - СПб.: Питер, 2014. - 688 с.
3. Юнкеров В.И., Григорьев С.Г. Математико-статистическая обработка данных медицинских исследований. – СПб.: ВМсдА, 2016 - 266 с.

- Электронные ресурсы:

1. Биостатистика [Электронный ресурс]: учебник/ К.Ж. Кудабаев [и др.].- Электрон. текстовые дан. (85,7Мб).- Шымкент: ЮКГФА, 2015.- 187с. эл. опт. диск (CD-ROM)
2. http://matstats.ru/

6. Контрольные вопросы:

1. Зачем в эпидемиологическом анализе используется корреляционный анализ?
2. В каких пределах изменяется коэффициент корреляции?
3. Зачем нужны диаграммы рассеяния?
4. По какой формуле рассчитывается парный коэффициент корреляции Пирсона?
5. Как определяется достоверность коэффициента корреляции?
6. В каких случаях используется ранговый коэффициент корреляции Спирмена?
7. По какой формуле рассчитывается ранговый коэффициент корреляции Спирмена?