ЛЕКЦИОННЫЙ КОМПЛЕКС

Дисциплина: Биостатистика

Код дисциплины: Biostat 2203

ОП: 6В10111 «Общественное здоровье»

Объем учебных часов (кредитов): 150/5

Курс и семестр изучения: 2/3

Объем лекций: 10

Шымкент, 2024 г.

Лекционный комплекс разработан в соответствии с МУП «Информатизация и цифровизация здравоохранения» и обсужден на заседании кафедры.

Протокол №11 от «30» 05 2024 г.

Зав.кафедрой _____ Иванова М.Б.

ЛЕКЦИЯ №1

1. **Тема: Введение. Основы биостатистики.**
2. **Цель:** сформировать у студентов основное представление о дисциплине «Биостатистика», ее предмете, задачах и истории становления. Ознакомить студентов с типами данных, методами их сбора и графическим представлением, а также с видами измерительных шкал, понятиями надежности и достоверности измерения.
3. **Тезисы лекции:**

Статистика - это общественная наука, изучающая количественную сторону массовых общественных явлений в неразрывной связи их с качественной стороной.

В статистике свойство объектов или явлений, которое может быть наблюдаемо или измерено, называется *признаком*.

Статистика, изучающая вопросы, связанные с биологией, медициной, фармацией, гигиеной и здравоохранением, называется *биостатистикой*.

Роль биостатистики в практической и научной работе врача, медсестры, фармацевта велика.

Биостатистика применяет различные методы: сбор данных, их обобщение, анализ и подведение итогов, основанных на полученных наблюдениях.

Статистический анализ помогает добывать информацию из данных и оценивать качество этой информации.

Задачи биостатистики:

- количественное представление биологических 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
- применение классических статистических методов для медицинских анализов;
- применение современных статистических методов для медицинских анализов;
- подготовка новых методов для медицинских анализов.

Основы биометрии начинаются с Фрэнсиса Гальтона (1822—1911 гг.).

Первоначально Гальтон готовился стать врачом. Однако, обучаясь в Кембриджском университете, он увлекся естествознанием, метеорологией, антропологией, наследственностью и теорией эволюции.

его книге, посвященной природной наследственности, изданной в 1889 г. им впервые было введено в употребление слово *biometry.*

это же время он разработал основы корреляционного анализа. Гальтон заложил основы новой науки и дал ей имя. Однако превратил её в научную дисциплину математик Карл Пирсон (1857—1936 гг.) (рисунок 1.1, b). В 1884 г. Пирсон получил кафедру прикладной математики в Лондонском университете,

в 1889 г. познакомился с Гальтоном и его работами.

Большую роль в жизни Пирсона сыграл английский зоолог, биометрик, первый организатор журнала «Биометрика» В.Уэлдон. Помогая Уэлдону в анализе зоологических данных, Пирсон ввёл в 1893 г. понятие среднего квадратического отклонения и коэффициента вариации.

Пытаясь математически оформить теорию наследственности Гальтона, Пирсон в 1898 г. разработал основы множественной регрессии.

1903 г. Пирсон разработал основы теории сопряженности признаков, а в 1905 г. опубликовал основы нелинейной корреляции и регрессии.

Следующий этап развития биометрии связан с именем великого английского статистика Рональда Фишера (1890—1962 гг.) (рисунок 1.1, с).

Во время обучения в Кембриджском университете Фишер познакомился с трудами Г. Менделя и К. Пирсона.

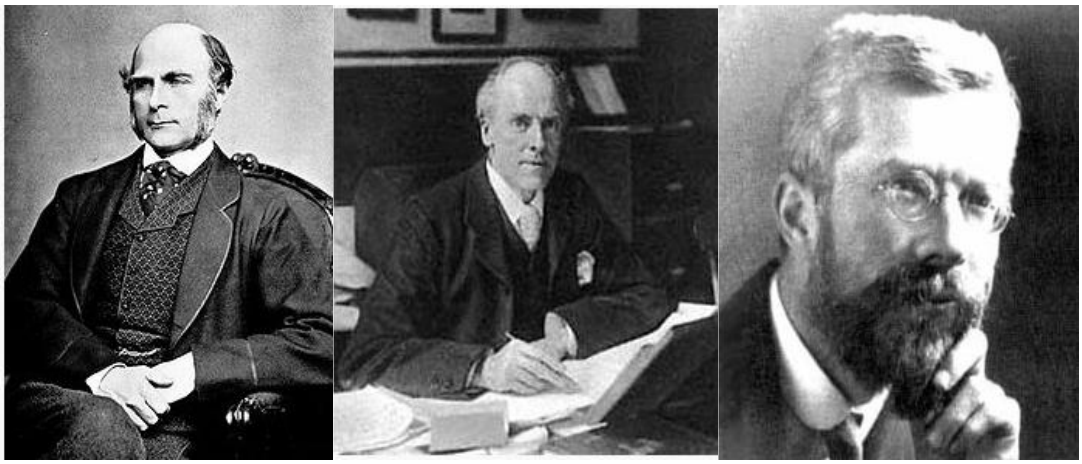
1913—1915 гг. Фишер работал статистиком на предприятии, а в 1915 — 1919 гг. преподавал физику и математику в средней школе.

1919 по 1933 г. Фишер работал статистиком на опытной сельскохозяйственной станции.

Затем, по 1943 г., Фишер занимал должность профессора в Лондонском университете, а с 1943 г. по 1957 г. заведовал кафедрой генетики в Кембридже.

Он является основоположником теории выборочных распределений, методов дисперсионного и дискриминантного анализа, теории планирования экспериментов, метода максимального правдоподобия и многого другого, что составляет основу современной прикладной статистики и математической генетики.

Основоположник понятия о средних величинах, бельгийский ученый А. Кетле применил статистические методы для решения задач биологии, медицины и социологии.



a b c

Рис.1.1. a – Ф. Гальтон, b – К.Пирсон, c – Р.Фишер

Первым этапом при проведении любого статистического исследования является сбор данных об анализируемом объекте или процессе в виде конкретных значений переменных.

Сбором статистических данных называется процесс получения информации об элементах исследуемой совокупности и их свойствах. Эти данные являются предметом статистической обработки и анализа.

Вторым этапом является анализ типов данных.

Типы данных: количественные, качественные и даты (рисунок 1.2).

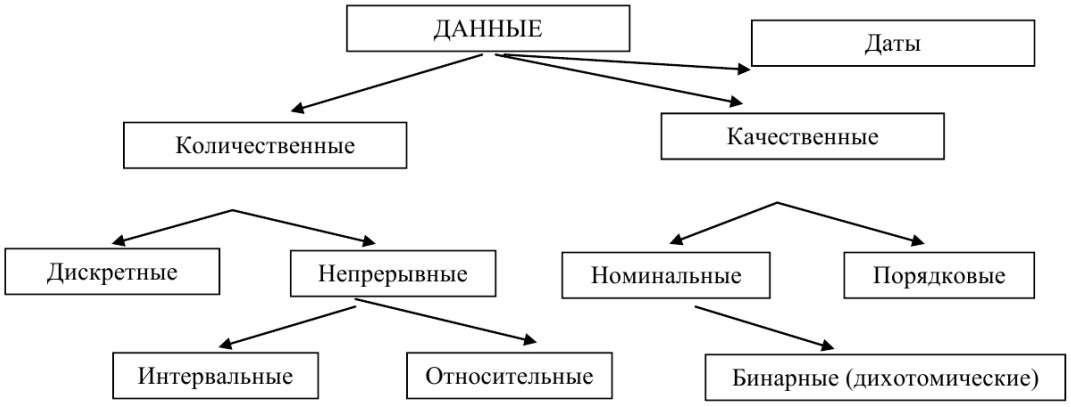


Рисунок 1.2. Классификация типов данных

Основные типы данных делятся на количественные и качественные.

Количественные данные в свою очередь подразделяются на дискретные (прерывные) и непрерывные.

Дискретные данные – количественные данные, которые представлены только в виде целого числа, т.е. не могут иметь дробную часть. Например:
количество детей.

Непрерывные данные – это данные, которые получают при измерении на непрерывной шкале, т.е. теоретически они могут иметь дробную часть. Например: масса тела, рост, артериальное давление и т.д.

Непрерывные данные бывают интервальными и относительными.

Интервальные данные – вид непрерывных данных, которые измеряются в абсолютных величинах, имеющих физический смысл.

Относительные данные – вид непрерывных данных, отражающих долю изменения (увеличения или уменьшения) значения признака по отношению к исходному (или к какому-либо другому) значению этого признака. Эти данные являются безразмерными величинами или выражаются в процентах.

Качественные данные – подразделяются на номинальные и порядковые.

Номинальные данные – вид качественных данных, которые отражают условные коды неизмеряемых категорий (коды диагноза).

Порядковые данные - вид качественных данных, которые отражают условную степень выраженности какого-либо признака (стадии онкологических заболеваний, степени сердечной недостаточности).

Их основное отличие от дискретных количественных данных заключается в отсутствии пропорциональной шкалы для измерения выраженности признака.

Бинарные (дихотомические) данные - особо выделяемый вид качественных данных. Признак такого типа имеет лишь два возможных значения (пол, наличие или отсутствие какого-либо заболевания).

Особым видом данных являются *даты.* Поскольку в ряде случаев бывает необходимо произвести с ними некоторые арифметические действия (вычисление абсолютного периода времени между двумя событиями по датам этих событий).

Иногда выделяют также некоторые особые подтипы данных, являющиеся частными случаями вышеперечисленных типов: ранги, очки, визуальные аналоговые шкалы, цензурированные данные.

Перед тем как проводить углубленный статистический анализ, важно провести

предварительный анализ данных. На этом этапе для сжатия и систематизации набора данных используют графические методы. Это позволяет оценить особенности набора данных и выявить аномалии, т.е. выбрать для дальнейшого анализа подходящие статистические методы.

Дискретные данные могут быть представлены в виде таблицы, столбиковой диаграммы, пиктограммы, круговой диаграммы, точечного рисунка.

Непрерывные данные могут быть представлены в виде группированной выборки, гистограммы, диаграммы «стебель с листьями» или «ящик с усами», кривой Лоренца и т.д.

Смешанные данные могут быть представлены в виде диаграммы рассеяния.

Графические методы представления данных.

График, в котором статистические данные изображаются различными геометрическими фигурами, называется *диаграммой*.

Виды наиболее часто используемых диаграмм:

 Диаграммы, изображающие динамику явления, выраженного в показателях интенсивности, соотношения, наглядности, средних или абсолютных величинах, называются *линейными* (рисунок 1.3).

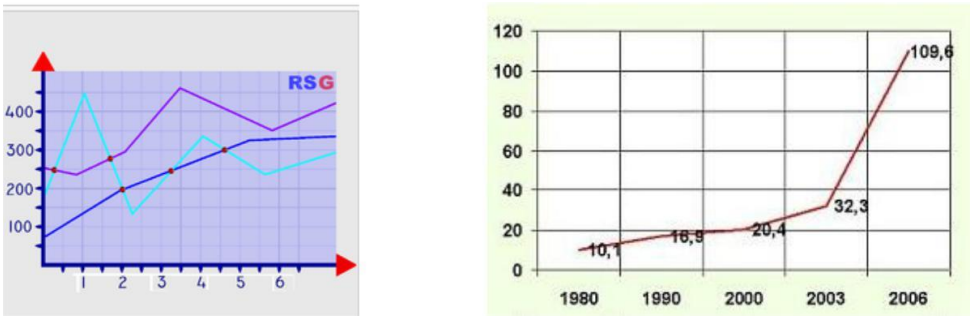


Рисунок 1.3. Линейные диаграммы (график)

Вид линейной диаграммы, применяемой для изображения динамики явления за замкнутый цикл времени (сутки, неделя, месяц, год), называется *радиальной* (рисунок 1.4).

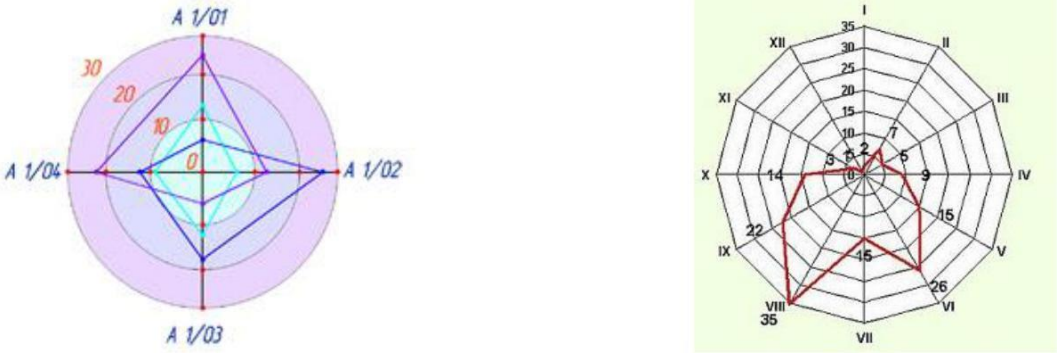
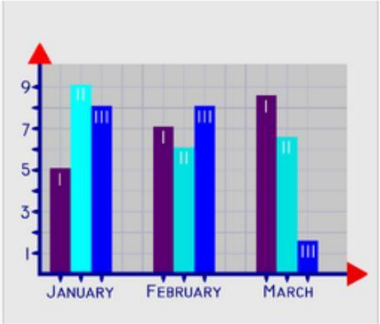


Рисунок 1.4. Радиальные диаграммы

 Диаграммы, изображающие динамику или статику явления в соответствии с избранным масштабом, называются *столбиковыми* (рисунок 1.5).

 Диаграммы, изображающие структуру явления, выраженного экстенсивными показателями, и представляющие собой прямоугольник, в котором цветом выделены составляющие его части в соответствии с их удельным весом, называются *внутристолбиковыми* (рисунок 1.6).



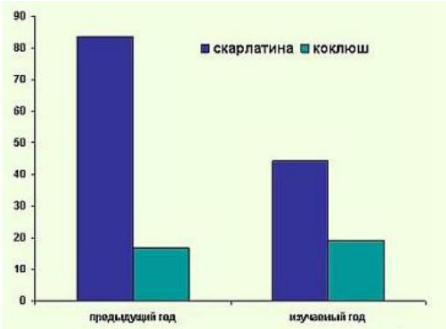


Рисунок 1.5. Столбиковые диаграммы

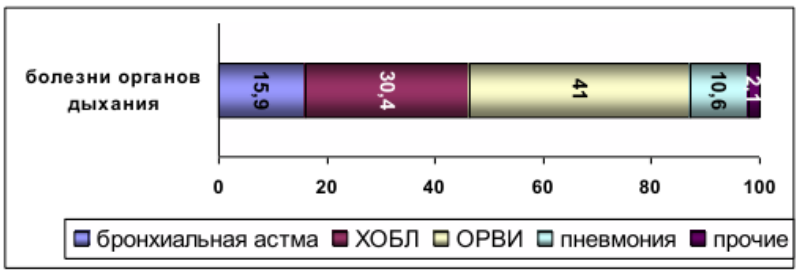


Рисунок 1.6. Внутристолбиковая диаграмма

График, который представляет собой смесь диаграммы и таблицы, эффективен для отображения данных по увеличению порядка величины, называется *графиком «стебель и листья»* (рисунок 1.7).

График, который представляет собой прямоугольник, где две параллельных стороны соответствуют верхнему и нижнему квартилям данных, а линии, начинающиеся в конце прямоугольника, показывают минимальные и максимальные значения, называется *график «ящик с усами»* (рисунок 1.8).

3	1,0	04
665	1,1	39
53	1,2	99
9751	1,3	1135677999
955410	1,4	0148
987665	1,5	00338899
9531100	1,6	0001355
731	1,7	00114569
99843110	1,8	6
654400	1,9	01
6	2,0	
7	2,1	19
10	2,2	

Рисунок 1.7. График «стебель и листья»

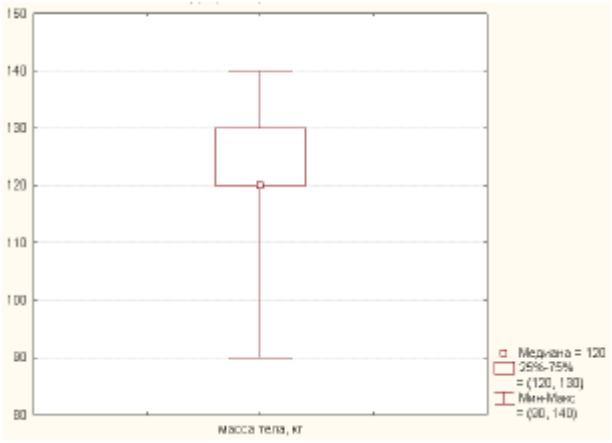


Рисунок 1.8. График *«ящик с усами»*

Измерение – это процедура сравнения объектов по определенным показателям или характеристикам (признакам, атрибутам).

Шкала – необходимый, обязательный элемент измерительной процедуры. Основные

типы измерительных шкал, применяемые в медико-биологических исследованиях:

 номинальная или *шкала наименований* используетя 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* – 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 – показывает «размах» отдельных измерений признака (время, шкала температур, тестовые баллы);

 шкала отношений – выявляет соотношение измеренных значений признака (рост, вес, время реакции, количество выполненных заданий теста).

В процессе измерения возникает вопрос его *надежности* и *достоверности*.

Надежность измерения зависит от:

- правильности (правильно ли выбрана шкала, правильно ли записываются показания, учитываются ли систематические ошибки и др.);
- устойчивости (совпадают ли результаты при повторных измерениях);
- обоснованности (измерено именно заданное свойство объекта, а не другое, на него похожее).

Достоверность измерения характеризует точность измерений величины по отношению к тому, что существует в реальности.

Главное направление проверки достоверности измерений заключается в получении информации из различных источников.

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: 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)

6. Контрольные вопросы:

1. Что такое «биостатистика»?
2. Какова роль ученых Ф. Гальтона, К. Пирсона, Р. Фишера в развитии биометрики?
3. Что называется сбором статистических данных?
4. Какие типы статистических данных Вы знаете?
5. Какие типы измерительных шкал применяются в медико-биологических исследованиях?

ЛЕКЦИЯ №2

1. Тема: Описательная статистика

2. Цель: Ознакомить студентов с понятиями генеральной и выборочной совокупностей, а также с процедурой оценки параметров совокупностей.

3. Тезисы лекции:

Статистическая совокупность – совокупность однородных по какому-либо признаку объектов, ограниченных пространством и временем (число детей, родившихся в стране в течение определенного года; число жителей одного города; число больных онкозаболеваниями в данной стране и т.д.).

При медико-биологических, клинических, фармацевтических и других исследованиях в распоряжении исследователя практически никогда нет полной группы объектов, т.е. невозможно провести сплошное наблюдение, поэтому для исследования используют *выборочный метод*.

Выборочный метод – метод статистического обследования, при котором из статистической совокупности выбирают ограниченное число объектов и их подвергают изучению.

Выборочный метод находит широкое применение в медицине, биологии, здравоохранении и фармации. Например: нет возможности обследовать всех больных с определенной патологией, поэтому обследуют их некоторое число; нет возможности проверить все лекарственные препараты на соответствие стандарту, поэтому проводят их выборочный контроль и т.д.

Генеральная статистическая совокупность - это совокупность, которая состоит из бесконечно большого числа элементов. Например: все больные с данной патологией; все жители данной территории и т.д.

Выборочная совокупность (выборка) - это совокупность, которая состоит из части выбранных элементов наблюдения, способных охарактеризовать всю генеральную совокупность.

Объем совокупности - это общее число элементов наблюдения. Объем генеральной совокупности обозначается «*N*», объем выборочной совокупности – «*n*». Если $n \leq 30$, то выборка считается малой.

Элемент наблюдения - это каждый частный случай явления, которое изучается.

Выборочный метод исследования является единственно возможным в случае бесконечной генеральной совокупности или в случае, когда исследование связано с уничтожением наблюдаемых объектов (например, проверка лекарственных препаратов). Кроме того, он позволяет существенно экономить затраты ресурсов. Недостатком этого метода является появление ошибок исследования, которые связаны с тем, что изучается только часть объекта.

Главным свойством выборки является *показательность (репрезентативность)*, т.е. ее свойство достаточно хорошо воспроизводить генеральную совокупность. Репрезентативность достигается, если объекты генеральной совокупности имеют одинаковую вероятностью попадания в выборку.

Виды показательности:

качественная – это соответствие признаков элементов наблюдения в выборочной и генеральной совокупностях.

количественная - это достаточное число наблюдений.

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

Статистическое распределение выборки (или *вариационный ряд*) представляет собой таблицу, состоящую из двух столбцов (таблица 2.1).

В первом столбце записываются значения варьирующего признака, называемые *вариантами* и обозначаемые «x_i», а во втором столбце записываются числа, называемые *частотами* и обозначаемые «v_i», показывающие сколько раз встречается каждый вариант.

Таблица 2.1.

Варианты (x_i)	Частоты (v_i)
...	...
...	...
...	...
Всего:	$n=\Sigma\, v_i$

Варьирующие признаки могут иметь дискретный и непрерывный характер. Варианты признаков, которые являются целыми числами, называют *дискретными*. Например, число детей в семье, число пациентов, количество ампул в упаковке и т.д.,

Если варианты вариационного ряда выражены в виде дискретных величин, то такой вариационный ряд называют *дискретным.*

Пример 2.1. В результате отдельных испытаний активности тетрациклина гидрохлорида были получены значения x_i (в *ED/мг*): 925, 940, 760, 905, 995, 965, 940, 925, 940, 940, 905.

Располагая значения активности и частоты в порядке возрастания, получим дискретный вариационный ряд (таблица 2.2):

Таблица 2.2.

Варианты (x_i)	Частоты (v_i)
760	1
905	2
925	2
940	3
965	1
995	1
Всего:	n=10

Дискретный вариационный ряд можно представить графически в виде многоугольника, называемого *полигоном* (рисунок 2.1).

Варианты признаков, которые могут принимать любые значения в определенном интервале, называют *непрерывными*. Например, артериальное давление, рост, вес, заработная плата и др.

Для непрерывных признаков строятся *интервальные* вариационные ряды. Минимальное значение признака в заданном интервале называют нижней границей интервала, а максимальное – верхней.

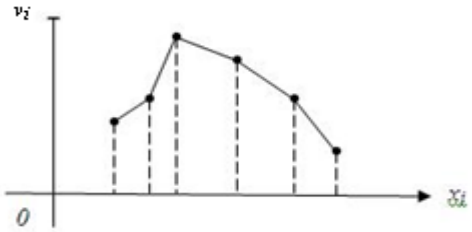


Рисунок 2.1. Полигон

Пример 2.2. В таблице 2.3 приведен интервальный вариационный ряд роста (x_i) мужчин.

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

Число интервалов для интервальных вариационных рядов определяется по формуле Стерджеса:

$$k=1+3{,}322\lg n, \qquad (2.1)$$

где n – объем выборки.

Для вычисления величины интервала используется формула:

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

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

За начало первого интервала берётся величина:

$$x_{нач}=x_{min}-0{,}5h. \qquad (2.3)$$

Интервальный вариационный ряд представляется графически в виде ступенчатой фигуры, называемой *гистограммой* (рисунок 2.2).

Интервальный вариационный ряд представляется графически в виде ступенчатой фигуры, называемой *гистограммой* (рисунок 2.2).

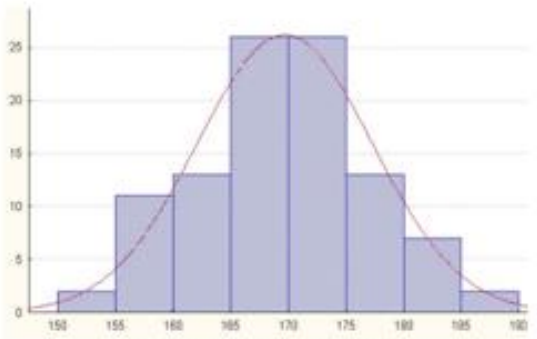


Рисунок 2.2. Гистограмма

При большом числе наблюдений интервальные вариационные ряды строят и для

дискретных признаков.

Вариационный ряд характеризуется *показателями центральной тенденции* и *показателями разнообразия.*

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

Средние величины:

средняя арифметическая простая $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, где *n* - общее число членов ряда;

средняя арифметическая взвешенная $\bar{x} = \frac{\sum_{i=1}^{n} x_i \nu_i}{\sum_{i=1}^{n} \nu_i}$, где ν_i – частоты;

средняя геометрическая простая $\bar{x} = \sqrt[n]{x_1 \cdot x_2 \ldots x_n}$;

средняя квадратическая простая $\bar{x} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}$;

средняя квадратическая взвешенная $\bar{x} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 \nu_i}{\sum_{i=1}^{n} \nu_i}}$.

Структурные величины:

1. *мода (Мо)* – варианта с наибольшей частотой.
2. *медиана (Ме)* – варианта, находящаяся в середине ряда.
3. *квантили* - отдельные равные части, на которые разбивается вариационный ряд:
   квартили – величины, делящие вариационный ряд на четыре равные части;
   квинтили - величины, делящие вариационный ряд на пять равных частей;
   децили - величины, делящие вариационный ряд на десять равных частей;
   процентили - величины, делящие вариационный ряд на сто равных частей

(рисунок 2.3).

3. *Нижний квартиль* (Q_1) или *25-й процентиль* (P_{25}) - это значение случайной величины, ниже которого находится 25% выборки.
4.
5. Номер нижнего квартиля определяется по формуле: $N_{Q_1} = \frac{n+1}{4}$.
6. *Верхний квартиль* (Q_3) или *75-й процентиль* (P_{75}) - это значение случайной величины, выше которого находится 25% выборки.
7.
8. Номер верхнего квартиля определяется по формуле: $N_{Q_3} = \frac{3(n+1)}{4}$.

9. Если номер квартиля получится дробным, то его можно округлить до ближайшего целого.

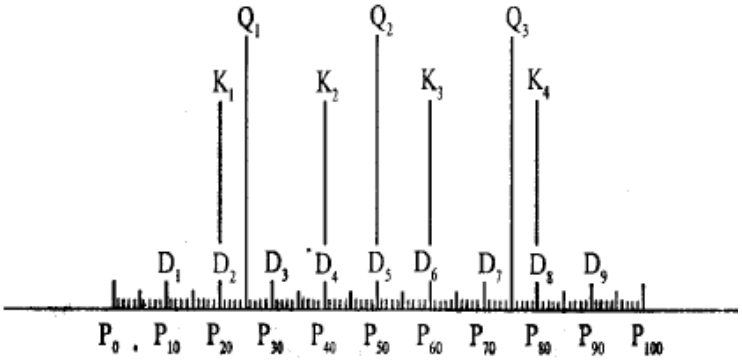


Рисунок 2.3. Структурные характеристики вариационного ряда

К показателям *разнообразия* относятся:

 размах вариационного ряда $R = x_{max} - x_{min}$, где x_{max} , x_{min} – наибольшее и наименьшее значения вариант соответственно;

дисперсия - мера разброса случайной величины от ее среднего значения:

если выборка задана вариационным рядом, то выборочная дисперсия определяется по формуле:

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n};$$

если выборка задана в виде таблицы, то выборочная дисперсия определяется по формуле:

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \nu_i}{\sum_{i=1}^{n} \nu_i};$$

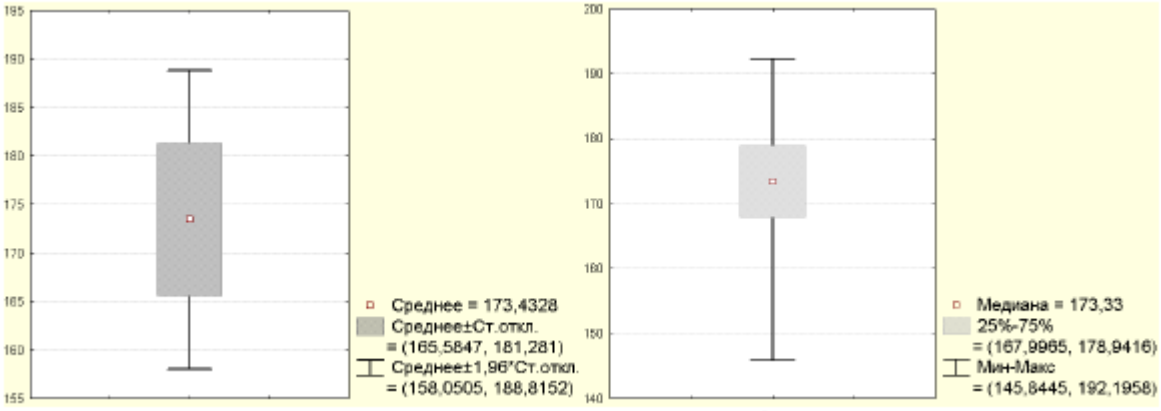
среднее квадратическое отклонение - мера разброса случайной величины от ее среднего значения, выраженная в тех же единицах, что и варианты: $S = \sqrt{S^2}$;

коэффициент вариации - мера разброса случайной величины, выраженная в процентах: $V = \frac{S}{\bar{x}} \cdot 100\%$

Если $V \leq 33\%$, то выборка считается однородной.

Статистические характеристики удобно показывать с помощью графика «ящик с усами».

Например, для выборочных данных о росте 30-летних женщин были построены следующие графики (рисунок 2.4 a, b).



a b

Рисунок 2.4. Отображение статистических характеристик на графике «ящик с усами»

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

На первом графике (рисунок 2.4, a) приведены среднее, минимальное и максимальное значения, а также стандартное отклонение. На втором графике (рисунок 2.4, b) приведены значения медианы, 25-го и 75-го процентилей.

Характеристики генеральной совокупности («*X*» - генеральная средняя, «*D*» - генеральная дисперсия, «*σ*» - среднее квадратическое отклонение) называются *параметрами.*

Параметры обычно неизвестны, и их можно оценить на основе выборочных данных лишь приближенно. Эти приближенные значения называются *оценками параметров генеральной совокупности*.

Оценкой генеральной средней «*X*» является выборочная средняя « $\bar{x}$ ».

Для того чтобы охарактеризовать рассеяние значений изучаемого признака выборки вокруг своего среднего значения « $\bar{x}$ » вводят характеристику, называемую *выборочной дисперсией* «S^2».

Оценкой генеральной дисперсии «*D*» является исправленная выборочная дисперсия «s^2».

Исправленная выборочная дисперсия определяется по формуле: $s^2 = \frac{n}{n-1} S^2$

Оценкой «*σ*» среднего квадратического отклонения генеральной совокупности является «*s*» - исправленное выборочное среднеквадратическое отклонение.

Исправленное выборочное среднеквадратическое отклонение определяется по формуле: $s = \sqrt{s^2} = \sqrt{\frac{n}{n-1} S^2}$.

Средней ошибкой или *средней квадратической ошибкой*, или *стандартной ошибкой среднего* называется величина « S_x », определяемая по формуле: $S_{\bar{x}} = \frac{s}{\sqrt{n}}$.

Эта величина характеризует стандартное отклонение выборочного среднего, рассчитанного по выборке объема «*n*» из генеральной совокупности.

Оценивание некоторого отдельного параметра дает *точечную оценку*.

Интервальной оценкой параметра генеральной совокупности называют интервал, который с заданной вероятностью «*γ*» накрывает истинное значение параметра.

Интервальную оценку называют *доверительным интервалом*, а связанную с ним вероятность «*γ*» – *доверительной вероятностью* или *надежностью* (в медицине и биологии *γ*=0,95).

Доверительный интервал для генеральной средней может быть получен из соотношения: $\bar{x} - t \cdot S_{\bar{x}} \le X \le \bar{x} + t \cdot S_{\bar{x}}$, где $\bar{x}$ - выборочная средняя из «*n*» наблюдений, *t* (*γ; n-1*) - табличная величина, зависящая от «*γ*» и «*n*», S_x - стандартная ошибка среднего.

Одним из важнейших этапов любого медико-биологического исследования это - определение объема рассматриваемой выборки.

Если объем выборки будет недостаточным, то это приведет к увеличению ошибки выборочных характеристик и к неправильным выводам.

Объем выборки зависит от среднего квадратического отклонения изучаемой величины «σ», мощности используемого критерия и уровня значимости «*p*».

На практике для расчета оптимального объема выборки можно использовать простую формулу: $n = \left(\frac{t \cdot \sigma}{\varepsilon}\right)^2$, где *n* - объем выборки, *σ* – среднее квадратическое отклонение, *t* =1,96 – критическое значение стандартного нормального распределения при *p*=0,05 - табличная величина, *ε* =0,01 – заданная точность оценки, согласно ГОСТа.

Если проводятся независимые испытания, в которых событие наступает с неизвестной вероятностью «*p*», то объем выборки определяется по формуле: $n = \frac{t^2}{\varepsilon^2}\,\tilde{p}(1-\tilde{p})$, где $\tilde{p} = \frac{m}{n}$ - оценка вероятности «*p*», «*m*» - число появлений события, «*n*» - общее число испытаний.

Чем больше объем выборки «*n*», тем точнее результат исследования.

4. Иллюстративный материал: презентация, слайды.

6. 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)

6. Контрольные вопросы:

1. Что такое генеральная и выборочная совокупности?
2. Что такое вариационный ряд?
3. Какие показатели вариационного ряда Вы знаете?
4. В чем различие между точечной и интервальной оценками?
5. Как определяется доверительный интервал для генеральной средней?

6. Как определяется необходимый объем выборки?

ЛЕКЦИЯ №3

1. Тема: Основы теории проверки статистических гипотез. Проверка гипотезы о нормальности распределения случайной величины.

2. Цель: Ознакомить студентов с основами теории проверки статистических гипотез.

3. Тезисы лекции:

В прикладных задачах 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*.

Гипотеза – это некоторое предположение о параметрах известных распределений (параметрическая) или о виде неизвестного закона распределения

(непараметрическая) случайных величин, выдвигаемое в качестве предварительного, условного объяснения.

Теория проверки статистических гипотез является основным инструментом доказательной, а не интуитивной медицины.

Задачи медицинских и биологических исследований, для решения которых необходимо сформулировать статистические гипотезы:

анализ соответствия распределения значений признака в изучаемой группе какому-либо определенному закону (анализ соответствия распределения нормальному закону);

сравнение групп по параметрам распределений признака (по средним значениям, дисперсиям).

Например, при проверке статистических гипотез можно получить ответ на следующий вопрос. В двух однородных группах больных гриппом была проведена вакцинация: одной лекарственным средством «*А*», а другой - «*В*», среднее время выздоровления в группах неодинаково. Указывает ли это обстоятельство на то, что одно противогриппозное средство по эффективности превосходит другое или же выявленное различие случайно?

Для решения любой подобной задачи выдвигаются две статистические гипотезы:

нулевая гипотеза H_0 - гипотеза об отсутствии различий между группами, либо об определенных значениях параметров, либо о соответствии распределения нормальному закону;

альтернативная гипотеза H_1 - гипотеза о существовании различий между группами, либо об отличающихся от заданных значениях параметров, либо о несоответствии распределения нормальному закону.

Нулевая гипотеза формулируется таким образом, чтобы она была противоположной той исследовательской (медицинской, биологической) гипотезе, которая послужила поводом для проведения исследования.

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

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

Статистическими критериями называются правила, согласно которым выясняется, соответствует или нет интересующая нас гипотеза опытным данным. Статистические критерии - это наиболее широко применяемые статистические средства.

Значение критерия, которое рассчитано по выборочной совокупности, подчиняющейся определённому закону распределения, называется наблюдаемым.

Множество возможных значений статистического критерия, при которых основная гипотеза принимается, называется областью принятия.

Множество возможных значений статистического критерия, при которых основная гипотеза отвергается, называется *критической областью*.

Точки, разграничивающие критическую область и область принятия гипотезы, называются *критическими точками*.

При проверке статистических гипотез возникают следующие виды ошибок:

- *ошибка первого рода* – это вероятность отвергнуть правильную нулевую гипотезу;
- *ошибка второго рода* – это вероятность принять неправильную нулевую гипотезу.

Уровень значимости - это максимально приемлемая для исследователя вероятность ошибочно отклонить нулевую гипотезу, когда на самом деле она верна, т.е. допускаемая исследователем величина ошибки первого рода.

При иссследованиях в фармации, медицине и биологии используется величина уровня значимости, равная 0,05. При разработке стандартов используют уровень значимости равный 0,01.

Уровень значимости или вероятность ошибки первого рода обозначается через «p», а вероятность ошибки второго рода - через «γ».

Доверительная вероятность (γ) - это вероятность не совершить ошибку первого рода и принять верную гипотезу H_0 ($\gamma=1-p$).

Важнейшей характеристикой любого статистического критерия является его *мощность*. *Мощностью критерия* называется его способность правильно исключать ложную гипотезу. Мощность оценивается вероятностью $1-\gamma$, где γ - вероятность ошибки второго рода.

Схема проверки статистических гипотез:

1. Выдвигаются две гипотезы: основная (нулевая) «H_0» и альтернативная (конкурирующая) «H_1».
2. Задается уровень значимости «p». Статистический вывод никогда не может быть сделан со стопроцентной уверенностью. Всегда допускается риск принятия неправильного решения. При проверке статистических гипотез мерой такого риска является уровень значимости «p».
3. По исходным данным, т.е. по выборке, вычисляется наблюдаемое (эмпирическое, расчетное) значение критерия.
4. По специальным статистическим таблицам определяется табличное (критическое) значение критерия.
5. Путем сравнения наблюдаемых и табличных значений делается вывод о правильности той или иной гипотезы.

В биостатистике часто проверяются гипотезы о виде распределения случайной величины.

Множество биологических и медицинских показателей (показатели физического развития, составляющие плазмы крови и др.), а также ошибки их измерения подчиняются нормальному распределению.

Поэтому важно уметь проверять гипотезы о параметрах нормально распределенных случайных величин.

Все предположения о характере того или иного распределения - являются гипотезами. Поэтому они должны подвергаться статистической проверке с помощью *критериев согласия.* Эти критерии дают возможность определить, когда расхождения между теоретическими и

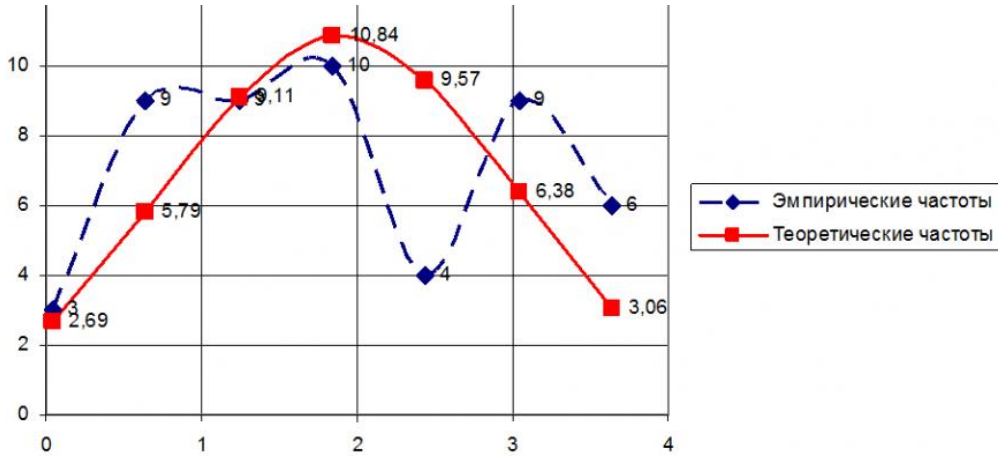
эмпирическими частотами следует признать несущественными, т.е. случайными, а когда – существенными, т.е. неслучайными.

Таким образом, критерии согласия позволяют отвергнуть или подтвердить правильность выдвинутой при выравнивании ряда гипотезы о характере распределения в эмпирическом ряду.

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

Эти критерии применяются в двух случаях:

- для сопоставления расчетного распределения признака с теоретическим распределением (нормальным, показательным, равномерным и т.д.) (рисунок 6.1);
- для сопоставления двух расчетных распределений одного и того же признака.



Сопоставление эмпирических и теоретических частот

5. *Схема применения критерия согласия χ^2-Пирсона:*

1) H_0: случайная величина «X» имеет функцию распределения $F(x)$.
H_1: случайная величина «X» не имеет функцию распределения $F(x)$.

2) p=0,05 - уровень значимости.

3) $$\chi^2_{расч} = \sum_{i=1}^{k} \frac{\left(\nu_i - \nu_i^*\right)^2}{\nu_i^*}, \qquad (6.1)$$

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

4) $\chi^2_{табл}(p; f)$

где $f = k - 1 - r$ - число степеней свободы (табличное значение), k - число групп выборки, r - число параметров предполагаемого распределения (для нормального распределения r=2).

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.

6. *Схема применения критерия согласия Колмогорова - Смирнова:*

1) H_0: случайная величина «*X*» имеет функцию распределения *F(x).*
   H_1: случайная величина «*X*» не имеет функцию распределения *F(x).*
2) p=0,05 - уровень значимости.
3) $\lambda_{расч} = d_{\max}\sqrt{n}$, (6.2)

где $d_{\max} = \max\left|F_n(x) - F(x)\right|$ - 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.

4) $\lambda_{табл}=1{,}36$ (табличное значение при p=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.

7. 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)

6. Контрольные вопросы:

1. Что называется статистической гипотезой?
2. Какие виды статистических гипотез Вы знаете?
3. Что называется ошибкой первого и второго рода?
4. Что называется доверительной вероятностью и уровнем значимости?
5. Какова общая схема проверки статистических гипотез?

ЛЕКЦИЯ №4

1. Тема: Сравнение средних значений признака двух групп.

2. Цель: Ознакомить студентов с *t*-критерием Стьюдента.

3. Тезисы лекции:

В процессе медико-биологических исследований часто возникает проблема сравнения результатов обследования (например, в контрольной и экспериментальной группах или до и после эксперимента).

Для решения этой проблемы существует большое количество статистических критериев. Каждый из них имеет свою специфику, отличаясь друг от друга

(например, типами данных, объемами выборок, количеством сравниваемых выборок, качеством сравниваемых выборок (зависимая и независимая) и др.).

Наиболее популярным из таких критериев является *t*-критерий Стьюдента, который применяется примерно в 30-40% научных медицинских работ (рисунок 4.1).

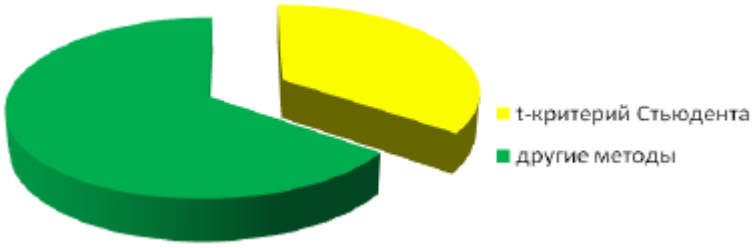


Рисунок 4.1. Соотношение статистических методов, используемых в медицинских научных работах

t-критерий Стьюдента - метод проверки однородности выборок, позволяет принять или отвергнуть гипотезу о равенстве средних двух выборок ($H_0 : \bar{x}_1 = \bar{x}_2$).

Данный критерий был разработан английским химиком Уильямом Госсеттом (1876-1936 гг.) для оценки качества пива в компании «Гиннес».

Статья Госсета вышла в 1908 г. в журнале «Биометрика» под псевдонимом «Student» (Студент).

t-критерий Стьюдента используется:

- при проверке гипотезы о равенстве средних двух *независимых* выборок (*двухвыборочный t-критерий*). В этом случае анализируются контрольная и экспериментальная выборки разных объемов. Например, группа больных сахарным диабетом и группа здоровых людей;
- при проверке гипотезы о равенстве средних двух зависимых выборок (парный t-критерий). В этом случае анализируется одна и та же выборка, но до и после эксперимента. Например, средняя частота пульса у одних и тех же пациентов до и после приема антиаритмического препарата.

Применение критерия Стьюдента возможно, если выполняются следующие два условия:

- рассматриваемые выборки имеют нормальное распределение;
- дисперсии рассматриваемых выборок равны.

Исследователями установлено, что оба условия одновременно выполняются лишь в 4-5% случаев медико-биологических экспериментов.

В то же время анализ диссертаций, научных работ, статей, опубликованных медицинских журналах, показывает, что t-критерий Стьюдента используется при проведении статистических расчетов в 50% работ.

Схема применения двухвыборочного *t*-критерия Стьюдента

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

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) $t_{табл}(p; f)$, где $f = n_1+n_2-2$- число степеней свободы.

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

Если $t_{расч} > t_{табл}$, то «H_0» отвергается.

Критерий Стьюдента применяется в случае малых выборок ($n_{1,2} \le 30$).

Пример 4.1. Если при родах шейка матки долго не раскрывается, то продолжительность родов увеличивается и может возникнуть необходимость кесарева сечения. Ученые решили выяснить, ускоряет ли гель с простагландином E_2 раскрытие шейки матки. В исследование вошло 2 группы рожениц. Роженицам первой группы вводили в шейку матки гель с простагландином E_2, роженицам второй группы вводили гель-плацебо. В обеих группах было по 21 роженице возраст, рост и сроки беременности были примерно одинаковы. Роды в группе, получавшей гель с простагландином E_2, длились в среднем 8,5 часов (стандартное отклонение 4,7 часа), в контрольной группе - 13,9 часа (стандартное отклонение 4,1 часа). Можно ли утверждать, что гель с простагландином E_2 сокращал продолжительность родов?

Решение.

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

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

2) p=0,05.

3) $t_{расч} = \dfrac{13{,}9 - 8{,}5}{\sqrt{(21-1)\cdot 4{,}1^2 + (21-1)\cdot 4{,}7^2}} \cdot \sqrt{\dfrac{21\cdot 21}{21+21}(21+21-2)} \approx 4$.

4) $t_{табл}(0{,}05;40) = 2{,}02$.

5) Т.к. $t_{расч} > t_{табл}$, то «H_0» отвергается, т.е. гель с простагландином E_2 сокращал продолжительность родов.

Схема применения парного критерия Стьюдента:

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

2) p=0,05 - уровень значимости.

3) $t_{расч} = |\bar{d}| \cdot \sqrt{\dfrac{n(n-1)}{\sum_{i=1}^{n} d_i^2 - n\bar{d}^2}}$, где $d = x_i - y_i$ - разности между соответствующими значениями пар переменных, $\bar{d}$ - среднее значение этих разностей, n - объем выборки.

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

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

Если $t_{расч} > t_{табл}$, то «H_0» отвергается.

Иногда сравнение выборочных средних проводится по следующей формуле:

$$t_{расч} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{m_1^2 + m_2^2}},$$

где $\bar{x}_1, \bar{x}_2$ - сравниваемые средние величины, m_1 и m_2 - ошибки сравниваемых средних величин.

Если $t_{расч} \geq 2$, то «H_0» отвергается.

Пример 4.2. Для оценки эффективности нового гипогликемического препарата были проведены измерения уровня глюкозы в крови пациентов, страдающих сахарным диабетом, до и после приема препарата:

№ пациента	Уровень глюкозы в крови, моль/л	
	до приема препарата	после приема препарата
1	9,6	5,7
2	8,1	4,2
3	8,8	6,4
4	7,9	5,5
5	9,2	5,3
6	8,0	5,2
7	8,4	5,1
8	10,1	5,9
9	7,8	7,5
10	8,1	5,0
Среднее значение	8,6	5,6

Можно ли считать, что после приема препарата уровень глюкозы в крови пациентов снижается?

Решение.

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

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

2) p=0,05- уровень значимости

№ пациента	Уровень глюкозы в крови, моль/л		Разность значений $d = x_i - y_i$	d^2
	до приема препарата	после приема препарата		
1	9,6	5,7	3,9	15,21
2	8,1	4,2	3,9	15,21
3	8,8	6,4	2,4	5,76
4	7,9	5,5	2,4	5,76
5	9,2	5,3	3,9	15,21
6	8,0	5,2	2,8	7,84
7	8,4	5,1	3,3	10,89
8	10,1	5,9	4,2	17,64
9	7,8	7,5	0,3	0,09
10	8,1	5,0	3,1	9,61
Сумма			30,2	103,22

$$\bar{d} = \frac{30,2}{10} = 3,02$$

3) $t_{расч} = 3,02 \cdot \sqrt{\frac{10(10-1)}{103,22 - 10 \cdot 3,02^2}} \approx 8,3.$

4) $t_{табл}(0,05;9) = 2,26$

Т.к. $t_{расч} > t_{табл}$, то «H_0» отвергается, т.е. уровень глюкозы в крови после приема препарата снизился, значит новое средство эффективно.

Пример 4.3. У студентов - медиков проводилось исследование пульса до и после сдачи экзамена. Частота пульса до экзамена составила 98,8±4,0, а после экзамена 84,0±5,0.

Можно ли считать, что после экзамена частота пульса снижается и приближается к норме?

Решение.

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

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

2) p=0,05.

3) $t_{расч} = \frac{98,8 - 84}{\sqrt{4^2 + 5^2}} \approx 2,3.$

Т.к. $t_{расч} > 2$, то «H_0» отвергается, т.е. после экзамена частота пульса снижается и приближается к норме.

4. Иллюстративный материал: презентация, слайды.

8. 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)

6. Контрольные вопросы:

1. Почему *t*-критерий Стьюдента пользуется большой популярностью при статистическом анализе медико-биологических данных?
2. Какие условия должны выполняться при использовании *t*-критерия Стьюдента?
3. В чем разница между двухвыборочным и парным критериями Стьюдента?

ЛЕКЦИЯ №5

1. Тема: Непараметрическая альтернатива.

2. Цель: Ознакомить студентов с некоторыми непараметрическими критериями, используемыми для проверки гипотезы об однородности средних.

3. Тезисы лекции:

Статистические критерии делятся на параметрические и непараметрические.

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* значений признака. Например, критерий Манна-Уитни, критерий Уилкоксона, критерий знаков и др.).

Ранг - порядковый номер значения признака.

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

Аналогом двухвыборочного *t*-критерия Стьюдента является *U*-критерий Манна-Уитни. Аналогом парного *t*-критерия Стьюдента является *W*-критерий Уилкоксона.

U-критерий Манна-Уитни - непараметрический статистический критерий, используемый для сравнения двух *независимых* выборок по уровню какого-либо признака, измеренного количественно.

Данный метод выявления различий между выборками был предложен в 1945 г. американским химиком и статистиком Ф. Уилкоксоном.

В 1947 г. метод был переработан и расширен математиками Х.Б. Манном и Д.Р. Уитни.

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.

W-критерий Уилкоксона - непараметрический статистический критерий, используемый для сравнения двух *зависимых* выборок по уровню какого-либо признака, измеренного количественно.

Критерий Уилкоксона применяется в случае, если объем выборки *«n»* удовлетворяет неравенству $5 \leq n \leq 50$.

Схема применения критерия Манна-Уитни:

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

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

2) p=0,05 - уровень значимости.

3) Из двух сравниваемых выборок составляется единый ранжированный ряд, который затем опять разделяется на два, состоящих из единиц первой и второй выборок, при этом отмечаются значения рангов для каждой единицы.

Подсчитываются 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) $U_{табл}(p; n_1; n_2)$

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

Если $U_{расч} \le U_{табл}$, то «H_0» отвергается.

Пример 5.1. Исследуется эффективность препарата, позволяющего сбросить лишнюю массу больным, страдающим ожирением. При этом группе добровольцев предписана определенная диета. Через месяц, с целью проверки соблюдения диеты и регулярного приема препарата, фиксируется величина потерянной массы (кг). Для проведения эксперимента отобрана группа из 8 человек. 3 из них получали исследуемый препарат (экспериментальная группа), а 5 получали плацебо (контрольная группа). Отбор 3 испытуемых из 8 в экспериментальную группу осуществлялся случайным образом. Все участники эксперимента считали, что принимают препарат.

Экспериментальная группа	Контрольная группа
Потерянная масса, кг	Потерянная масса, кг
6,2	4,0
3,0	-0,5
3,9	3,3
	1,5
	3,0

Решение.

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

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

2) p=0,05 - уровень значимости.

3) Составим единый ряд.

4)

Потерянная масса, кг	6,2	3,0	3,9	4,0	-0,5	3,3	1,5	3,0
Ранг	8	3,5	6	7	1	5	2	3,5

Разделим единый ранжированный ряд на два, состоящих из единиц первой и второй выборок.

Экспериментальная группа		Контрольная группа	
Потерянная масса, кг	Ранг	Потерянная масса, кг	Ранг
6,2	8	4,0	7
3,0	3,5	-0,5	1
3,9	6	3,3	5
		1,5	2
		3,0	3,5
	T_1=17,5		T_2=18,5

T_1 и T_2 – суммы рангов; $T_1 < T_2$, значит $T_2 = T_x$, $n_x = n_2 = 5$.

$$U_{расч} = 3 \cdot 5 + \frac{5 \cdot (5+1)}{2} - 18{,}5 = 11{,}5.$$

5) $U_{табл}(0{,}05;\ 3;\ 5)=1$.

6) $U_{расч} > U_{табл}$, то «H_0» принимается, т.е. препарат неэффективен.

Схема применения критерия Уилкоксона:

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

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

2) $p \approx 0{,}05$ - уровень значимости.

3) Вычисляется разность между индивидуальными значениями во втором и первом замерах.

Абсолютные величины разностей упорядочиваются по рангу (меньшему значению присваивается меньший ранг).

Каждому рангу ставится знак «+» или «–» в зависимости от знака соответствующей ему разности, получаются знаковые ранги.

Расчетное значение критерия $W_{расч}$ определяется из суммы знаковых рангов.

4) $W_{табл}(p;\ n)$, где n - объем выборки.

5) Если $|W_{расч}| \leq W_{табл}$, то «H_0» принимается.

Если $|W_{расч}| > W_{табл}$, то «H_0» отвергается.

Пример 5.2. Проверить есть ли разница в содержании сахара в крови натощак до работы и через три часа после работы у 12 работающих на ультразвуковых установках.

№	Содержание сахара до работы	Содержание сахара после работы
1	112	54
2	82	67
3	101	96
4	72	59
5	79	79
6	82	76
7	64	66
8	70	66
9	88	48
10	81	50
11	66	61
12	88	61

Решение.

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

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

2) $p \approx 0{,}05$ - уровень значимости.

3)

№	x	y	x-y	$\|x-y\|$	$r_{\|x-y\|}$	Знаковые ранги
1	112	54	58	58	12	12
2	82	67	15	15	8	8
3	101	96	5	5	4,5	4,5
4	72	59	13	13	7	7
5	79	79	0	0	1	1
6	82	76	6	6	6	6
7	64	66	-2	2	2	-2
8	70	66	4	4	3	3
9	88	48	40	40	11	11
10	81	50	31	31	10	10
11	66	61	5	5	4,5	4,5
12	88	61	27	27	9	9
				Сумма знаковых рангов		$W_{расч}$ =74

4) $W_{табл}$ (0,052; 12)=50.

5) $|W_{расч}| > W_{табл}$, то «H_0» отвергается, значит есть разница в содержании сахара в крови у работников до и после работы.

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: 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)

6. Контрольные вопросы:

1. В чем заключается разница между параметрическими и непараметрическими статистическими критериями?
2. Почему критерий Манна-Уитни называют аналогом двухвыборочного t-критерия Стьюдента?
3. Почему критерий Уилкоксона называют аналогом парного t-критерия Стьюдента?

ЛЕКЦИЯ №6

1. **Тема: Однофакторный дисперсионный анализ.**
2. **Цель:** Ознакомить студентов с основами дисперсионного анализа.
3. **Тезисы лекции:**

Дисперсионным анализом называют группу статистических методов, разработанных английским математиком и генетиком Р. Фишером в 20-х годах XX-го века для ряда экспериментальных задач биологии и сельского хозяйства.

Постановка задачи. Пусть даны генеральные совокупности $X_1, X_2, \ldots, X_k$, где:

- все «*k*» генеральных совокупностей распределены нормально;
- дисперсии всех генеральных совокупностей одинаковы.

При этих условиях и заданном уровне значимости «*p*» требуется проверить нулевую гипотезу о равенстве выборочных средних, т.е. $H_0: \bar{x}_1 = \bar{x}_2 = \ldots = \bar{x}_k$.

Каждая из генеральных совокупностей подвержена влиянию одного или нескольких *факторов*, которые могут изменять их средние значения.

Фактором называется показатель, который оказывает влияние на конечный результат.

Конкретную реализацию фактора называют *уровнем фактора*.

Значение измеряемого признака называют *откликом* на фактор.

Например, некоторое количество больных гипертонией разбиты случайным образом на «*k*» групп, каждой из которых назначен прием определенного лекарства. В результате контролируется среднее значение показателя изменения артериального давления.

В данном примере:

- значения показателя в «*i*»-ой группе, состоящей из «n_i» больных – это «*i*» - выборка объема «n_i»;
- лекарство - это *фактор,* влияющий на величину контролируемого показателя;
- показатель изменения артериального давления - это *отклик* на воздействие фактора.

Предполагается, что по группам принимаемые лекарства различаются либо видом, либо дозой, либо еще каким-либо образом. Тогда воздействующий фактор подразделяется на некоторые составляющие, называемые *уровнями фактора*.

Если фактор оказывает воздействие на величину отклика, то нулевая гипотеза о равенстве средних $H_0: \bar{x}_1 = \bar{x}_2 = \ldots = \bar{x}_k$ отвергается.

В зависимости от количества изучаемых факторов дисперсионный анализ делится на *однофакторный* и *многофакторный*.

В примере с изменением артериального давления можно исследовать:

- фактор времени года (уровни: зима, весна, лето, осень);
- фактор места эксперимента (уровни: лечение в стационаре или дома);
- фактор режима (уровни: постельный, обычный или регулярные пешие прогулки на свежем воздухе) и т.п.

Выборочные данные для однофакторного дисперсионного анализа оформляют в виде таблицы (таблица 6.1).

Номер испытания	Уровень фактора A			
	A_1	A_2	...	A_k
1	x_{11}	x_{12}	...	x_{1k}
2	x_{21}	x_{22}	...	x_{2k}
...	...	...	...	...
R	x_{r1}	x_{r2}	...	x_{rk}
Групповая средняя	$\bar{x}_{гр1}$	$\bar{x}_{гр2}$	...	$\bar{x}_{грk}$

Основная цель дисперсионного анализа состоит в разбиении выборочной дисперсии на две компоненты:

- первая – это *факторная дисперсия,* она соответствует влиянию фактора на изменчивость средних значений;
- вторая – это *остаточная дисперсия,* она обусловлена случайными причинами и не влияет на изменчивость средних значений.

Для численной оценки влияния исследуемого фактора используют сравнение этих компонент с помощью *F*-критерия Фишера.

Факторная дисперсия ($S_{факт}^2$) – это дисперсия, которая соответствует влиянию фактора на изменение средних значений выборки:

$$S_{факт}^2 = \frac{SS_{фатк}}{k-1} = \frac{r\sum_{j=1}^{k}\left(\bar{x}_{гр\,j} - \bar{x}\right)^2}{k-1},$$

где $SS_{факт}$ - факторная сумма квадратов отклонений, k - количество уровней фактора, r - количество значений в каждой группе, $\bar{x}$ - общая средняя, $\bar{x}_{гр}$ - групповая средняя.

Остаточная дисперсия ($S_{ост}^2$) – это дисперсия, возникающая по случайными причинами и не влияющая на изменение средних значений выборки:

$$s_{ост}^2 = \frac{SS_{ост}}{k(r-1)} = \frac{\sum_{i=1}^{r}\left(x_{i1} - \bar{x}_{гр1}\right)^2 + \sum_{i=1}^{r}\left(x_{i2} - \bar{x}_{гр2}\right)^2 + \ldots + \sum_{i=1}^{r}\left(x_{ik} - \bar{x}_{грk}\right)^2}{k(r-1)},$$

где $SS_{ост}$ - остаточная сумма квадратов отклонений.

Общая дисперсия ($S_{общ}^2$) – это сумма факторной и остаточной дисперсий:

$$S_{общ}^2 = \frac{SS_{общ}}{n-1} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{r}\left(x_{ij} - \bar{x}\right)^2}{n-1},$$ где $SS_{общ} = SS_{факт} + SS_{ост}$.

Схема применения однофакторного дисперсионного анализа:

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) p=0,05- уровень значимости.
3) $F_{расч} = \frac{S^2_{факт}}{S^2_{ост}}$.
4) $F_{табл}(p, f_1, f_2)$, где $f_1=k-1$, $f_2=k(r-1)$ – число степеней свободы (табличные значения), k - количество уровней фактора, r - количество значений в каждой группе.
5) Если $F_{расч} \leq F_{табл}$, то «H_0» принимается.

Если $F_{расч} > F_{табл}$, то «H_0» отвергается.

Пример 6.1. Среди взрослого населения определенной возрастной категории фиксировалось число заболеваний дыхательных путей за два года. Цель исследования – статистическое доказательство влияния курения на заболеваемость дыхательных путей. Случайным образом были отобраны 3 группы по 4 человека каждая, из них: 1 группа - некурящие, 2 группа - стаж курильщика - до 5 лет, 3 группа - стаж курильщика более 5 лет. Таким образом, исследуемый фактор «A» - курение, уровни фактора, A_1, A_2, A_3 - стаж курильщика. Отклик на фактор курения - число заболеваний дыхательных путей. Были получены 12 значений числа заболеваний – x_{ij} , где j - номер уровня фактора (j=1, 2, 3), i - номер элемента в соответствующей выборке (группе), i =1, 2, 3, 4:

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43} \end{bmatrix} = \begin{bmatrix} 1 & 3 & 3 \\ 0 & 2 & 4 \\ 1 & 2 & 5 \\ 2 & 1 & 3 \end{bmatrix}.$$

Предполагаем, что $\{x_{ij}\}$ - выборка из нормальной генеральной совокупности.
Все данные необходимо занести в таблицу:

Номер испытания	Уровень фактора «А»		
	A_1	A_2	A_3
1	1	3	3
2	0	2	4
3	1	2	5
4	2	1	3
Групповая средняя	4/4=1	8/4=2	15/4=3,75

Решение.

H_0: $\bar{x}_1 = \bar{x}_2 = \bar{x}_3$.

1) H_1: $\bar{x}_1 \neq \bar{x}_2 \neq \bar{x}_3$.

2) p=0,05 - уровень значимости.

3) Вычисляется:

3.1) Общая средняя: $\bar{x} = \dfrac{\bar{x}_1 + \bar{x}_2 + \bar{x}_3}{3} = \dfrac{1+2+3,75}{3} = 2,25$.

3.2) Факторная сумма квадратов отклонений:

$$SS_{факт} = r\sum_{j=1}^{3}\left(\bar{x}_{гр\ j} - \bar{x}\right)^2 = 4\cdot\left[(1-2,25)^2 + (2-2,25)^2 + (3,75-2,25)^2\right] = 15,5\,.$$

3.3) Остаточная сумма квадратов отклонений:

$$SS_{ост} = \sum_{i=1}^{4}\left(x_{i1} - \bar{x}_{гр1}\right)^2 + \sum_{i=1}^{4}\left(x_{i2} - \bar{x}_{гр2}\right)^2 + \sum_{i=1}^{4}\left(x_{ik} - \bar{x}_{гр3}\right)^2 =$$

$$= \left[(1-1)^2 + (0-1)^2 + (1-1)^2 + (2-1)^2\right] + \left[(3-2)^2 + (2-2)^2 + (2-2)^2 + (1-2)^2\right] +$$

$$+ \left[(3-3,75)^2 + (4-3,75)^2 + (5-3,75)^2 + (3-3,75)^2\right] = 6,75.$$

3.4) Факторная дисперсия: $S^2_{факт} = \frac{SS_{фатк}}{k-1} = \frac{15,5}{3-1} = 7,75$.

3.5) Остаточная дисперсия: $S^2_{ост} = \frac{SS_{ост}}{k(r-1)} = \frac{6,75}{3(4-1)} = 0,75$.

3.6) $F_{набл} = \frac{S^2{}_{факт}}{S^2{}_{ост}} = \frac{7,75}{0,75} = 10,3$.

4) $F_{табл}(0,05; 2; 9)=4,26$.

5) $F_{набл} > F_{табл}$, то «H_0» отвергается, значит фактор курения значимо влияет на заболеваемость дыхательных путей.

H-критерий Крускала–Уоллиса является непараметрическим аналогом однофакторного дисперсионного анализа для сравнения трех и более независимых групп. Данный критерий рассчитывается с использованием не фактических значений данных, а их рангов. *H*-критерий используется, если распределение в группах не является нормальным.

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 Крускала-Уоллиса асимптотически приближается к распределению «χ^2».

Схема применения H-критерия Крускала – Уоллиса:

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

2) p=0,05 - уровень значимости.

3) $H_{расч} = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$, где $n = \sum_{i=1}^{k} n_i$ - общее число наблюдений по всем группам, R_i - сумма рангов i-ой выборки.

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

4) В случае, когда число выборок $k>5$ $H_{табл} = \chi^2_{табл}(p; f)$, 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.

Пример 6.2. Для оценки дозовой нагрузки химическими веществами, загрязняющими питьевую воду, изучалось количество потребляемой для питья водопроводной воды среди разных возрастных групп населения. В результате получены следующие данные:

Дети (n_1=8)	Подростки (n_2=7)	Взрослые (n_3=9)
Вода, л/день	Вода, л/день	Вода, л/день
1,22	1,47	1,56
1,24	1,52	1,58
1,31	1,55	1,81
1,31	1,70	1,89
1,45	1,93	2,00
1,52	2,00	2,00
1,84	3,00	2,55
2,52		2,58
		4,00
$\bar{x}_1 = 1{,}55$	$\bar{x}_1 = 1{,}88$	$\bar{x}_1 = 2{,}22$

Проверить гипотезу о равенстве средних значений количества потребляемой для питья воды в популяциях детей, подростков и взрослого населения.

Решение:

Исследуемые группы являются независимыми, а данные имеют ненормальное распределение, то для проверки нулевой гипотезы используем критерий Крускала-Уоллиса.

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

1) H_1: $\bar{x}_1 \neq \bar{x}_2 \neq ... \neq \bar{x}_k$.
2) p=0,05 - уровень значимости.
3)

Дети (n_1=8)		Подростки (n_2=7)		Взрослые (n_3=9)	
Вода, л/день	Ранг	Вода, л/день	Ранг	Вода, л/день	Ранг
1,22	1	1,47	6	1,56	10
1,24	2	1,52	7,5	1,58	11
1,31	3,5	1,55	9	1,81	13
1,31	3,5	1,70	12	1,89	15
1,45	5	1,93	16	2,00	18
1,52	7,5	2,00	18	2,00	18
1,84	14	3,00	23	2,55	21
2,52	20			2,58	22
				4,00	24
Сумма рангов	56,5		91,5		152

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

$$= \frac{12}{24\cdot(24+1)}\cdot\left(\frac{56{,}5^2}{8} + \frac{91{,}5^2}{7} + \frac{152^2}{9}\right) - 3\cdot(24+1) = 0{,}02\cdot(399+1196+2567{,}1) - 75 = 8{,}24.$$

4) $H_{табл} = \chi^2_{табл}(0{,}05;\ 2) = 5{,}99$.
5) $H_{расч} > H_{табл}$, то H_0 отвергается, т.е. в разных возрастных группах ежедневно потребляется разное количество питьевой воды.

Двухфакторный дисперсионный анализ – система статистических методов

исследования действия на признак двух организованных факторов.

Двухфакторный дисперсионный анализ позволяет оценить не только влияние каждого из факторов в отдельности, но и их взаимодействие.

Т.к. вычисления при проведении двухфакторного дисперсионного анализа достаточно громоздки, рекомендуется пользоваться специальным программным обеспечением (Statistica 10, SPSS и др.).

Пример 6.3. В химической лаборатории проверяется влияние температуры (фактор «*A*») и катализатора (фактор «*B*») на выход продукта химического синтеза «*Y*». Полученные результаты приведены в таблице. Требуется проверить гипотезу о влиянии факторов «*A*» и «*B*» и их комбинации на указанный признак.

	B_1	B_2	B_3
A_1	16; 19; 17; 16	18; 16; 17; 14	16; 16; 18; 13
A_2	22; 22; 19; 23	18; 19; 23; 24	18; 16; 19; 20
A_3	20; 16; 18; 19	18; 17; 19; 19	20; 20; 16; 16
A_4	23; 20; 22; 23	19; 18; 19; 22	20; 19; 20; 22

1) Формулировка гипотез.

Гипотезы для фактора «*A*»:

H_0: Для всех режимов температуры «A_i» нет разницы между средним результатом продукта химического синтеза «*Y*».

H_1: Для всех типов «A_i» существует разница между средним результатом «*Y*».

Гипотезы для фактора «*B*»:

H_0: Для всех типов катализатора «B_j» нет разницы между средним результатом «*Y*».

H_1: Для всех типов «B_j» существует разница между средним результатом «*Y*».

Гипотезы для взаимодействия факторов «*A*» и «*B*»:

H_0: Фактор «*A*» (температура) и фактор «*B*» (катализатор) не оказывают эффекта взаимодействия на результат «*Y*» (продукт химического синтеза).

H_1: Фактор «*A*» и фактор «*B*» оказывают эффект взаимодействия на результат «*Y*».

2) Реализация двухфакторного дисперсионного анализа с помощью программы «Statistica 10» (рисунок 6.1).

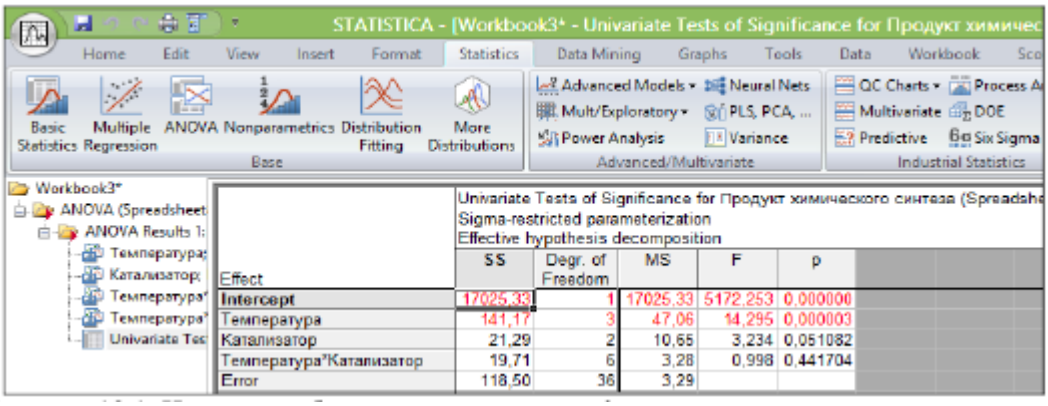


Рисунок 6.1. Итоговая таблица результатов двухфакторного дисперсионного анализа, выполненного в программе «Statistica 10»

3) Выводы:

Режим температуры оказывает влияние на средний результат химического синтеза (выделено красным цветом).

Тип катализатора не оказывает влияния на средний результат химического синтеза.

Сочетание температуры и катализатора не оказывает существенного влияния на продукт химического синтеза.

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: 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://www.statsoft.ru/

6. Контрольные вопросы:

1. Какая нулевая гипотеза проверяется с помощью дисперсионного анализа?
2. Какие условия должны выполняться при использовании дисперсионного анализа?
3. Какова основная идея дисперсионного анализа?
4. Когда используется *H*-критерий Крускала–Уоллиса?

ЛЕКЦИЯ № 7

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.

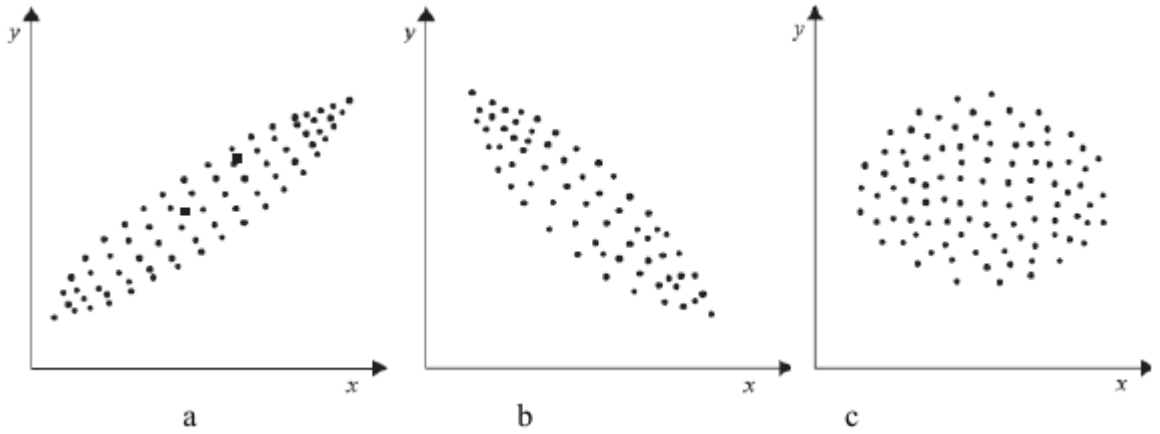


Рис 7.1. Диаграммы рассеяния: a - прямая связь; b - обратная связь; c - связь отсутствует

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?

ЛЕКЦИЯ №8

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

Впервые термин «регрессия» был введен основателем биометрии Ф. Гальтоном (XIX в.), идеи которого были развиты его последователем К. Пирсоном.

Регрессионный анализ - метод статистической обработки данных, позволяющий измерить связь между одной или несколькими причинами (факторными признаками) и следствием (результативным признаком).

Признак - это основная отличительная черта, особенность изучаемого явления или процесса.

Результативный признак - исследуемый показатель.

Факторный признак - показатель, влияющий на значение результативного признака.

Целью регрессионного анализа является оценка функциональной зависимости среднего значения результативного признака (y) от факторных ($x_1, x_2, \ldots, x_n$), выражаемой в виде *уравнения регрессии*: $y = f(x_1, x_2, \ldots, x_n)$.

Различают два вида регрессии: парную и множественную.

Парная (простая) регрессия - уравнение вида: $y = f(x)$.

Результативный признак при парной регрессии рассматривается как функция от одного аргумента, т.е. одного факторного признака.

Регрессионный анализ включает в себя следующие этапы:

1. определение типа функции;
2. определение коэффициентов регрессии;
3. расчет теоретических значений результативного признака;
4. проверку статистической значимости коэффициентов регрессии;
5. проверку статистической значимости уравнения регрессии.

Множественная регрессия - уравнение вида: $y = f(x_1, x_2, \ldots, x_n)$.

Результативный признак рассматривается как функция от нескольких аргументов, т.е. много факторных признаков.

Для того чтобы правильно определить тип функции нужно на основании теоретических данных найти направление связи.

По направлению связи регрессия делится на:

- *прямую регрессию,* возникающую при условии, что с увеличением или уменьшением независимой величины «*x*» значения зависимой величины «*y*» также соответственно увеличиваются или уменьшаются;
- *обратную регрессию,* возникающую при условии, что с увеличением или уменьшением независимой величины «*x*» зависимая величина «*y*» соответственно уменьшается или увеличивается.

Для характеристики связей используют следующие виды уравнений парной регрессии:

1. $y=a+bx$ – *линейное;*
2. $y=e^{ax+b}$ – *экспоненциальное;*
3. $y=a+b/x$ – *гиперболическое;*

4. $y=a+b_1x+b_2x^2$ – *параболическое;*
5. $y=ab^x$ – *показательное* и др.

где a, b_1, b_2 - коэффициенты (параметры) уравнения; y - результативный признак; x - факторный признак.

Построение уравнения регрессии сводится к оценке его коэффициентов (параметров), для этого используют *метод наименьших квадратов* (МНК).

Метод наименьших квадратов позволяет получить такие оценки параметров, при которых сумма квадратов отклонений фактических значений результативного признака «y» от теоретических «y_x» минимальна, то есть $\sum(y-y_x)^2 \to \min$.

Параметры уравнения регрессии $y=a+bx$ по методу наименьших квадратов оцениваются с помощью формул:

$$a=\bar{y}-b\bar{x}, \qquad b=\frac{\overline{yx}-\bar{y}\cdot\bar{x}}{\overline{x^2}-\bar{x}^2},$$

где a – свободный коэффициент, b - коэффициент регрессии, показывает на сколько изменится результативный признак «y» при изменении факторного признака «x» на единицу измерения.

Для оценки статистической значимости коэффициентов регрессии используется t-критерий Стьюдента.

Схема проверки значимости коэффициентов регрессии:

1. H_0: $a=0$, $b=0$ - коэффициенты регрессии незначимо отличаются от нуля.
   H_1: $a\neq0$, $b\neq0$ - коэффициенты регрессии значимо отличаются от нуля.
2. $p=0{,}05$ - уровень значимости.

$$t_{b\ расч}=\frac{b}{m_b}, \qquad t_{a\ расч}=\frac{a}{m_a},$$

3. где m_b, m_a - случайные ошибки:

$$m_b=\sqrt{\frac{\sum(y-y_x)^2}{n-2}}:\sum(x-\bar{x})^2; \qquad m_a=\sqrt{\frac{\sum(y-y_x)^2}{n-2}\cdot\frac{\sum x^2}{n\sum(x-\bar{x})^2}}.$$

4. $t_{табл}(p;f)$, где $f=n-k-1$-число степеней свободы (табличное значение), n – число наблюдений, k - число параметров в уравнении при переменных «x».
5. Если $t_{расч} > t_{табл}$, то H_0 отклоняется, т.е. коэффициент значимый.
   Если $t_{расч} < t_{табл}$, то H_0 принимается, т.е. коэффициент незначимый.

Для проверки правильности построенного уравнения регрессии применяется критерий Фишера.

Схема проверки значимости уравнения регрессии:

1) H_0: уравнение регрессии незначимо.
   H_1: уравнение регрессии значимо.
2) $p=0{,}05$ - уровень значимости.

$$F_{расч}=\frac{\frac{\sum(y_x-\bar{y})^2}{k}}{\frac{\sum(y-y_x)^2}{n-k-1}}=(n-2)\frac{r_{xy}^2}{1-r_{xy}^2},$$

3) где n - число наблюдений; k - число параметров в уравнении при переменных «x»; y - фактическое значение результативного признака; y_x -

теоретическое значение результативного признака; r_{xy} - коэффициент парной кореляции.

4) $F_{табл}(p; f_1; f_2)$, где $f_1=k$, $f_2=n-k-1$- число степеней свободы (табличные значения).

5) Если $F_{расч}>F_{табл}$, то уравнение регрессии подобрано верно и может применяться на практике.

Если $F_{расч}<F_{табл}$, то уравнение регрессии подобрано неверно.

Основным показателем, отражающим меру качества регрессионного анализа, является коэффициент детерминации (R^2).

Коэффициент детерминации показывает, какая доля зависимой переменной «у» учтена в анализе и вызвана влиянием на нее факторов, включенных в анализ.

Коэффициент детерминации (R^2) принимает значения в промежутке [0, 1].

Уравнение регрессии является качественным, если $R^2 \geq 0,8$.

Коэффициент детерминации равен квадрату коэффициента корреляции $R^2 = r_{xy}^2$.

Пример 8.1. По следующим данным построить и проанализировать уравнение регрессии:

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

Решение.

1) Вычислить коэффициент корреляции: $r_{xy}=0,47$. Связь между признаками прямая и умеренная.

2) Построить уравнение парной линейной регрессии.

2.1) Составить расчетную таблицу.

№	*X*	*y*	*xy*	x^2	y_x	$(y_x-\bar{y})^2$	$(y-y_x)^2$	$(x-\bar{x})^2$
1	352	64	22528	123904	55,89	47,54	65,70	6241
2	228	60	13680	51984	45,07	15,42	222,83	2025
3	340	52	17680	115600	54,85	34,19	8,11	4489
4	300	48	14400	90000	51,36	5,55	11,27	729
5	196	46	9016	38416	42,28	45,16	13,84	5929
6	258	41	10578	66564	47,69	1,71	44,77	225
7	237	32	7584	56169	45,86	9,87	192,05	1296
Сумма	1911	343	95466	542637	343	159,45	558,55	20934
Среднее	273	49	13638	77519,6	49	22,78	79,79	2990,6

2.2) Рассчитать коэффициенты регрессии:

$$b = \frac{\overline{yx} - \bar{y}\cdot\bar{x}}{\overline{x^2} - \bar{x}^2} = \frac{13638 - 49\cdot 273}{77519,6 - 273^2} = 0,087,$$

$$a = \bar{y} - b\bar{x} = 49 - 0,087\cdot 273 = 25,17.$$

Уравнение парной линейной регрессии: $y_x=25,17+0,087x$.

3) Найти теоретические значения «y_x» путем подстановки в уравнение регрессии фактических значений «*x*».

4) Построить графики фактических «у» и теоретических значений «y_x» результативного признака (рисунок 8.1):

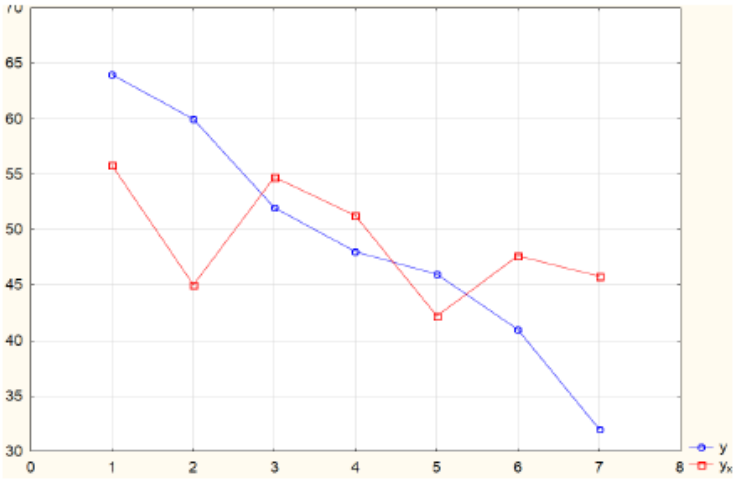


Рис. 8.1. Графики фактических «*y*» и теоретических значений «y_x» результативного признака

5) Проверить статистическую значимость коэффициентов регрессии:

5.1) Вычислить случайные ошибки:

$$m_b = \sqrt{\frac{\sum (y - y_x)^2}{n-2}} \Big/ \sum (x - \bar{x})^2 = \sqrt{\frac{558{,}55}{7-2}} \Big/ 20934 \approx 0{,}073;\quad m_a = \sqrt{\frac{\sum (y - y_x)^2}{n-2} \cdot \frac{\sum x^2}{n \sum (x - \bar{x})^2}} = \sqrt{\frac{79{,}8}{7-2} \cdot \frac{542637}{7 \cdot 2990{,}6}} \approx 20{,}34.$$

5.2)

$$t_{b\ расч} = \frac{b}{m_b} = \frac{0{,}087}{0{,}073} \approx 1{,}19, \qquad t_{a\ расч} = \frac{a}{m_a} = \frac{25{,}17}{20{,}34} \approx 1{,}24.$$

5.3) $t_{табл}(0{,}05;\ 5)=2{,}57$

5.4) $t_{b\ расч} < t_{табл}$, значит коэффициент *b* - незначим,

$t_{a\ расч} < t_{табл}$, значит коэффициент *a* - незначим.

6) Проверить статистическую значимость уравнения регрессии:

6.1)
$$F_{расч} = \frac{\sum (y_x - \bar{y})^2}{k} \Big/ \frac{\sum (y - y_x)^2}{n-k-1} = \frac{159{,}45}{1} \Big/ \frac{558{,}55}{5} = 1{,}43.$$

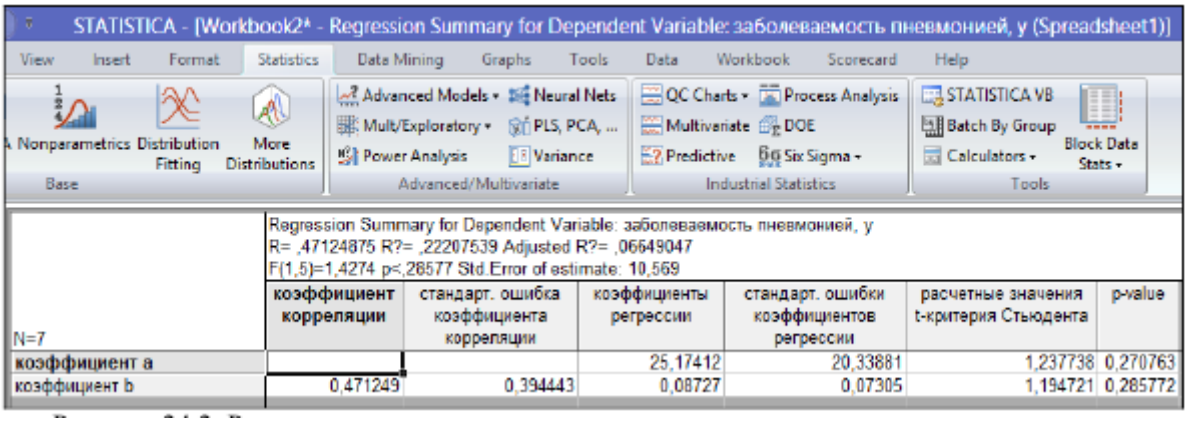
6.2) $F_{табл}(p;\ k;\ n\text{-}k\text{-}1)=(0{,}05;\ 1;\ 5)=6{,}61.$

6.3) $F_{расч}<F_{табл}$, значит уравнение регрессии подобрано неверно. Этот результат можно объяснить невысокой теснотой зависимости (r_{xy}=0,47) и небольшим числом наблюдений.

7) Вычислить коэффициент детерминации: $R^2=(0{,}47)^2=0{,}22$. Построенное уравнение некачественное.

Т.к. вычисления при проведении регрессионного анализа достаточно объемные, рекомендуется пользоваться специальными программами («Statistica 10», SPSS и др.).

На рисунке 8.2 приведена таблица с результатами регрессионного анализа, проведенного с помощью программы «Statistica 10».



Regression Summary for Dependent Variable: заболеваемость пневмонией, y
R= ,47124875 R?= ,22207539 Adjusted R?= ,06649047
F(1,5)=1,4274 p<,28577 Std.Error of estimate: 10,569

N=7	коэффициент корреляции	стандарт. ошибка коэффициента корреляции	коэффициенты регрессии	стандарт. ошибки коэффициентов регрессии	расчетные значения t-критерия Стьюдента	p-value
коэффициент a			25,17412	20,33881	1,237738	0,270763
коэффициент b	0,471249	0,394443	0,08727	0,07305	1,194721	0,285772

Рис.8.2. Результаты регрессионного анализа,
проведенного с помощью программы «Statistica 10».

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: 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)

6. Контрольные вопросы

1. Из каких этапов состоит регрессионный анализ?
2. Какие виды регрессии Вы знаете?
3. Как определяются коэффициенты линейного парного уравнения регрессии?
4. Как проверяется значимость коэффициентов регрессии?
5. Как проверяется значимость уравнения регрессии?

ЛЕКЦИЯ №9

1. Тема: Анализ качественных признаков.

2. Цель: Ознакомить студентов с методами анализа качественных признаков.

3. Тезисы лекции:

Признак – измеренное свойство объекта наблюдения. Различают количественные (рост, вес, САД, возраст и др.) и качественные (пол, национальность, диагноз, семейное положение и др.) признаки *сопряженности.*

Предположим, что имеется есть два качественных признака A и B, признак А имеет r градаций, а признак B - s градаций.

Пусть имеется выборка объема n.

Каждый объект выборки может обладать одним из уровней признака *A* и одновременно каким-либо уровнем признака *B*. По этой выборке можно определить частоты событий A_i и B_j по отдельности и в любых комбинациях.

Обозначим через ν_{ij} частоту события A_iB_j. Число появлений признака A_i (частота признака A_i) равно: $\nu_{i.} = \sum_{j=1}^{s} \nu_{ij} = \nu_{i1} + \nu_{i2} + ... + \nu_{is}$

Аналогично, частота появления события B_j равна: $\nu_{.j} = \sum_{i=1}^{r} \nu_{ij} = \nu_{1j} + \nu_{2j} + ... + \nu_{rj}$

Общее число наблюдений, т.е. объем выборки: $\nu_{..} = \sum_{i=1}^{r} \nu_{i.} = \sum_{j=1}^{s} \nu_{.j} = \sum_{i=1}^{r} \sum_{j=1}^{s} \nu_{ij}$

Замена индекса точкой (ν.i , ν.j , ν..) означает результат суммирования по этому индексу. Полученные частоты представляют в виде таблицы сопряженности (табл. 9.1).

Таблица 9.1.

	B_1	B_2	…	B_j	…	B_s	
A_1	ν_{11}	ν_{12}	…	ν_{1j}	…	ν_{1s}	$\nu_{1.}$
A_2	ν_{21}	ν_{22}	…	ν_{2j}	…	ν_{2s}	$\nu_{2.}$
…	…	…	…	…	…	…	…
A_i	ν_{i1}	ν_{i2}	…	ν_{ij}	…	ν_{is}	$\nu_{i.}$
…	…	…	…	…	…	…	…
A_r	ν_{r1}	ν_{r2}	…	ν_{rj}	…	ν_{rs}	$\nu_{r.}$
	$\nu_{.1}$	$\nu_{.2}$	…	$\nu_{.j}$	…	$\nu_{.s}$	$\nu_{..}=n$

При анализе таблиц сопряженности нулевая гипотеза формулируется следующим образом: связи между признаками *A* и *B* нет.

Для каждой клетки таблицы сопряженности (т.е. для каждой комбинации A_iB_j) рассчитываются теоретические частоты (таблица 9.2) по формуле:

$$\nu^*_{ij} = \nu_{.i} \cdot \frac{\nu_{.j}}{\nu_{..}}$$

	B1	B2	…	Bj	…	Bs	
A1	ν^*_{11}	ν^*_{12}	…	ν_{1j}	…	ν^*_{1s}	$\nu^*_{1.} = \nu 1.$
A2	ν^*_{21}	ν^*_{22}	…	ν_{2j}	…	ν^*_{2s}	$\nu^*_{2.} = \nu 2.$
…	…	…	…	…	…	…	…
Ai	ν^*_{i1}	ν^*_{i2}	…	ν^*_{ij}	…	ν^*_{is}	$\nu^*_{i.} = \nu i.$
…	…	…	…	…	…	…	…
Ar	ν^*_{r1}	ν^*_{r2}	…	ν^*_{rj}	…	ν^*_{rs}	$\nu^*_{r.} = \nu r.$
	$\nu^*_{.1} = \nu.1$	$\nu^*_{.2} = \nu.2$	…	$\nu^*_{.j} = \nu.j$	…	$\nu^*_{.s} = \nu.s$	ν..

При выполнении гипотезы H_0 наблюдаемые частоты ν_{ij} не должны сильно отличаться

от теоретических частот ν^*_{ij}.

Чтобы сопоставить теоретические и наблюдаемые частоты применяют χ^2-критерий Пирсона.

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

Если число наблюдений невелико и в клетках таблицы встречается частота меньше 5, критерий χ^2 *неприменим,* вместо него используется точный критерий Фишера.

Схема применения χ^2-критерия Пирсона:

1) H_0: связи между признаками нет.
H_1: связь между признаками есть.
2) p=0,05 - уровень значимости.
3) $$\chi^2_{расч} = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(\nu_{ij}-\nu^*_{ij})^2}{\nu^*_{ij}}$$ где ν_{ij} – наблюдаемые частоты, ν^*_{ij} – теоретические частоты.
4) $\chi^2_{табл}(p, f)$, 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.

Пример 9.1. Имеются данные о количестве наблюдений и случаев летальности для четырех форм острых гнойных деструкций легких. С помощью χ^2-критерия Пирсона требуется оценить значимость различия между группами по числу случаев летальных исходов.

Номер группы	Форма заболевания	Число случаев		Число больных
		летальных исходов	выздоровления	
1	Гнойный абсцесс	5	136	141
2	Гангренозный абсцесс	11	37	48
3	Гангрена доли	7	8	15
4	Тотальная гангрена	6	5	11

Решение.

1) H_0: связи между признаками нет.
H_1: связь между признаками есть.
2) p=0,05 – уровень значимости.
3)

3.1)Рассчитаем теоретические частоты, используя формулу $\nu^*_{ij} = \nu_{.i}\cdot\frac{\nu_{.j}}{\nu_{..}}$

	B_1	B_2	Всего
A_1	5	136	141
A_2	11	37	48
A_3	7	8	15
A_4	6	5	11
Всего	29	186	215

	B_1	B_2	Всего
A_1	29*141/215=19	186*141/215=122	141
A_2	29*48/215=6,5	186*48/215=41,5	48
A_3	29*15/215=2	186*15/215=13	15
A_4	29*11/214=1,5	186*11/214=9,5	11
Всего	29	186	215

3.2) Вычислим величины $\frac{(\nu_{ij}-\nu^*_{ij})^2}{\nu^*_{ij}}$

	B_1	B_2
A_1	$(5-19)^2/19=10,3$	$(136-122)^2/122=1,6$
A_2	$(11-6,5)^2/6,5=3,1$	$(37-41,5)^2/41,5=0,5$
A_3	$(7-2)^2/2=12,5$	$(8-13)^2/13=1,9$
A_4	$(6-1,5)^2/1,5=13,5$	$(5-9,5)^2/9,5=2,1$

3.3)Вычислим расчетное значение критерия:

$$\chi^2_{расч}=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(\nu_{ij}-\nu^*_{ij}\right)^2}{\nu^*_{ij}}=10,3+3,1+12,5+13,5+1,6+0,5+2+2,1=45,5\ .$$

4) $\chi^2_{табл}(p, f)$, где $f=(r-1)(s-1)=(4-1)(2-1)=3$ - число степеней свободы, $\chi^2_{табл}(0,05;3)=7,8$

5) Если $\chi^2_{расч} > \chi^2_{табл}$, то «$H_0$» отвергается, значит различия между группами по числу случаев летальных исходов статистически значимые.

Т.к. процесс вычисления критерия довольно трудоемкий, то целесообразно производить анализ с помощью специальных статистических программ.

В медицине очень часто используются таблицы сопряженности размера 2x2.

Предположим, что имеется два качественных признака (каждый из которых имеет две градации), характеризующие обследованных лиц (таблица 9.3).

Таблица 9.3.

	Первый признак (первая градация)	Первый признак (вторая градация)	Всего
Второй признак (первая градация)	Частота встречаемости a	Частота встречаемости b	$a+b$
Второй признак (вторая градация)	Частота встречаемости c	Частота встречаемости d	$c+d$
Всего	$n_1=a+c$	$n_2=b+d$	$n=a+b+c+d$

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

1) H_0: связи между признаками нет.
   H_1: связь между признаками есть.
2) p=0,05 - уровень значимости.
3) $\chi^2_{расч}=\frac{(ad-bc)^2 n}{(a+b)(c+d)(a+c)(b+d)}$, где a, b, c, d - частоты из таблицы сопряженности.
4) $\chi^2_{табл}(p, f)$, где $f=(r-1)(s-1)$ - число степеней свободы
5) Если $\chi^2_{расч} \le \chi_{табл}$, то «$H_0$» принимается.
   Если $\chi^2_{расч} > \chi_{табл}$, то «$H_0$» отвергается.

Пример 9.2. Исследуется взаимосвязь между приемом контрацептивных таблеток матерями, и желтухой у детей, получающих грудное вскармливание. Данные для исследования представлены в таблице.

Прием матерью таблеток	Есть желтуха	Нет желтухи	Всего
Принимала таблетки	33	24	57
Не принимала таблетки	14	45	59
Всего	47	69	116

Решение.

1) H_0: связи между признаками нет.
H_1: связь между признаками есть.
2) p=0,05 – уровень значимости.
3) $\chi^2_{расч} = \frac{(ad-bc)^2 n}{(a+b)(c+d)(a+c)(b+d)} = \frac{(33\cdot 45 - 24\cdot 14)^2 \cdot 116}{57\cdot 59\cdot 47\cdot 69} = 14{,}04$
4) $\chi^2_{расч}(0{,}05;1) = 3{,}8$
5) Т.к. $\chi^2_{расч} > \chi^2_{табл}$ то гипотеза о независимости между желтухой и приемом контрацептивных таблеток отвергается, т.е. зависимость существует.

Поправка Йетса

Формула $\chi^2_{расч} = \frac{(ad-bc)^2 n}{(a+b)(c+d)(a+c)(b+d)}$ для χ^2 в случае таблицы 2x2 дает завышенные значения. На практике это приводит к тому, что нулевая гипотеза будет отвергаться слишком часто. Чтобы компенсировать этот эффект, в формулу вводят поправку Йетса:

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

Для рассмотренного выше примера 9.2 расчетное значение критерия с поправкой Йетса: $\chi^2_{расч} = \frac{n\left(ad-bc-\frac{n}{2}\right)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{116\left(33\cdot 45-24\cdot 14-\frac{116}{2}\right)^2}{57\cdot 59\cdot 47\cdot 69} = 12{,}66$.

Критерий χ2 Пирсона применяется для независимых выборок. Если выборки зависимые, то применяется критерий χ2 Макнемара.

Выборки называются зависимыми, если над одними и теми же объектами проводятся два наблюдения: «до» и «после».

Критерий χ2 Макнемара применяется только для таблиц сопряженности размера 2x2.

Признак «до»	Признак «после»	
	Вторая градация «после» (-)	Первая градация«после» (+)
Первая градация «до» (+)	Число изменений от (+) к (-) a	*b* Число сохранивших (+)
Вторая градация «до» (-)	Число сохранивших (-) *c*	*d* Число изменений от (-) к (+)

Схема применения χ^2-критерия Макнемара (2x2)

1) H_0: частота встречаемости градаций признака после воздействия фактора не изменилась.
H_1: частота встречаемости градаций признака после воздействия фактора изменилась.
2) p=0,05 – уровень значимости.
3) $\chi^2_{расч} = \frac{(|a-d|-1)^2}{(a+d)}$
4) $\chi^2_{табл}(p, f)$, где f – число степеней свободы. Для таблицы спряженности 2x2 f=1.
5) Если $\chi^2_{расч} > \chi^2_{табл}$, то «$H_0$» принимается.
Если $\chi^2_{расч} > \chi^2_{табл}$, то «$H_0$» отвергается.

Пример 9.3. Исследуется эффективность пробиотика метаболитного типа в комплексной терапии при осложненной смешанной респираторной вирусной инфекции и его влияние на микробиоценоз кишечника. В исследовании приняли участие 32 больных. Данные для исследования представлены в таблице.

До лечения пробиотиком	После лечения пробиотиком	
	Нет дисбактериоза	Есть дисбактериоз
Есть дисбактериоз	9	5
Нет дисбактериоза	18	0

Решение.

1) H_0: частота заболеваний дисбактериозом после применения пребиотика не изменилась.

H_1: частота заболеваний дисбактериозом после применения пребиотика изменилась.

2) p=0,05 – уровень значимости

3) $\chi^2_{расч} = \frac{(|a-d|-1)^2}{(a+d)} = \frac{(|9-0|-1)^2}{(9+0)} = 7{,}11$

4) $\chi^2_{табл}(0{,}05;1) = 3{,}8$

5) Т.к. $\chi^2{}_{расч} > \chi^2{}_{табл}$, то гипотеза о том, что число пациентов с дисбактериозом после применения пребиотика не изменилась 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: 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)

6. Контрольные вопросы:

1. В чем особенность анализа качественных признаков?
2. Что из себя представляет таблица сопряженности 2х2?
3. Какие условия должны выполняться при применении критерия χ^2 Пирсона?
4. Для чего используется поправка Йетса?
5. В каких случаях применяется критерий χ^2 Макнемара?

ЛЕКЦИЯ №10

1. Тема: Анализ динамических рядов.

2. Цель: Ознакомить студентов с понятием динамические ряды, научить выравнивать

динамический ряд с помощью уравнения тренда, строить краткосрочные прогнозы и вычислять показатели динамического ряда.

3. Тезисы лекции:

Динамический (временной) ряд — совокупность значений какого-либо показателя за несколько последовательных моментов или периодов.

Каждый временной ряд состоит из двух элементов:

1) моменты или периоды времени, к которым относятся приводимые статистические данные;

2) статистические показатели, которые характеризуют изучаемый объект на определенный момент или за указанный период времени.

Статистические показатели, характеризующие изучаемый объект, называют *уровнями ряда.*

Динамические ряды классифицируются по времени, по полноте обхвата во времени, по форме представления уровней ряда (рис. 10.1).

Рисунок 10.1. Виды динамических рядов

1) Виды динамических рядов (по времени):

- *моментные ряды динамики* отображают состояние изучаемых явлений на определенные даты (моменты) времени;

Дата	1.01.2019	1.04.2019	1.07.2019	1.10.2019	1.01.2020
Количество сотрудников поликлиники (человек).	192	190	195	198	200

- *интервальные ряды динамики* отражают итоги развития изучаемых явлений за отдельные периоды (интервалы) времени.

Год	2016	2017	2018	2019	2020
Количество детей, вакцинированных против кори (тыс. человек)	88,5	93,2	98,0	102,8	108,8

2) Виды динамических рядов (по полноте обхвата во времени):

- *полные ряды динамики* имеют равные интервалы;
- *неполные ряды динамики* не имеют равных интервалов.

3) Виды динамических рядов (по форме представления уровней ряда):

- *ряды абсолютных величин* – уровни ряда выражены в соответствующих единицах измерения (кг, л, км, ч, тг и др.);
- *ряды относительных величин* – уровни ряда выражены в процентах, долях, промилле и др.;
- *ряды средних величин* - уровни ряда выражены числами, которые являются

усредненным показателями.

Тренд - это функция от времени, определяющая основную тенденцию развития показателя во времени.

Для установления тренда динамический ряд выравнивают. Выравнивание осуществляется следующими способами: укрупнение периодов, расчет групповой средней, расчет скользящей средней, метод наименьших квадратов.

Метод наименьших квадратов применяется для более точной количественной оценки динамики изучаемого явления: $\sum(y-y_t)^2 \to \min$, где y - фактические (эмпирические) уровни ряда, y_t - теоретические значения уровней ряда, т.е. вычисленные по соответствующему аналитическому уравнению на момент времени «*t*».

Обычно строят *линейный тренд* – это уравнение прямой линии, выражающее тенденцию изменения временного ряда, которое имеет вид: $y_t = a + bt$, где a, b - коэффициенты, рассчитываемые по формулам: $a = \bar{y} - b\bar{t}$, $b = \dfrac{\overline{yt} - \bar{y}\bar{t}}{\overline{t^2} - \bar{t}^2}$.

Расчет параметров уравнения (a и b) можно упростить, если отсчет времени производить так, чтобы сумма показателей времени изучаемого ряда динамики была равна нулю: $\sum t = 0$. При этом используют следующие формулы:

- если ряд содержит нечетное число членов $t_{н/ч} = k - \dfrac{n+1}{2}$
- если ряд содержит четное число членов $t_{ч} = 2k - (n+1)$

где k – порядковый номер года, n – число лет в периоде.

Тогда, формулы для нахождения коэффициентов уравнения линейного тренда примут вид: $a = \dfrac{\sum y_i}{n}$, $b = \dfrac{\sum y_i \cdot t_i}{\sum t_i^2}$.

По рассчитанным параметрам записывают уравнение прямой линии для ряда динамики, представляющей собой *трендовую модель* искомой функции: $y_t^* = a + bt$. Подставляя последовательно в уравнение значения «*t*», находят *теоретические значения* уровней ряда.

Для определения прогнозных значений уровней ряда динамики на будущее используют метод экстраполяции.

Под экстраполяцией понимают нахождение уровней за пределами изучаемого ряда, т.е. продление в будущее тенденции, наблюдавшейся в прошлом.

На практике результат экстраполяции прогнозируемых явлений обычно получают в виде интервальных оценок – доверительных интервалов прогноза.

Для определения границ интервалов используют формулу: $y_t^* \pm t_p S$, где y_t^* - точечная оценка прогнозного значения уровня ряда в момент времени t, S – остаточное среднее квадратическое отклонение от тренда.

$$S = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - y_i^*\right)^2}{n-m}},$$

где n – число уровней ряда динамики, m –число параметров модели тренда (для линейного m=2), t_p - коэффициент доверия по распределению Стьюдента, при уровне значимости p=0,05 и числе степеней свободы f=n-m.

Пример 10.1. На основании данного динамического ряда:

1. построить уравнение линейного тренда, отражающего тенденцию заболеваемости;
2. построить на графике теоретическую кривую по выровненным уровням ряда динамики и сделать вывод о характере общей тенденции заболеваемости;
3. определить прогнозируемую заболеваемость ветряной оспой в 2018 году
4. доверительной вероятностью 95%.

Год	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Заболеваемость ветряной оспой на 10 000 тыс. населения	3,5	4,9	3,6	5,7	6,5	5,5	8,1	7,2	5,0	7,3

Решение.

Составим расчетную таблицу:

Год	**Период (*k*)**	**Показатель (y_i)**	**Период (t_i)**	$y_i t_i$	t^2
2008	1	3,5	-9	-31,5	81
2009	2	4,9	-7	-34,3	49
2010	3	3,6	-5	-18	25
2011	4	5,7	-3	-17,1	9
2012	5	6,5	-1	-6,5	1
2013	6	5,5	1	5,5	1
2014	7	8,1	3	24,3	9
2015	8	7,2	5	36	25
2016	9	5,0	7	35	49
2017	10	7,3	9	65,7	81
Сумма		**57,3**	**0**	**59,1**	**330**

1) Определим коэффициенты уравнения линейного тренда:

$$a = \frac{\sum y_i}{n} = \frac{57,3}{10} = 5,73 \quad b = \frac{\sum y_i \cdot t_i}{\sum t_i^2} = \frac{59,1}{330} = 0,18$$

Уравнение линейного тренда: $y_t^* = a + bt \Rightarrow \quad y_t^* = 5,73 + 0,18 \cdot t$

2) Подставляя в полученное уравнение значения *t*, найдем *выровненные уровни* (y_t^*).

Год	**Период (*k*)**	**Показатель (y_i)**	**Период (t_i)**	$y_i t_i$	t^2	y_t^*
2008	1	3,5	-9	-31,5	81	**4,11**
2009	2	4,9	-7	-34,3	49	**4,47**
2010	3	3,6	-5	-18	25	**4,83**
2011	4	5,7	-3	-17,1	9	**5,19**
2012	5	6,5	-1	-6,5	1	**5,55**
2013	6	5,5	1	5,5	1	**5,91**
2014	7	8,1	3	24,3	9	**6,27**
2015	8	7,2	5	36	25	**6,63**
2016	9	5,0	7	35	49	**6,99**
2017	10	7,3	9	65,7	81	**7,35**
Сумма		**57,3**	**0**	**59,1**	**330**	**57, 3**

Построим на графике теоретическую кривую по выровненным уровням ряда

динамики (рис. 10.2).

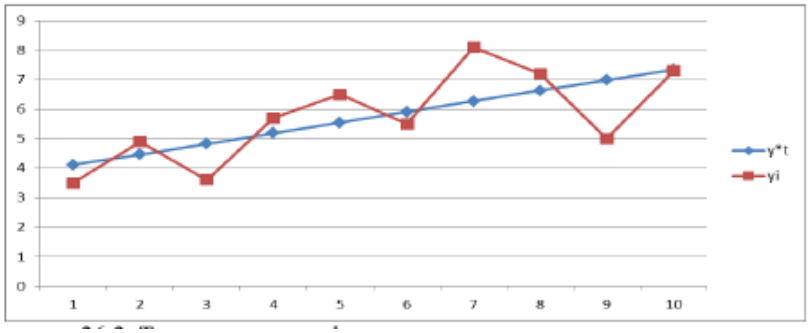


Рисунок 10.2. Теоретическая и фактическая кривые для динамического ряда

На основании полученных данных за период 10 лет можно сделать вывод о тенденции к росту заболеваемости ветряной оспой в данном регионе.

3) Построим прогноз на 2018 год с доверительной вероятностью 95%.

3.1) Определим величину: $S = \sqrt{\dfrac{\sum_{i=1}^{n}\left(y_i - y_t^*\right)^2}{n-m}}$

Год	Период (k)	Показатель (y_i)	Период (t_i)	$y_i t_i$	t^2	y_t^*	$y_i - y_t^*$	$\left(y_i - y_t^*\right)^2$
2008	1	3,5	-9	-31,5	81	4,11	-0,61	0,37
2009	2	4,9	-7	-34,3	49	4,47	0,43	0,18
2010	3	3,6	-5	-18	25	4,83	-1,23	1,51
2011	4	5,7	-3	-17,1	9	5,19	0,51	0,26
2012	5	6,5	-1	-6,5	1	5,55	0,95	0,90
2013	6	5,5	1	5,5	1	5,91	-0,41	0,17
2014	7	8,1	3	24,3	9	6,27	1,83	3,35
2015	8	7,2	5	36	25	6,63	0,57	0,32
2016	9	5,0	7	35	49	6,99	-1,99	3,96
2017	10	7,3	9	65,7	81	7,35	-0,05	0,00
Сумма		**57,3**	**0**	**59,1**	**330**	**57, 3**		**11,04**

$$S = \sqrt{\frac{11{,}04}{10-2}} = 1{,}17$$

Рассчитаем точечную оценку прогнозного значения уровня динамики в момент времени t=11: $y_{11}^* = 5{,}73 + 0{,}18 * 11 = 7{,}71$

По таблице найдем значение $t(0{,}05;\ 8)$= 2,31.

Определим границы прогнозного интервала:

$$y_t^* \pm t_p S \ \Rightarrow\ 7{,}71 \pm 2{,}31 \cdot 1{,}17 \ \Rightarrow\ 5{,}01 \le y_{прогноз} \le 10{,}41$$

Вывод: С вероятностью 95% можно сказать, что в 2018 году заболеваемость ветряной оспой в данном регионе будет не менее чем 5,01 и не более чем 10,41 чел. на 10 000 тыс. населения.

Показатели динамического ряда

Анализ скорости и интенсивности развития явлений во времени осуществляется с помощью статистических показателей, которые получаются в результате сравнения уровней между собой.

К таким показателям относятся: абсолютный прирост, темп прироста, абсолютное

значение одного процента прироста.

При этом принято сравниваемый уровень называть *отчетным*, а уровень, с которым производится сравнение - *базисным*.

Показатели динамики с *постоянной* базой (*базисные показатели*) характеризуют окончательный результат всех изменений в уровнях ряда от периода, к которому относится базовый уровень, до данного (*i*-го) периода.

Показатели динамики с *переменной* базой (*цепные показатели*) характеризуют интенсивность изменения уровня от периода к периоду в пределах изучаемого промежутка времени.

1. *Абсолютный прирост (Δi)* - показатель, определяемый как разность между двумя уровнями динамического ряда. Он показывает, на сколько данный уровень ряда превышает уровень, принятый за базу сравнения: $\Delta_i^б = y_i - y_0$, где $\Delta_i^б$ - абсолютный базисный прирост; y_i - уровень сравниваемого периода; y_0 - уровень базисного периода.

При сравнении с переменной базой абсолютный прирост будет равен: $\Delta_i^ц = y_i - y_{i-1}$, где y_{i-1} - уровень непосредственно предшествующего периода.

Год	**Показатель (y_i)**	$\Delta_i^б$ **(База 2008 года)**	Δ_i^T
2008	3,5	-	-
2009	4,9	4,9-3,5=1,4	4,9-3,5=1,4
2010	3,6	3,6-3,5=0,1	3,6-4,9=-1,3
2011	5,7	5,7-3,5=2,2	5,7-3,6=2,1
2012	6,5	6,5-3,5=3	6,5-5,7=0,8
2013	5,5	5,5-3,5=2	5,5-6,5=-1
2014	8,1	8,1-3,5=4,6	8,1-5,5=2,6
2015	7,2	7,2-3,5=3,7	7,2-8,1=-0,9
2016	5,0	5,0-3,5=1,5	5,0-7,2=-2,2
2017	7,3	7,3-3,5=3,8	7,3-5,0=2,3

Абсолютный прирост с переменной базой называют *скоростью роста*.

2. *Коэффициент роста* определяется как отношение двух сравниваемых уровней и показывает, во сколько раз данный уровень превышает уровень базисного периода:

базисный $k_i^б = \frac{y_i}{y_0}$, цепной $k_i^ц = \frac{y_i}{y_{i-1}}$.

Год	**Показатель (y_i)**	$k_i^б$ **(База 2008 года)**	k_i^T
2008	3,5	-	-
2009	4,9	4,9/3,5=1,40	4,9/3,5=1,40
2010	3,6	3,6/3,5=1,03	3,6/4,9=0,73
2011	5,7	5,7/3,5=1,62	5,7/3,6=1,58
2012	6,5	6,5/3,5=1,86	6,5/5,7=1,14
2013	5,5	5,5/3,5=1,57	5,5/6,5=0,85
2014	8,1	8,1/3,5=2,31	8,1/5,5=1,47
2015	7,2	7,2/3,5=2,06	7,2/8,1=0,89
2016	5,0	5/3,5=1,43	5/7,2=0,69
2017	7,3	7,3/3,5=2,09	7,3/5=1,46

3. Если коэффициенты роста выражают в процентах, то их называют *темпами роста*, т.е. они характеризуют скорость изменения показателя в единицу времени,

выраженную в процентах: $T_P = k \cdot 100\%$

Год	**Показ. (y_i)**	$k_i^б$	$T_p^б$ %	k_i^T	T_p^T %
2008	3,5	-	-	-	-
2009	4,9	4,9/3,5=1,40	140	4,9/3,5=1,40	140
2010	3,6	3,6/3,5=1,03	103	3,6/4,9=0,73	73
2011	5,7	5,7/3,5=1,62	162	5,7/3,6=1,58	158
2012	6,5	6,5/3,5=1,86	186	6,5/5,7=1,14	114
2013	5,5	5,5/3,5=1,57	157	5,5/6,5=0,85	85
2014	8,1	8,1/3,5=2,31	231	8,1/5,5=1,47	147
2015	7,2	7,2/3,5=2,06	206	7,2/8,1=0,89	89
2016	5,0	5/3,5=1,43	143	5/7,2=0,69	69
2017	7,3	7,3/3,5=2,09	209	7,3/5=1,46	146

4. *Темп прироста* показывает, на сколько процентов уровень данного периода больше (или меньше) базисного уровня. Этот показатель может быть рассчитан двояко:

- как отношение абсолютного прироста к базисному уровню:

$$\text{базисный } T_П^б = \frac{y_i - y_0}{y_0} \cdot 100\%; \quad \text{цепной } T_П^ц = \frac{y_i - y_{i-1}}{y_{i-1}} \cdot 100\%$$

- как разность между темпом роста (в %) и 100% : $T_П = T_P - 100\%$

5. Чтобы правильно оценить значение полученного темпа прироста, его рассматривают в сопоставлении с показателем абсолютного прироста.

Результат выражают показателем, который называют *абсолютным значением одного процента прироста (A_i)*: $A_i = \frac{y_i - y_{i-1}}{T_П} = \frac{\Delta_i^ц}{T_П}$.

6. При сопоставлении динамики развития двух явлений можно использовать показатели, представляющие собой отношения темпов прироста за одинаковые отрезки времени по двум динамическим рядам.

$$k_{оп} = \frac{T'_p}{T''_p} \quad \text{или} \quad k_{оп} = \frac{T'_П}{T''_П},$$

Эти показатели называют коэффициентами опережения:

где T', T''_p, $T'_П$, $T''_П$ - соответственно темпы роста и темпы прироста.сравниваемых динамических рядов.

С помощью этих коэффициентов могут сравниваться:

- ряды одинакового содержания но относящиеся к разным территориям (странам, регионам, районам и т.п.) или различным организациям;

- ряды разного содержания, характеризующие один и тот же объект.

Для обобщающей характеристики динамики исследуемого явления за ряд периодов определяют различного рода средние показатели.

1. *Средний абсолютный прирост* - средняя величина изменения показателя за интервал времени.

Рассчитывается как средняя арифметическая величина из показателя скорости роста за отдельные промежутки времени: $\overline{\Delta} = \frac{\sum_{i=1}^{n} \Delta_i^ц}{n-1} = \frac{y_n - y_1}{n-1}$.

где n - число уровней ряда; $\Delta^{ц}_{i}$ - абсолютные изменения по сравнению с предшествующим уровнем.

Средний абсолютный прирост дает возможность рассчитать, на сколько в среднем за единицу времени должен увеличиваться уровень ряда.

2. *Средний темп роста* – это характеристика интенсивности изменения уровней ряда динамики. Он показывает во сколько раз в среднем за единицу времени изменился уровень динамического ряда: $\overline{T} = \bar{k} \cdot 100\%$, где $\bar{k} = \sqrt[n-1]{\frac{y_n}{y_1}}$ - средний коэффициент роста; n-число уровней ряда.

3. *Средний темп прироста* вычисляется по следующей формуле: $\overline{T}_{Пр} = \overline{T}_{Р} - 100\%$.

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: 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)

6. Контрольные вопросы

1. Что такое динамический ряд? Из каких элементов он состоит?
2. Какие виды динамических рядов Вы знаете?
3. Что такое тренд?
4. Какими способами осуществляется выравнивание динамического ряда?
5. Как определяются коэффициенты линейного тренда?
6. В чем разница между базисными и цепными показателями?
7. Какие относительные, средние показатели динамики Вы знаете?