

**ФМ-Ф-14**

**Федеральное государственное бюджетное образовательное учреждение высшего образования «Северо-Осетинская государственная медицинская академия»  
Министерства здравоохранения Российской Федерации**

**МЕТОДИЧЕСКАЯ РАЗРАБОТКА  
ПО МЕТОДАМ СТАТИСТИЧЕСКИХ ИССЛЕДОВАНИЙ В  
МЕДИЦИНЕ И БИОЛОГИИ**

основной профессиональной образовательной программы высшего образования – программы подготовки научно-педагогических кадров в аспирантуре по направлению подготовки 30.06.01 Фундаментальная медицина по специальности 14.03.06 Фармакология, клиническая фармакология

Форма обучения: **заочная (заочная)**

Срок освоения: **3 года (4 года)**

Кафедра Химии и физики

Квалификация (степень) выпускника: исследователь

Преподаватель-исследователь

**Владикавказ, 2020 г**

## СОДЕРЖАНИЕ

1. **ТЕМА 1** Случайное событие. Вероятность случайного события. Классическое статистическое определение вероятности. Условия нормировки. Условная вероятность. Теорема умножения вероятностей. Случайные величины. Законы распределения случайных величин. Нормальный закон распределения
2. **ТЕМА 2.** Элементы организации медико-статистического исследования. Статистическая совокупность. Статистические величины. Вычисление статистических величин
3. **ТЕМА 3** Графический анализ данных. Изучение распределения случайных величин, подчиняющихся нормальному закону распределения Гаусса
4. **ТЕМА 4:** Статистические гипотезы. Компьютеры в медико-биологической статистике»
5. **ТЕМА 5:** Критерии Стьюдента для двух несвязанных выборок, F- критерии Фишера, U-критерии Манна-Уитни, критерии Краскела-Уолиллиса для выявления различий в уровне признака
6. **ТЕМА 6:** Параметрические коэффициенты корреляции. Применение критериев Стьюдента, Вилкоксона, Фридмана.
7. **ТЕМА 7:** Применение линейной корреляции. Пирсона. Коэффициент Спирмена
8. **ТЕМА 8:** Регрессионные коэффициенты t-критерию Стьюдента. Коэффициент множественной детерминации. Стандартный и пошаговый метод. Регрессионный анализ с помощью метода ввода.
9. **ТЕМА 9:** «Дисперсионный анализ. ОДА. ДДА.»
10. **ТЕМА 10** Многомерные статистические методы.

**ТЕМА 1: Случайное событие. Вероятность случайного события. Классическое статистическое определение вероятности. Условия нормировки. Условная вероятность. Теорема умножения вероятностей. Случайные величины. Законы распределения случайных величин. Нормальный закон распределения.**

**1. Научно-методическое обоснование темы:**

*Теория вероятностей* – математическая наука, изучающая закономерности в явлениях и опытах, результаты которых не могут быть заранее предсказаны.

Возникновение теории вероятностей как науки относят к средним векам. Первоначальным толчком к развитию теории вероятностей послужили задачи, относящиеся к азартным играм, таким, как орлянка, кости, карты, рулетка, когда в них начали применять количественные подсчеты и прогнозирование шансов на успех.

Другим толчком для развития теории вероятностей послужило страховое дело, а именно с конца XVII века на научной основе стало производиться страхование от несчастных случаев и стихийных бедствий. Стала зарождаться новая наука, вырисовываться ее специфика и методология: определения, теоремы, методы.

В ТВ вводятся специальные понятия и строятся специфические математические модели. Исходными понятиями в ТВ являются понятие случайного события и вероятности.

Слово «статистика» происходит от латинского слова «status» - состояние, положение. Впервые это слово в середине XVIII века применил немецкий ученый Ахенваль при описании состояния государства (нем. Statistik, от итал. stato - государство).

**Статистика:**

1) вид практической деятельности, направленной на сбор, обработку, анализ и публикацию статистической информации, характеризующей количественные закономерности жизни общества (экономики, культуры, политики и др.).

2) отрасль знаний (и соответствующие ей учебные дисциплины), в которой излагаются общие вопросы сбора, измерения и анализа массовых количественных данных.

*Статистика как наука включает разделы:* общая теория статистики, экономическая статистика, медицинская статистика, отраслевые статистики и др.

Общая теория статистики излагает общие принципы и методы статистической науки.

Как каждая наука, статистика имеет свой *предмет исследования* – массовые явления и процессы общественной жизни, свои *методы исследования* - статистические, математические, разрабатывает системы и подсистемы показателей, в которых отражаются размеры и качественные соотношения общественных явлений.

Статистика изучает количественные уровни и соотношения общественной жизни в неразрывной связи с их качественной стороной.

Статистика возникла на базе математики, и широко пользуется *математическими методами*. Это выборочный метод исследования, основанный на математической теории вероятности и законе больших чисел, различные методы обработки вариационных и динамических рядов, измерение корреляционных связей между явлениями и др.

Статистика разрабатывает и *специальную методологию исследования и обработки материалов*: массовые статистические наблюдения, метод группировок, средних величин, индексов, метод графических изображений.

Главная задача статистики, как и всякой другой науки, заключается в установлении закономерностей изучаемых явлений.

## 2. Краткая теория.

### Элементы теории вероятности. Основные понятия теории вероятности.

**Испытание** (опыт, эксперимент) – это выполнение определенного комплекса условий, в которых наблюдается то или иное явление, фиксируется тот или иной результат.

В теории вероятностей рассматриваются **испытания**, результаты которых нельзя предсказать заранее, а сами испытания можно повторять, хотя бы теоретически, произвольное число раз при неизменном комплексе условий. Испытаниями, например, являются: подбрасывание монеты, выстрел из винтовки, проведение денежно-вещевой лотереи.

Испытание, в котором событие  $A$  наступило, называется **успешным**, в противном случае – **неудачным**.

**Случайным событием** (возможным событием или просто *событием*) называется любой факт, который в результате испытания может произойти или не произойти.

Случайное событие – это не какое-нибудь происшествие, а лишь возможный **исход**, результат испытания (опыта, эксперимента). События обозначаются прописными (заглавными) буквами латинского алфавита:  $A, B, C$ .

1. Событие называется **достоверным**, если оно всегда происходит.
2. Событие, которое никогда не произойдет, называется **невозможным**.
3. **Суммой** двух событий называется такое событие, которое наступает тогда и только тогда, когда наступает хотя бы одно из слагаемых ( $A \cup B$ ).
4. **Произведением** двух событий называется такое событие, которое наступает тогда и только тогда, когда наступают оба события ( $A \cap B$ ).
5. **Разностью** двух событий называется событие, которое наступает тогда, когда наступает одно событие и не наступает другое ( $A \setminus B$ ).
6. Говорят, что событие  $A$  **влечет за собой** событие  $B$ , если из наступления  $A$  всегда следует наступление  $B$  ( $B \subset A$ ).
7. Если одновременно  $B \subset A$  и  $A \subset B$ , то в этом случае события  $A$  и  $B$  называются **равносильными**.
8. События  $A$  и  $B$  называются **несовместными**, если наступление одного из них исключает появление другого в одном и том же испытании.
9. События  $A$  и  $B$  называются **совместными** если они могут произойти вместе в одном и том же испытании.

**Пример 1.** Испытание состоит в однократном подбрасывании игральной кости с шестью гранями. Событие  $A$  – появление трех очков, событие  $B$  – появление четного числа очков,  $C$  – появление нечетного числа очков. События  $A$  и  $C$  совместны, поскольку число 3 – нечетное, а значит, если выпало 3 очка, то произошло и событие  $A$  и событие  $C$ . Кроме того, событие  $A$  влечет за собой событие  $C$ . События  $A$  и  $B$  несовместны, т.к. если произошло и событие  $A$ , то не произойдет событие  $B$ , а если произошло событие  $B$ , то не произойдет событие  $A$ . События  $B$  и  $C$  также являются несовместными.

10. События называются **парно несовместными** (или **взаимоисключающими**), если любые два из них несовместны.

**Пример 2.** Испытание – сдача студентом экзамена по определенной дисциплине. События – соответственно студент получит на экзамене один балл, два, три и т.д. Эти события являются парно несовместными.

События образуют **полную группу** для данного испытания, если они парно несовместны и в результате испытания обязательно появится одно из них.

В примере 2 события образуют полную группу, а события – нет.

11. События называются **равновозможными**, если нет оснований считать, что одно из них является более возможным, чем другое.

Примеры равновозможных событий: выпадение любого числа очков при броске игральной кости, появление любой карты при случайном извлечении из колоды, выпадение герба или цифры при броске монеты и т.п.

### **Классическое определение вероятности.**

Для практической деятельности важно уметь сравнивать события по степени возможности их наступления. Например, интуитивно ясно, что при последовательном извлечении из колоды пяти карт более возможна ситуация, когда появились карты разных мастей, чем появление пяти карт одной масти; при десяти бросках монеты более возможно чередование гербов и цифр, нежели выпадение подряд десяти гербов, и т.д. поэтому для сравнения событий нужна определенная мера.

*Численная мера степени объективной возможности наступления события называется **вероятностью события** и является, наряду с понятием случайного события, вторым основным понятием теории вероятности.*

Множество всех взаимоисключающих исходов эксперимента называется **пространством элементарных событий**. Пространство элементарных событий будем обозначать буквой  $\Omega$ , а его исходы – буквой  $\omega$ , т.е.  $\omega \in \Omega$ .

Пусть производится испытание с конечным числом равновозможных исходов  $\omega_1, \omega_2, \dots, \omega_n$ , образующих полную группу событий.

Пусть число возможных исходов равно  $n$  (общее число элементарных исходов), а при  $m$  из них происходит некоторое событие  $A$  (число благоприятных исходов), тогда при сделанных ранее предположениях на испытание, **вероятностью  $P(A)$  случайного события  $A$** , наступившего в данном испытании вычисляется по формуле:

$$P(A) = \frac{m}{n}$$

Это, так называемое, **классическое определение вероятности**.

**Пример 3.** из урны, содержащей 6 белых и 4 черных шара, наудачу вынут шар. Найти вероятность того, что он белый.

Решение. Будем считать элементарными событиями, или исходами опыта, извлечение из урны каждого из имеющихся в ней шаров. Число возможных исходов равно 10, а число благоприятных исходов (появлению белого шара) – 6 (количество белых шаров). Значит,

$$P(A) = \frac{m}{n} = \frac{6}{10} = 0.6$$

Данная формула справедлива только в случае всех равновозможных исходов. Она может применяться только для очень узкого класса задач. В случае, когда исходы не равновозможные, требуется определять вероятность события другим способом. Для этого введем вначале понятие **относительной частоты (частоты)  $W(A)$**  события  $A$  как отношение числа опытов, в которых наблюдалось событие  $A$ , к общему количеству проведенных испытаний

$$W(A) = \frac{M}{N}$$

где  $N$  – общее число опытов,  $M$  – число опытов, в которых появилось событие  $A$ . Т.е. **статистической вероятностью события** считается его относительная частота или число, близкое к ней.

### **Основные теоремы.**

#### **Теорема сложения вероятностей.**

**Теорема 1.** Вероятность  $P(A+B)$  суммы событий  $A$  и  $B$  равна

$$P(A+B) = P(A) + P(B) - P(AB).$$

$$P(A + B) = \frac{m_A + m_B + m_{AB}}{n} = \frac{m_A}{n} + \frac{m_B}{n} - \frac{m_{AB}}{n} = P(A) + P(B) - P(AB)$$

что и требовалось доказать.

**Следствие.** Сумма вероятностей противоположных событий равна 1:

$$P(A) + P(\bar{A}) = 1$$

### Теорема умножения вероятностей.

Остановимся более подробно на следующем примере иллюстративного характера. Допустим, что студент из 30 билетов успел выучить билеты с 1-го по 3-й и с 28-го по 30-й. На экзамен он пришел одиннадцатым, и оказалось, что к его приходу остались только билеты с 1-го по 20-й (событие  $A$ ). Вероятность события  $B = \{\text{студент получил выученный билет}\}$  без дополнительной информации о том, что событие  $A$  произошло, может быть вычислена по классическому определению с  $\Omega = \{1, 2, \dots, 30\}$ . Согласно формуле

$$P(B) = \frac{6}{30} = \frac{1}{5}$$

При дополнительной информации (событие  $A$  произошло) множество возможных исходов  $A$  состоит из 20 элементарных исходов, а событие  $B$  вместе с  $A$  наступает в 3 случаях. Следовательно, в рассматриваемом примере естественно определить *условную*

*вероятность*  $P(B|A) = P_A(B) = P(B \setminus A) = P_A(B) = \frac{P(AB)}{P(A)}$  *события  $B$  при условии, что*

*событие  $A$  произошло, как*  $P_A(B) = \frac{3}{20}$

**Теорема 2 (теорема умножения).** Вероятность произведения двух событий равна произведению вероятности одного из них на условную вероятность другого при условии, что первое событие произошло:

$$P(AB) = P(A) \cdot P_A(B)$$

**Доказательство.** Воспользуемся обозначениями теоремы 1. Тогда для вычисления  $P_A(B)$  множеством возможных исходов нужно считать  $m_A$  (так как  $A$  произошло), а множеством благоприятных исходов – те, при которых произошли и  $A$ , и  $B$  ( $m_{AB}$ ). Следовательно,

$$P_A(B) = \frac{m_{AB}}{m_A} = \frac{m_{AB}}{n} \cdot \frac{n}{m_A} = \frac{P(AB)}{P(A)}$$

откуда следует утверждение теоремы.

### Формула полной вероятности

**Теорема 3 (формула полной вероятности).** Если событие  $A$  может произойти только при условии появления одного из событий (гипотез)  $H_i = 1, \dots, n$  образующих полную группу, то вероятность события  $A$  равна

$$P(A) = \sum_{i=1}^n P(H_i) \cdot P_{H_i}(A)$$

Группа событий  $H_1, \dots, H_n$  называется *полной группой событий*, если выполняются следующие условия:

1.  $P(A_i) > 0, i = \overline{1, n}$
2.  $A_i \cdot A_j \neq \emptyset, i, j = \overline{1, n}, i \neq j$
3.  $\sum_{i=1}^n A_i = \Omega$

**Теорема 3 (формула Байеса).** Пусть  $H_1, \dots, H_n$  полная группа,  $(P(A) > 0)$ . Тогда имеем место формула

$$P_A(H_i) = \frac{P(H_i) \cdot P_{H_i}(A)}{P(A)} = \frac{P(H_i) \cdot P_{H_i}(A)}{\sum_{i=1}^n P(H_i) \cdot P_{H_i}(A)}$$

**Доказательство.** По определению условной вероятности  $P_A(B) = \frac{P(AB)}{P(A)}$ . В

числителе применим теорему умножения, а в знаменателе – формулу полной вероятности.

### Случайные величины.

Наряду с понятием случайного события в теории вероятности используется понятие случайной величины.

**Случайной величиной** называется переменная величина, которая в результате испытания в зависимости от случая принимает одно из возможного множества своих значений, причем заранее неизвестно, какое именно.

Примеры:

- число очков, выпавших при броске игральной кости;
- число появлений герба при 10 бросках монеты;
- число выстрелов до первого попадания в цель;
- расстояние от центра мишени до пробойны при попадании.

Случайные величины подразделяются на две группы: *дискретные* и *непрерывные*.

### Дискретные случайные величины

Случайная величина называется **дискретной** (ДСВ), если множество  $\{x_1, x_2, \dots, x_n, \dots\}$  ее возможных значений конечно или счетно (т.е. если все ее значения можно занумеровать).

Такие из перечисленных выше случайных величин, как количество очков, выпадающих при бросании игрального кубика, число посетителей аптеки в течение дня, количество яблок на дереве являются дискретными случайными величинами.

Наиболее полную информацию о дискретной случайной величине дает **закон распределения** этой величины.

**Закон распределения** – это соответствие между всеми возможными значениями этой случайной величины и соответствующими им вероятностями.

Закон распределения дискретной случайной величины часто задают в виде двухстрочной таблицы, в первой строке которой перечислены все возможные значения этой величины (в порядке возрастания), а во второй – соответствующие этим значениям вероятности:

$X$	$x_1$	$x_2$	$\dots$	$x_n$
$P$	$p_1$	$p_2$	$\dots$	$p_n$

**Пример.** Два стрелка делают по одному выстрелу по мишени. Вероятности их попадания при одном выстреле равны соответственно 0,6 и 0,7. Составить ряд распределения случайной величины  $X$  – числа попаданий после двух выстрелов.

**Решение.** Очевидно, что  $X$  может принимать три значения: 0, 1 и 2. Найдем их вероятности: Пусть события  $A_1$  и  $A_2$  – попадание по мишени соответственно первого и второго стрелка. Тогда

$$P(X = 0) = P(\overline{A_1} \overline{A_2}) = 0.4 * 0.3 = 0.12$$

$$P(X = 1) = P(\overline{A_1} A_2 + A_1 \overline{A_2}) = 0.4 * 0.7 + 0.6 * 0.3 = 0.46$$

$$P(X = 2) = P(A_1 A_2) = 0.6 * 0.7 = 0.42$$

Следовательно, ряд распределения имеет вид:

$x_i$	0	1	2
$p_i$	0.12	0.46	0.42

Для описания определенных особенностей дискретной случайной величины используют ее *основные числовые характеристики*: математическое ожидание, дисперсию и среднее квадратическое отклонение (стандарт).

**Математическим ожиданием**  $M(X)$  (используется также обозначение « $\mu$ ») дискретной случайной величины  $x$  называется сумма произведений каждого из всех ее возможных значений на соответствующие вероятности:

$$M(x) = \mu = \sum_{i=1}^n x_i p_i = x_1 p_1 + x_2 p_2 + \dots x_n p_n$$

Основной смысл математического ожидания дискретной случайной величины состоит в том, что оно представляет собой *среднее значение* данной величины. Другими словами, если произведено некоторое количество испытаний, по результатам которых найдено среднее арифметическое всех наблюдавшихся значений дискретной случайной величины  $X$ , то это среднее арифметическое приближенно равно (тем точнее, чем больше количество испытаний) математическому ожиданию данной случайной величины. Приведем *некоторые свойства математического ожидания*.

1. Математическое ожидание постоянной величины равно этой постоянной величине:

$$M(C) = C$$

2. Математическое ожидание произведения постоянного множителя на дискретную случайную величину равно произведению этого постоянного множителя на математическое ожидание данной случайной величины:

$$M(kX) = kM(X)$$

3. Математическое ожидание суммы двух случайных величин равно сумме математических ожиданий этих величин:

$$M(X+Y) = M(X) + M(Y)$$

4. Математическое ожидание произведения независимых случайных величин равно произведению их математических ожиданий:

$$M(X \cdot Y) = M(X) \cdot M(Y)$$

Отдельные значения дискретной случайной величины группируются около математического ожидания как центра. Для характеристики степени разброса возможных



значений дискретной случайной величины относительно ее математического ожидания вводят понятие *дисперсии дискретной случайной величины*:

**Дисперсией**  $D(X)$  (используется также обозначение « $\sigma^2$ ») дискретной случайной величины  $X$  называется математическое ожидание квадрата отклонения этой величины от ее математического ожидания:

$$D(X) = \sigma^2 = M((X - \mu)^2),$$

На практике дисперсию удобнее вычислить по формуле

$$D(X) = \sigma^2 = M(X^2) - \mu^2,$$

Перечислим основные свойства дисперсии.

1. *Дисперсия постоянной величины равна нулю:*

$$D(C) = 0$$

2. *Дисперсия любой случайной величины есть число неотрицательное:*

$$D(X) \geq 0$$

3. *Дисперсия произведения постоянного множителя  $k$  на дискретную случайную величину равна произведению квадрата этого постоянного множителя на дисперсию данной случайной величины:*

$$D(kX) = k^2 \cdot D(X).$$

В вычислительном отношении более удобна не дисперсия, а другая мера рассеивания случайной величины  $X$ , которая чаще всего и используется – *среднее квадратическое отклонение (стандартное отклонение или просто стандарт)*.

**Средним квадратическим отклонением** дискретной случайной величины называется квадратный корень из ее дисперсии:

$$\sigma(x) = \sqrt{D(X)}$$

Удобство стандартного отклонения состоит в том, что оно имеет размерность самой случайной величины  $X$ , в то время как дисперсия имеет размерность, представляющую квадрат размерности  $X$ .

### Непрерывные случайные величины

Случайная величина называется **непрерывной** (НСВ), если множество ее возможных значений целиком заполняет некоторый конечный или бесконечный интервал или системы интервалов на числовой оси.

Например, непрерывными случайными величинами являются: температура больного в фиксированное время суток, масса наугад выбранной таблетки некоторого препарата, рост наугад выбранного студента и др.

Непрерывную случайную величину нельзя задать в виде таблицы ее закона распределения, поскольку невозможно перечислить и выписать в определенной последовательности все ее значения, а также потому, что вероятность любого конкретного значения непрерывной случайной величины равна нулю.

Одним из возможных способов задания непрерывной случайной величины является использование с этой целью соответствующей **функции распределения**.

Функция  $F(x)$ , равная вероятности того, что случайная величина  $X$  в результате испытания примет значение, меньшее  $x$ , называется **функцией распределения** данной случайной величины:

$$F(x) = P(X < x),$$

**Свойства функции распределения:**

1. *Функция распределения удовлетворяет неравенству:*

$$0 \leq F(x) \leq 1$$

2. Функция распределения является неубывающей функцией, т.е. из  $x_2 > x_1$  следует  $F(x_2) \geq F(x_1)$ .

3. Функция распределения стремится к 0 при неограниченном убывании ее аргумента и стремится к 1 при его неограниченном возрастании.

**Плотностью распределения вероятностей** (плотностью вероятности)  $f(x)$  непрерывной случайной величины  $X$  называется производная функции распределения  $F(x)$  этой величины:

$$f(x) = F'(x)$$

Под основными числовыми характеристиками непрерывной случайной величины понимают, как и в случае дискретной случайной величины, математическое ожидание, дисперсию и среднее квадратическое отклонение.

**Математическое ожидание** непрерывной случайной величины:

$$M(X) = \mu = \int_{-\infty}^{+\infty} xf(x)dx$$

**Дисперсия** непрерывной случайной величины:

$$D(X) = \sigma^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x)dx$$

**Среднее квадратическое отклонение**, как и для дискретной случайной величины, определяется формулой:

$$\sigma(x) = \sqrt{D(X)}$$

### **Законы распределения случайных величин.**

**Биномиальное распределение.** Дискретная случайная величина  $X$  имеет *биномиальное распределение*, если ее возможные значения  $0, 1, 2, \dots, m, \dots, n$ , а соответствующие им вероятности равны:

$$P_m = P\{X = m\} = C_n^m p^m q^{n-m},$$

где  $0 < p < 1$ ,  $q = 1 - p$ ;  $m = 0, 1, 2, \dots, n$ .

Как видно из формулы, вероятности  $P_m$  вычисляются, как члены разложения бинома Ньютона  $(p + q)^n$ , откуда и название «биномиальное распределение».

Примером является выборочный контроль качества производственных изделий, при котором отбор изделий для пробы производится по схеме случайной *повторной выборки*, т.е. когда проверенные изделия возвращаются в исходную партию. Тогда количество нестандартных изделий среди отобранных есть случайная величина с биномиальным законом распределения вероятностей.

Биномиальное распределение определяется двумя параметрами:  $n$  и  $p$ . Случайная величина, распределенная по биномиальному закону, имеет следующие основные числовые характеристики:

$$m = np, D = npq, \sigma = \sqrt{npq}.$$

**Распределение Пуассона.** Дискретная случайная величина  $X$  имеет *распределение Пуассона*, если она имеет бесконечное счетное множество возможных значений  $0, 1, 2, \dots, m, \dots$ , а соответствующие им вероятности определяются формулой:

$$P_m = \frac{a^m}{m!} e^{-a}, \quad m = 0, 1, 2, \dots$$

Примерами случайных явлений, подчиненных закону распределения Пуассона, являются: последовательность радиоактивного распада частиц, последовательность отказов при работе сложной компьютерной системы, поток заявок на телефонной станции и многие другие.

Закон распределения Пуассона (23) зависит от одного параметра  $a$ , который одновременно является и математическим ожиданием, и дисперсией случайной величины  $X$ , распределенной по закону Пуассона. Таким образом, для распределения Пуассона имеют место следующие основные числовые характеристики:

$$m = D = a, \quad \sigma = \sqrt{a}.$$

**Геометрическое распределение.** Дискретная случайная величина  $X$  имеет *геометрическое распределение*, если ее возможные значения  $0, 1, 2, \dots, m, \dots$ , а вероятности этих значений:

$$P_m = q^m p,$$

где  $0 < p < 1$ ,  $q = 1 - p$ ;  $m = 0, 1, 2, \dots$ .

Вероятности  $P_m$  для последовательных значений  $m$  образуют геометрическую прогрессию с первым членом  $p$  и знаменателем  $q$ , откуда и название «геометрическое распределение».

В качестве примера рассмотрим стрельбу по некоторой цели *до первого попадания*, причем вероятность попадания при каждом выстреле не зависит от результатов предыдущих выстрелов и сохраняет постоянное значение  $p$  ( $0 < p < 1$ ). Тогда количество произведенных выстрелов будет случайной величиной с геометрическим распределением вероятностей.

Геометрическое распределение определяется одним параметром  $p$ . Случайная величина, подчиненная геометрическому закону распределения, имеет следующие основные числовые характеристики:

$$m = q / p, \quad D = q / p^2, \quad \sigma = \sqrt{q} / p.$$

**Гипергеометрическое распределение.** Дискретная случайная величина  $X$  имеет *гипергеометрическое распределение* с параметрами  $a, b, n$ , если ее возможные значения  $0, 1, 2, \dots, m, \dots, a$  имеют вероятности:

$$P_m = P\{X = m\} = (C_a^m C_b^{n-m}) / C_{a+b}^n, \quad m = 0, \dots, a.$$

Гипергеометрическое распределение возникает, например, когда из урны, содержащей  $a$  черных и  $b$  белых шаров, вынимают  $n$  шаров. Случайной величиной,

подчиненной гипергеометрическому закону распределения, является число белых шаров среди вынутых. Основные числовые характеристики этой случайной величины:

$$m = na / (a + b), \quad D = nab / (a + b)^2 + n(n - 1) \left[ \frac{a}{a + b} \cdot \frac{a - 1}{a + b - 1} - \frac{a^2}{(a + b)^2} \right].$$

**Равномерное распределение.** Непрерывная величина  $X$  распределена равномерно на интервале  $(a, b)$ , если все ее возможные значения находятся на этом интервале и плотность распределения вероятностей постоянна:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{при } x \in (a, b), \\ 0 & \text{при } x \notin (a, b). \end{cases}$$

Для случайной величины  $X$ , равномерно распределенной в интервале  $(a, b)$  (рис. 1), вероятность попадания в любой интервал  $(x_1, x_2)$ , лежащий внутри интервала  $(a, b)$ , равна:

$$P\{x_1 < X < x_2\} = \frac{x_2 - x_1}{b - a}. \quad (30)$$

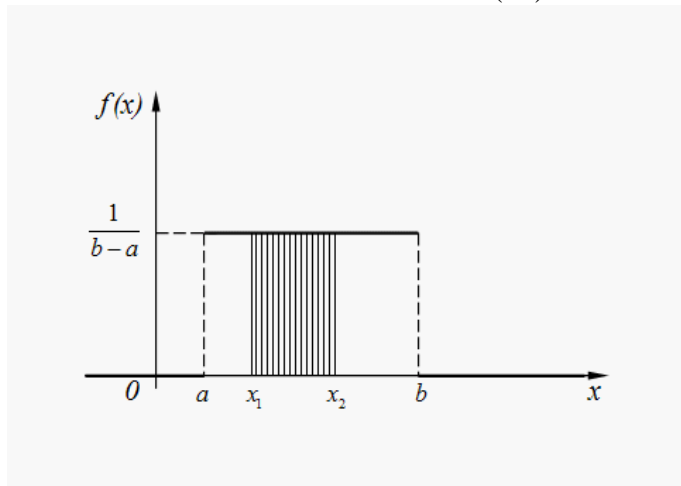


Рис. 1. График плотности равномерного распределения

Примерами равномерно распределенных величин являются ошибки округления. Так, если все табличные значения некоторой функции округлены до одного и того же разряда  $10^{-m}$ , то выбирая наугад табличное значение, мы считаем, что ошибка округления выбранного числа есть случайная величина, равномерно распределенная в интервале  $(-\varepsilon, +\varepsilon)$ , где  $\varepsilon = 0.5 \cdot 10^{-m}$ .

**Показательное распределение.** Непрерывная случайная величина  $X$  имеет *показательное распределение*, если плотность распределения ее вероятностей выражается формулой:

$$f(t) = \begin{cases} \lambda \cdot e^{-\lambda t} & \text{при } t > 0, \\ 0 & \text{при } t \leq 0. \end{cases}$$

График плотности распределения вероятностей представлен на рис. 2.

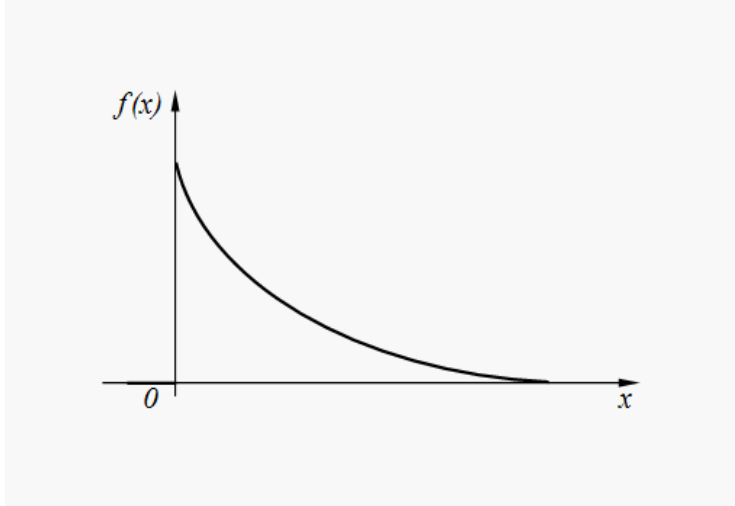


Рис. 2. График плотности показательного распределения

Время  $T$  безотказной работы компьютерной системы есть случайная величина, имеющая показательное распределение с параметром  $\lambda$ , физический смысл которого – среднее число отказов в единицу времени, не считая простоев системы для ремонта.

**Нормальное (гауссово) распределение.** Случайная величина  $X$  имеет *нормальное (гауссово) распределение*, если плотность распределения ее вероятностей определяется зависимостью:

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

где  $m = M(X)$ ,  $\sigma = \sqrt{D(X)}$ .

При  $m = 0$ ,  $\sigma = 1$  нормальное распределение называется *стандартным*.

График плотности нормального распределения представлен на рис. 3.

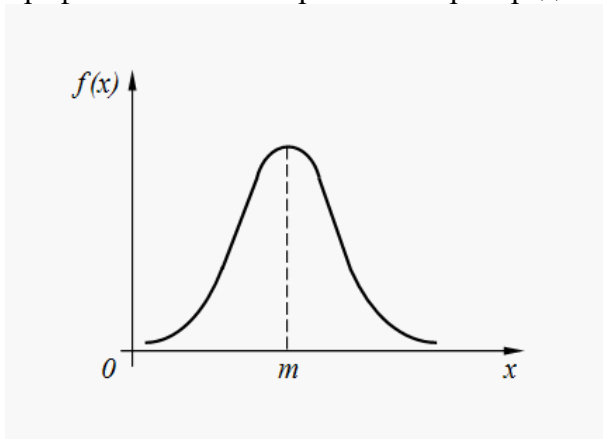


Рис. 3. График плотности нормального распределения

Нормальное распределение является наиболее часто встречающимся в различных случайных явлениях природы. Так, ошибки выполнения команд автоматизированным устройством, ошибки вывода космического корабля в заданную точку пространства,

ошибки параметров компьютерных систем и т.д. в большинстве случаев имеют нормальное или близкое к нормальному распределение. Более того, случайные величины, образованные суммированием большого количества случайных слагаемых, распределены практически по нормальному закону.

**Гамма-распределение.** Случайная величина  $X$  имеет гамма-распределение, если плотность распределения ее вероятностей выражается формулой:

$$f(x) = \begin{cases} \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} & \text{при } x > 0, \\ 0 & \text{при } x \leq 0. \end{cases}$$

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t} dt$$

где  $\Gamma(\alpha)$  – гамма-функция Эйлера.

Основные свойства гамма-функции:

$$1) \Gamma(\alpha + 1) = \alpha \cdot \Gamma(\alpha);$$

$$2) \Gamma(1) = \int_0^{\infty} e^{-t} dt = 1,$$

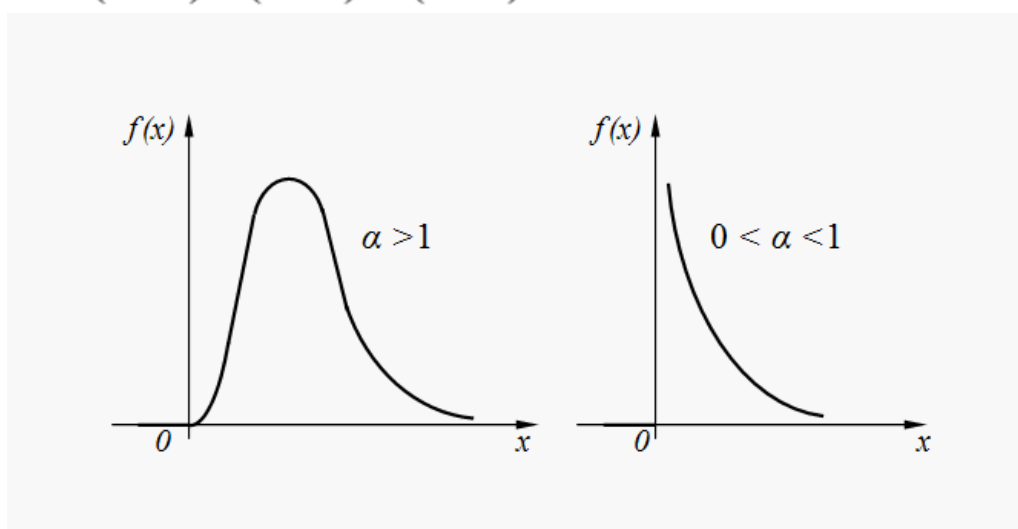
откуда для любого целого  $n > 0$

$$\Gamma(n+1) = n \cdot \Gamma(n) = n \cdot (n-1) \cdot \dots \cdot 2 \cdot 1 \cdot \Gamma(1) = n!;$$

$$3) \Gamma\left(\frac{1}{2}\right) = \int_0^{\infty} e^{-t} \frac{dt}{\sqrt{t}} = \sqrt{\pi},$$

откуда для любого целого  $n > 0$

$$\Gamma\left(n + \frac{1}{2}\right) = \left(n - \frac{1}{2}\right) \cdot \Gamma\left(n - \frac{1}{2}\right) = \frac{1 \cdot 3 \cdot 5 \cdot \dots \cdot (2n-1)}{2^n} \cdot \sqrt{\pi}.$$



Параметры  $\alpha, \lambda$  – любые положительные числа. Гамма-распределение является также [распределением Пирсона типа III](#). При  $\alpha = 1$  гамма-распределение превращается в показательное распределение с параметром  $\lambda$ , так как  $\Gamma(1) = 1$ . Гамма-распределение

широко используется в математической статистике. На рис. 4 представлены графики плотности гамма-распределения (33) при  $\alpha > 1$  и  $0 < \alpha < 1$ .

Рис. 4. Графики плотности гамма-распределения

### Нормальный закон распределения

Из известных видов распределения непрерывных случайных величин наиболее часто используют **нормальное распределение**, которое задается **законом Гаусса**. К нормальному закону распределения при весьма часто встречающихся условиях приближаются другие законы. Так, если мы имеем сумму большого числа независимых величин, подчиненных каким угодно законам распределения, то при некоторых общих условиях она будет приближенно подчиняться нормальному закону.

Непрерывная случайная величина называется **распределенной по нормальному закону (закону Гаусса)**, если ее плотность вероятности имеет вид

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

где  $\mu$  - математическое ожидание;  $\sigma^2$  - дисперсия;  $\sigma$  - среднее квадратическое отклонение этой величины.

### 3. Цель деятельности аспирантов на занятии:

**Аспирант должен знать:**

1. Что изучает Теория вероятности.
2. Что основные определения теории вероятности.
3. Основные формулы теории вероятности.

**Аспирант должен уметь:**

1. Вычислять вероятность случайного события.
2. Находить Сумму и произведение вероятностей.
3. Находить Математическое ожидание, Дисперсию и Среднее квадратическое отклонение дискретной случайной величины.

### 4. Содержание обучения:

1. Элементы теории вероятности.
  - 1.1. Основные понятия теории вероятности.
  - 1.2. Классическое определение вероятности.
  - 1.3. Основные теоремы.
  - 1.4. Формула полной вероятности.
  - 1.5. Случайные величины.
  - 1.6. Нормальный закон распределения.

## 5. Перечень вопросов для проверки уровня знаний:

1. Теория вероятности. Предмет и методы.
2. Формула вероятности случайного события.
3. Формула полной вероятности.

## 6. Перечень вопросов для проверки конечного уровня знаний:

1. Дайте определение теории вероятности.
2. Что такое испытание, случайное событие. Свойства случайного события.
3. Классическое определение вероятности.
4. Сформулируйте теоремы сложения и умножения.
5. Формула полной вероятности.
6. Случайная величина, виды случайных величин.
7. Дайте определения математического ожидания, дисперсии, среднего квадратического отклонение дискретной случайной величины.
8. Перечислите законы распределения случайных величин и коротко охарактеризуйте.

## Практическая часть

### Основные законы распределения. Вероятностный калькулятор

Вероятностный калькулятор – процедура, предназначенная для работы с наиболее известными законами распределения. Используя его, можно строить графики интегральной и дифференциальной функций распределения, для непрерывных случайных величин – вычислить процентные точки, определить вероятность попадания значений в заданный интервал, для дискретных случайных величин – вычислить вероятности и строить ряды распределения.

Нормальное распределение – наиболее важный закон распределения непрерывных случайных величин. С помощью нормального распределения можно описать большинство явлений окружающего мира, например, распределение некоторых физических параметров представителей животного, растительного мира. Нормальное распределение используется для моделирования экономических процессов – распределение заработной платы, налоговых поступлений, продолжительности жизни и т.д. нормальный закон также находит широкое применение для приближения распределения дискретных случайных величин – объёмов производства или продаж того или иного вида продукции, числа посетителей тех или иных учреждений и т.д. иногда это распределение называют распределением ошибок, так как ошибки всевозможных измерений также приближаются нормальным законом. Главной особенностью нормального распределения, выделяющего его среди других, является то, что оно – предельный закон, к которому приближаются другие законы распределения при выполнении определённых условий. Если из значений нормально распределённой случайной величины вычесть математическое ожидание и разделить на стандартное отклонение, то полученные случайные величины имеют стандартное нормальное распределение.

**Распределение  $\chi^2$  (Пирсона).** Случайная величина, имеющая распределение  $\chi$  с  $k$  степенями свободы, определяется как сумма квадратов  $k$  независимых случайных величин со стандартным нормальным распределением. В частном случае, когда  $k=1$ , случайная величина  $\chi^2$  равна квадрату стандартной нормальной



величины. Это распределение асимметрично, обладает положительной правосторонней асимметрией (сосредоточено только на положительной полуоси). При увеличении числа степеней свободы пик плотности распределения уменьшается и смещается вправо. Это распределение играет важную роль при проверке зависимостей в таблицах сопряжённости и в критериях согласия.

**Распределение Стьюдента (*t*-распределение).** Случайная величина, имеющая *t*-распределение с  $k$  степенями свободы, определяется как отношение случайной величины со стандартным нормальным распределением на корень квадратный из среднего арифметического квадратов  $k$  случайных величин, имеющих также нормальное стандартное отклонение. Кривая *t*-распределение, как и стандартная нормальная кривая, симметрична относительно оси ординат, но по сравнению с нормальной более пологая. При увеличении  $k$  это распределение приближается к нормальному. Данное распределение применяется при оценке среднего, в регрессионном анализе, при использовании временных рядов.

**F-распределение Фишера.** Случайная величина, имеющая *F*-распределение с парой степеней свободы  $m$  и  $n$ , определяется как отношение двух независимых случайных величин, имеющих распределение  $\chi^2$  со степенями свободы  $m$  и  $n$ , умноженным на нормировочный сомножитель  $n/m$ . Распределение асимметрично, обладает положительной правосторонней асимметрией. При увеличении  $m$  и  $n$  распределение приближается к нормальному. Распределение Фишера используется при оценке дисперсии случайной величины, в регрессионном, дисперсионном и дискриминантном анализе, а также в других видах многомерного анализа данных.

**Логарифмически-нормальное распределение.** Неотрицательная случайная величина  $X$  имеет логарифмически-нормальное (логнормальное) распределение, если случайная величина  $\ln(X)$  имеет нормальное распределение. Кривая плотности распределения асимметрична и располагается на положительной полуоси. Логнормальное распределение используется для описания распределения доходов, банковских вкладов, месячной заработной платы, посевных площадей под разные культуры, долговечности изделий в режиме износа и старения и др.

**Биномиальное распределение.** Биномиальное распределение представляет собой закон распределения числа наступлений  $m$  некоторого события  $A$  в  $n$  независимых испытаниях, в каждом из которых оно может произойти с одной и той же вероятностью  $p(P_n(m) = C_n^m p^m (1-p)^{n-m})$ . Этот закон распределения широко используется в теории и практике статического контроля качества продукции, при моделировании систем массового обслуживания, в теории стрельбы и других областях.

**Распределение Пуассона.** Дискретная случайная величина  $X$  имеет распределение Пуассона, если она принимает значения  $0, 1, 2, \dots$  с вероятностями  $P(X = m) = h^m e^{-h} / m!$  где  $m = 0, 1, 2, \dots$ . Закон распределения Пуассона является хорошим приближением биномиального распределения при достаточно больших  $n$  и малых значениях вероятности  $p$  (при условии, что произведение  $n p$  — постоянная величина). По закону Пуассона распределён, например, число рождений четверней, число сбоев на автоматической линии, число отказов сложной системы в «нормальном режиме», число требований на обслуживание, поступивших в единицу времени в системах массового обслуживания и т.д.

**Распределение Бернулли.** Случайная величина имеет распределение Бернулли, если она принимает значение 0 или 1 с вероятностями  $P(X = m) = p^m(1-p)^{1-m}$ , где  $m \in \{0,1\}$ ,  $p$  – вероятность наступления события. Это распределение наилучшим образом описывает ситуации, где результатами являются успех или неуспех.

**Геометрическое распределение.** Дискретная случайная величина имеет геометрическое распределение, если она принимает значения  $m = 1, 2, \dots$  с вероятностями  $P(X = m) = (1-p)^{m-1}p$ , где  $m$  – число неуспехов;  $p$  – вероятность успеха в одном испытании. Это распределение используют тогда, когда моделируют ситуации, в которых испытания проводятся до первого наступления успеха.

Рассмотрим основные принципы работы процедуры **Probability calculator** (вероятный калькулятор). Для запуска процедуры надо в модуле **Basic Statistics/Tables** выбрать команду **Probability calculator Distributions**. Откроется рабочее окно команды **Probability Distributions Calculator** (калькулятор вероятностных распределений).

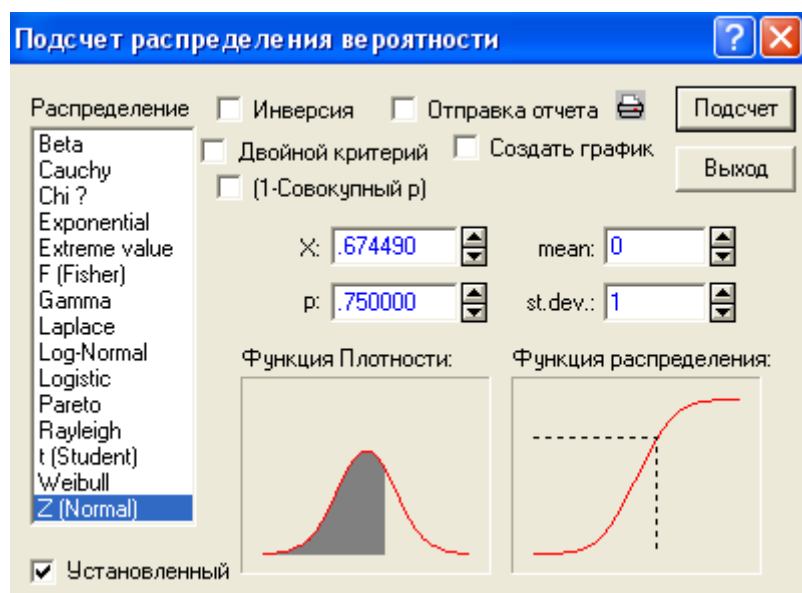


Рис.1

В левой части расположен список распределений **Distributions** (распределение). Многие стандартные распределения в этом окне можно выбрать, высвечивая их названия в списке слева: Бета, Коши,  $\chi^2$ , нормальное, логнормальное, распределение Стьюдента и т.д. Выберите, например, в списке нижнюю строчку *Z Normal* (нормальное распределение). Автоматически справа появляются поля, где можно задать параметры нормального распределения: *mean* (среднее) и *std.dev* (стандартное отклонение) рис1. По умолчанию система запишет в них стандартные значения: среднее = 0, стандартное отклонение = 1. Данные значения можно изменить: надо поместить курсор мыши в эти поля, щёлкнуть левой кнопкой и ввести с клавиатуры нужные величины. Одновременно с выбором распределения в левом списке справа в калькуляторе появятся графики нормальной плотности и функции распределения: *Density Function* (функция плотности), *Distribution Function* (функция распределения). В поле  $p$  надо задать уровень вероятности, при этом флажок автоматически установится на *Inverse* (инверсия). После нажатия на кнопку *Compute* (подсчёт) (в правом верхнем углу калькулятора) в поле  $Z$  появится значение квантиля,

соответствующее выбранному уровню вероятности. То же можно сделать и в обратную сторону – по заданному значению  $Z$  вычислить уровень вероятности  $p$ . Для этого надо задать значение квантиля, щёлкнуть по кнопке *Compute*; в поле  $p$  можно сделать и в обратную сторону – по заданному значению  $Z$  вычислить уровень вероятности  $p$ . Для этого надо задать значение квантиля щёлкнуть по кнопке *Compute*; в поле  $p$  появиться значение вероятности, соответствующее данному значения  $Z$ . Если установить флажок на *CreateGraph* (создать график) и нажать на кнопку *Compute*, то на экране появятся графики плотности и функции распределения (рис.2) с выделенными на них значениями вероятности и квантили.

Если установить флажок на two-tailed (двойной критерий), то расчёт будет проведён для отрезка  $[m-x; m+x]$  (где  $m$ -среднее значение), в противном случае – для отрезка  $[-\infty; x]$ . Если установить флажок на 1-Cumulativep(1-совокупный  $p$ ), то расчёт будет проведён для отрезка, противоположного указанному. Так, например, если рассматривался отрезок  $[-\infty; x]$ , то расчёт будет проведён для отрезка  $[x; +\infty]$ .

Флажок на Fixed Scaling (фиксированная шкала) под списком распределений Distributions указывает, что выбрана фиксированная шкала.

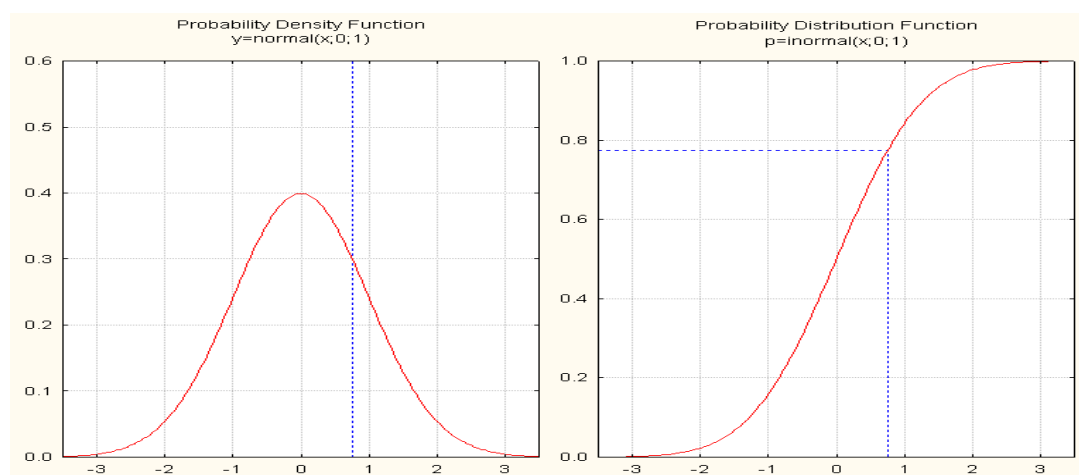


Рис.2

Помимо вычисления уровня вероятности, квантили и построения кривых распределений, **Probability calculator** может быть использован для изучения поведения кривых распределений при изменении параметров распределений, а так же для решения некоторых задач. Так, например, увеличивая mean(среднее) нормального распределения, можно увидеть, как кривая плотности нормального распределения сдвигается по оси ординат вправо. При увеличении стандартного отклонения плотность нормального распределения расплывается или рассеивается относительно среднего значения. При уменьшении они, наоборот, сжимается, концентрируясь возле одной точки – точки максимального значения.

Известно, что рост студентов имеет нормальное распределение со средним 175,6 см и стандартным отклонением 7,63 см. Произвольным образом выбирается студент, например, первый вошедший в аудиторию. Какова вероятность, что рост этого студента не больше 185 см и не меньше 175 см?

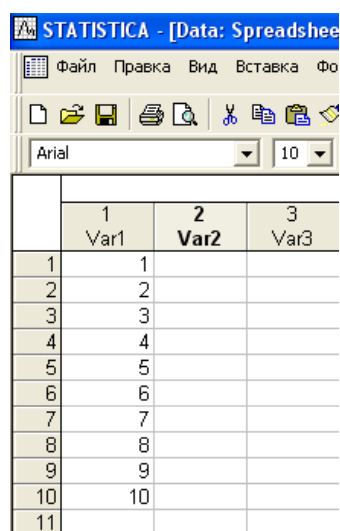
Выберите в списке распределений Z Normal. Задайте в поле mean – 175,6 в поле std.dev.- 7,63. В поле X-185. Нажмите кнопку *Подсчёт*. В поле p появится значение 0,891022. Запомните его как p.

В поле X задайте 175. Нажмите кнопку *Compute*. В поле p появится значение 0,468661. Запомните это значение как  $p_2$  Вычтите  $p_2$  из  $p_1$ .Получите 0,422361. Итак, с вероятностью 0,422361 случайный студент имеет рост не ниже 175 и не выше 185 см.

Рассмотрим решение задачи с использованием дискретного распределения.

В среднем 30% студентов сдают экзамен по дискретному программированию на отлично. Найдите вероятность того, что в группе, состоящей из 15 человек, не более 5 человек получают отлично.

Создайте пустую электронную таблицу. В первом столбце переменной **Var1**проставьте возможное число студентов, сдавших на отлично (количество испытаний).



	1 Var1	2 Var2	3 Var3
1	1		
2	2		
3	3		
4	4		
5	5		
6	6		
7	7		
8	8		
9	9		
10	10		
11			

Рис.3

Дважды щёлкните по имени переменной **Var2**. В нижней части окна в поле **Longname** запишите функцию с указанием параметров.

Запись должна начинаться со знака =. Функцию можно также выбрать из списка, нажав на кнопку **Functions**. В списке предложенных функций выберите нужную функцию (в данном случае Binom) и два раза щёлкните по ней. Как видно из подсказки, функция Binom (x;p;n) использует три параметра, которые перечислены в круглых скобках через точку с запятой. Первый параметр x – ссылка на переменную, в строках которой указано количество проводимых испытаний (в нашем случае «V1»). Второй параметр p – вероятность удачного исхода в одном испытании (в нашем случае  $p=0,3$  – вероятность сдать экзамен на отлично). Третий параметр  $n=15$  – количество испытаний. Нажмите **ОК**.

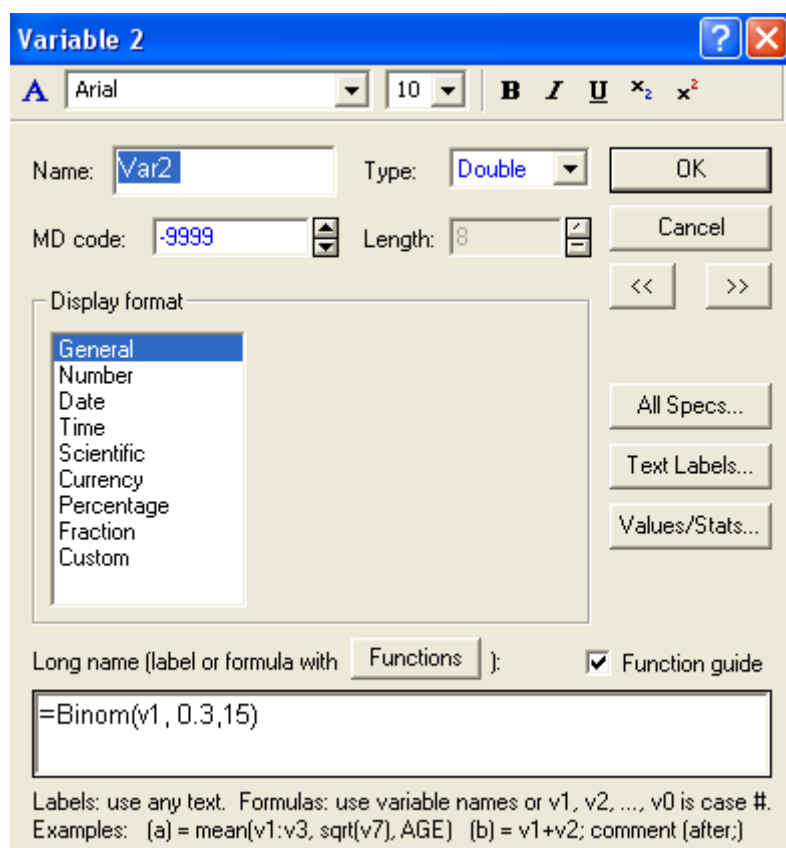


Рис.4

Согласно формуле Бернулли программа вычислит вероятность успеха и занесёт их в столбец таблицы, соответствующий второй переменной Var2. Для определения искомой вероятности надо выделить курсором мыши первые 5 элементов столбца Var2, далее щёлкнуть правой кнопкой мыши и в открывшемся контекстном меню щёлкнуть на *Statistics of Block Data* → *Block Columns* → *Sums* (статистика блока данных → блок столбцов → суммы). Получите значение вероятности  $p = 0,72$  того, что не более 5 студентов сдадут экзамен на отлично.

Рассмотрим решение задачи с использованием распределения Пуассона.

На факультете прикладной математики обучаются 685 студентов. Какова вероятность того, что 7 октября является днем рождения одновременно 7 студентов факультета, днем рождения более чем 2 студентов. Вероятность того, что день рождения студента 7 октября, равна  $p = 1/365$ . Так как вероятность  $p = 1/365$  – мала, а  $n = 685$  велико, применим формулу Пуассона при  $\lambda = n p = 685/365 \approx 1,88$ . Создайте пустую электронную таблицу. В первом столбце переменной Var1 проставьте возможное число студентов – 0, 1, 2..., 7. Дважды щёлкните по имени переменной Var2. Откроется диалоговое окно спецификации Var2. В нижней части окна в поле **Longname** запишите функцию Poissonc указанием параметров. Этих параметров всего два. Первый, как и в предыдущем примере, - ссылка на количество успешных испытаний (Var1), а второй –  $\lambda = 1,88$ . Нажмите ОК.

Согласно формуле Пуассона программа вычислит вероятности и занесёт их в столбец, соответствующий второй переменной Var2. Получите, что вероятность того, что день рождения семи студентов из 685 придётся на 7 октября мала и составит  $p = 0,0025$ . для нахождения вероятности того, что 7 октября является днём

рождения более чем у 2 студентов, надо из единицы вычесть сумму первых трёх элементов столбца Var2. Получите вероятность  $p = 0,29$ .

Рассмотрим вариант решения задачи с использованием геометрического распределения.

Для изучения вкусов и предпочтений студентов на факультете прикладной математики проведены маркетинговые исследования. Исследования показали, что 65% студентов предпочитают по утрам пить растворимый кофе, 15% - натуральный кофе, остальные 20% пьют чай. Компания *Nescafe* решила провести повторные исследования среди любителей растворимого кофе для определения того, каким сортам кофе студенты отдают наибольшее предпочтение. Потенциальных участников опроса выбирали случайным образом. Какова вероятность того, что только  $k$ -й из опрошенных является любителем растворимого кофе ( $k$  может принимать любое из значений 1,2,3...). Только  $k$ -й означает, что все опрошенные до него, начиная с 1-го заканчивая  $k-1$ -м, не являются любителями кофе.

Создайте пустую электронную таблицу. Для обозначения числа неудач используйте столбец переменной Var1. Впишите в него значения 0,1,...9. Число неудач может быть сколь угодно большим, но ограничимся наибольшим значением 9. В столбец Var2 программа запишет посчитанные вероятности. Дважды щёлкните по имени переменной Var2. Откроется диалоговое окно спецификации переменной Var2. В нижней части окна в поле **Longname** запишите функцию Geom ( $x;p$ ) с указанием параметров. Этих параметров всего два. Первый параметр  $x$ -ссылка на переменную, в строках которой указано количество неудачных испытаний (в нашем случае "V1") , второй – вероятность  $p=0,65$ . нажмите **ОК**. По формуле геометрического распределения программа вычислит вероятности и занесёт их в столбец Var2.

Вероятность ВО будет содержать число проведённых испытаний, завершившихся успехом. Для нашего примера – это число опрошенных, последний из которых окажется любителем растворимого кофе. Так, например, вероятность 0,0097 соответствует случаю, когда всего опросили 5 человек, причём первые 4 – не любители кофе, а последний оказался любителем.

**Задача 1.** Абонент забыл последнюю цифру номера телефона и поэтому набирает её наугад. Определить вероятность того, что ему придётся звонить не более чем в 3 места.

**Задача 2.** Абонент забыл последние 2 цифры телефонного номера, но помнит, что они различны и образуют двузначное число, меньшее 30. С учетом этого он набирает наугад 2 цифры. Найти вероятность того, что это будут нужные цифры.

#### **7. Самостоятельная работа аспиранта.**

Решение медицинских задач с использованием распределения Бернулли.

#### **8. Хронокарта учебного занятия:**

1. Организационный момент – 5 мин.
2. Текущий контроль знаний – 30 мин.
3. Разбор темы – 20 мин.
4. Практическая работа – 30 мин.
5. Подведение итогов занятия – 10 мин.

#### **9. Перечень учебной литературы к занятию:**

1. Кобринский Б.А., Зарубина Т.В. «Медицинская информатика», М., Издательский дом «Академия», 2009.
2. Жижин К.С. «Медицинская статистика», Высшее образование, 2007.

## **ТЕМА 2: Элементы организации медико-статистического исследования.**

### **Статистическая совокупность. Статистические величины. Вычисление статистических величин**

#### **1. Научно-методическое обоснование темы:**

Отечественная медицина всегда признавала большое значение медицинской статистики, особенно для научных медицинских исследований.

Не менее широко пользуются медицинской статистикой и современные ученые-медики. Однако врачам приходится соприкасаться со статистикой не только в научной работе. Почти нет ни одной врачебной должности в больнице (в клинике), в поликлинике, занимая которую, врачу не пришлось бы выполнять какие-нибудь статистические работы. Поэтому представляется важным, чтобы лечащие врачи знали, как взяться за дело, умели собирать и обрабатывать верные цифры, годные для сравнения и сопоставления. К этому можно добавить, что не только врачи-клинисты, но и врачи-экспериментаторы и врачи гигиенисты и организаторы здравоохранения не в меньшей мере нуждаются в умении правильно применять статистические методы исследования и правильно толковать и использовать результаты этих исследований.

Однако умение врачей пользоваться медико-статистическими приемами исследования, к сожалению, значительно отстает от общего признания ими важности статистики для научной медицинской работы и от стремления научных работников-медиков к использованию медико-статистических методов в своих работах. О медицинской статистике часто вспоминают в конце научного исследования, т.е. тогда, когда допущенные ранее статистические ошибки в планировании и организации этого исследования исправить уже почти невозможно. На самом же деле знание основных принципов медико-статистической методики и основных правил ее применения в экспериментальных, клинических и санитарно-гигиенических исследованиях нужно врачу не в конце, а в начале его научной работы. Это знание избавит от трудно исправимых впоследствии ошибок в организации научного наблюдения, поможет определить необходимое для наблюдения или эксперимента количество объектов и предотвратит излишнюю затрату сил, времени и средств, вызванных погоней за «большими числами». Современная статистическая наука позволяет во многих случаях пользоваться так называемой малой выборкой и получать при ее помощи не менее, а иногда и более достоверные результаты, чем от чрезмерно больших и трудоемких статистических работ. Для ознакомления с ней требуется литература, помогающая врачам освоить медико-статистические приемы исследования и, следовательно, рассчитанная не на читателя «вообще», а на врача, ведущего научно-исследовательскую работу.

#### **2. Краткая теория:**

##### **Элементы организации медико-статистического исследования**

В содержании медико-статистического исследования выделяют 4 последовательных этапа:

- 1-й – составление плана и программы исследования;
- 2-й – статистическое наблюдение;
- 3-й – статистическую группировку и сводку наблюдений;
- 4-й – статистическую обработку и анализ полученных материалов, оформление результатов исследования.

В процессе реализации **1-го этапа** исследования формулируется цель исследования, составляется его организационный план, кратко излагается содержание последующих этапов статистического исследования, то есть – программа исследования.

В процессе реализации **2-го этапа** исследования можно точно определить объект и единицу наблюдения, опираясь на четко сформулированную цель исследования. Под **объектом наблюдения** понимается явление, подлежащее исследованию. Правильное проведение наблюдения требует, чтобы объект наблюдения был точно ограничен во времени и пространстве и ясно определено в качественном отношении. **Единица наблюдения** – это отдельный случай изучаемого явления, и эти случаи составляют объект наблюдения.

В процессе реализации **3-го этапа** единицы наблюдения регистрируются сообразно целям исследования по общим и важным качественным характеристикам, называемым **признаками**. Уточнение и формулирование признаков производится на основе следующих общих правил:

1. Признаки отбирают с учетом целей изучения и возможности обработки и анализа полученных при наблюдении данных.
2. Отобранных признаков не должно быть много. Чрезмерное их количество затрудняет наблюдение и препятствует его проведению.
3. Признаки нужно так комбинировать, чтобы они взаимно дополняли и контролировали друг друга, давая возможность легко выявлять допущенные ошибки.
4. Отобранные признаки должны учитывать возможности тех, кто проводит исследование. Некоторые признаки таковы, что для их регистрации необходимо иметь специальную аппаратуру.

А также вычисляют статистических показатели, необходимые для оценки данных, полученных в результате данного статистического исследования, и строят графические изображения, которые служат для наглядного представления статистических величин, позволяет глубже их проанализировать.

На **4-м этапе** медико-статистического исследования проводится статистический анализ данных. Данному анализу могут быть подвергнуты абсолютные, средние и относительные величины, их графические изображения, различные коэффициенты и т.д.

В следующих главах данного пособия, предлагается проводить статистическую обработку, анализ полученных данных и оформление результатов исследования, используя программу Microsoft Excel.

### **Статистическая совокупность. Статистические величины**

Объектом любого статистического исследования является **статистическая совокупность** — группа или множество относительно однородных элементов, т. е. единиц, взятых вместе в конкретных границах времени и пространства и обладающих признаками сходства и различия.

Различают два вида статистической совокупности: **генеральную**, состоящую из всех единиц наблюдения, которые могут быть к ней отнесены в зависимости от цели исследования, и **выборочную** — часть генеральной совокупности, отобранную специальным выборочным методом. Каждую статистическую совокупность можно рассматривать как генеральную и как выборочную.

Целью изучения любой статистической совокупности является выявление общих свойств, общих закономерностей различных явлений, так как эти свойства не могут быть обнаружены при анализе единичных явлений.

Признаки сходства служат основанием для объединения единиц в совокупность, признаки различия, называемые учетными признаками, являются предметом их особого анализа. По своему характеру учетные признаки могут быть качественными (например, пол, профессия). Они могут быть также количественными, выраженными числом (например, возраст).



Одним из типов распределения признака в статистической совокупности является **вариационный ряд** - ряд числовых значений какого-то определенного признака, отличающихся друг от друга по своей величине и расположенных в ранговом порядке. Характеристиками вариационного ряда являются:

- ♦ числовое значение изучаемого признака ( $X$ );
- ♦ частота, с которой встречается каждый признак ( $p$ );
- ♦ общее число наблюдений ( $N$ ).

Вариационный ряд может быть *простым*, где значение каждого признака обозначается отдельно, или *сгруппированным*, где значения признака объединяются в группы с указанием частоты встречаемости признака, входящего в данную группу.

Простой вариационный ряд составляется обычно при малом числе наблюдений ( $N \leq 30$ ), а сгруппированный — при большом числе наблюдений ( $N > 30$ ).

Для оценки изучаемых явлений, составляющих статистическую совокупность, используют следующие статистические средние величины:

**мода** ( $M_o$ ) — величина признака, чаще других встречающегося в совокупности;

**медиана** ( $M_e$ ) — величина, которая делит распределение пополам: половина значений больше медианы, половина меньше.

**среднее по совокупности** ( $M$ ) — величина, равная отношению суммы всех значений признака к их общему числу членов совокупности:

$$M = \frac{\sum_{i=1}^N X_i}{N}, \quad (1)$$

где:  $M$  — среднее по совокупности;

$X_i$  — значение признака одного члена совокупности;

$N$  — число членов совокупности.

**Разнообразие признака в вариационном ряду.** Имеются следующие критерии разнообразия признака:

1. Характеризующие границы совокупности.

2. Характеризующие внутреннюю структуру совокупности - **стандартное отклонение** ( $\sigma$ ) — это показатель разброса значений относительно среднего.

Для того чтобы охарактеризовать разброс значений относительно среднего введем показатель разброса, который носит название **дисперсии** и обозначается  $\sigma^2$ . Ясно, что для характеристики разброса признака все равно в какую сторону отклоняется значение — в большую или в меньшую. Иными словами, отрицательные и положительные отклонения должны вносить равный вклад в характеристику разброса. Воспользуемся тем, что квадраты двух равных по абсолютной величине чисел равны между собой, и вычислим средний квадрат отклонения от среднего, т.е.  $\sigma^2$ . Чем больше разброс значений, тем больше дисперсия.

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - M)^2}{N} \quad (2)$$

Дисперсия измеряется в единицах, равных квадрату единицы измерения соответствующей величины. Поэтому чаще используется квадратный корень из дисперсии - **стандартное отклонение**:

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (X_i - M)^2}{N}} \quad (3)$$

3. **Коэффициент вариации** ( $C_x$ ), который характеризует разнообразие признака в вариационном ряду и вычисляется по формуле:

$$C_x = \frac{\sigma}{M} \cdot 100\% \quad (4)$$

Если $C_x < 10\%$	– слабое разнообразие признака,
$C_x = 10 - 20 \%$	– среднее разнообразие признака,
$C_x > 20 \%$	– сильное разнообразие признака.

### 3. Цель деятельности аспирантов на занятии:

#### Аспирант должен знать:

1. Элементы организации медико-статистического исследования.
2. Статистическая совокупность. Виды.
3. Вариационный ряд. Виды.
4. Статистические средние величины.
5. Критерии разнообразия признака.

#### Аспирант должен уметь:

1. Проводить статистическую обработку, анализ полученных данных и оформление результатов исследования, используя программу Microsoft Excel.

#### Содержание обучения:

##### Теоретическая часть:

1. 4 этапа медико-статистического исследования.
2. Вариационный ряд. Характеристики вариационного ряда.
3. Формула вычисления коэффициента вариации.

##### Практическая часть:

Пример выполнения задания.

**Цель работы:** Определение среднего уровня изучаемого признака - средней величины, использующихся для анализа медицинских данных, заключенных в вариационном ряду.

**Необходимо:**

1. В Excel набрать заданный вариационный ряд статистических данных.
2. Расположить значения признака в ранговом порядке (отсортировать по возрастанию).
3. Вычислить следующие статистические величины с помощью встроенных статистических функций Microsoft Excel: среднее по совокупности, стандартное отклонение, моду, медиану.
4. Коэффициент вариации вычислить по формуле (4).
5. По полученному значению коэффициента вариации определить характер разнообразия признака.

Данные: в результате измерения длины тела у 32 мальчиков при рождении были получены следующие данные (в см): 49, 52, 54, 49, 52, 54, 50, 49, 53, 52, 54, 50, 50, 54, 49, 51, 51, 53, 51, 52, 53, 48, 48, 55, 56, 55, 49, 53, 52, 52, 50, 51.

Введем исходные данные в таблицу Excel и отсортируем их в ранговом порядке по значению признака.

Книга1 - Excel

ФАЙЛ ГЛАВНАЯ ВСТАВКА РАЗМЕТКА СТРАНИЦЫ ФОРМУЛЫ ДАННЫЕ

Буфер обмена Шрифт Выравнивание Число Условное форматирование Форматировать как таблицу Стили ячеек Ячейки Ряд Стили

122 : X ✓ fx

	A	B	C	D	E	F	G	H
1	<b>Задание 1 Вариант 1</b>							
2	<b>Исходные данные:</b>							
3		<b>№</b>	<b>РОСТ, см</b>					
4		1	49					
5		2	52					
6		3	54					
7		4	49					
8		5	52					
9		6	54					
10		7	50					
11		8	49					
12		9	53					
13		10	52					
14		11	54					
15		12	50					
16		13	50					
17		14	54					
18		15	49					
19		16	51					
20		17	51					
21		18	53					
22		19	51					
23		20	52					
24		21	53					
25		22	48					
26		23	48					
27		24	55					
28		25	56					
29		26	55					
30		27	49					
31		28	53					
32		29	52					
33		30	52					
34		31	50					
35		32	51					
36								
37								

Лист1 Лист2 Лист3

Лист1

ГОТОВО 100%

	A	B	C	D	E
1	<b>Задание 1 Вариант 1</b>				
2	<b>Исходные данные:</b>				
3	<b>№</b>	<b>Рост, см</b>			
4	22	48			
5	23	48			
6	1	49			
7	4	49			
8	8	49			
9	15	49			
10	27	49			
11	7	50			
12	12	50			
13	13	50			
14	31	50			
15	16	51			
16	17	51			
17	19	51			
18	32	51			
19	2	52			
20	5	52			
21	10	52			
22	20	52			
23	29	52			
24	30	52			
25	9	53			
26	18	53			
27	21	53			
28	28	53			
29	3	54			
30	6	54			
31	11	54			
32	14	54			
33	24	55			
34	26	55			
35	25	56			
36					
37					

Используя встроенные функции Excel, вычислим значения статистических величин, характеризующих статистическую совокупность и по полученному значению коэффициента вариации определим характер разнообразия признака.

	A	B	C	D	E
1	<b>Задание 1 Вариант 1</b>				
2	<b>Исходные данные:</b>				
3	<b>№</b>	<b>Рост, см</b>	<b>Вычисление статистических величин</b>		
4	1	49	Среднее по совокупности	=СРЗНАЧ(B4:B35)	
5	2	52	Стандартное отклонение	=СТАНДОТКЛОН(B4:B35)	
6	3	54	мода	=МОДА(B4:B35)	
7	4	49	медiana	=МЕДИАНА(B4:B35)	
8	5	52	коэффициент вариации	=E4/E3	
9	6	54			
10	7	50			
11	8	49			
12	9	53			
13	10	52			
14	11	54			
15	12	50			
16	13	50			
17	14	54			
18	15	49			
19	16	51			
20	17	51			
21	18	53			
22	19	51			
23	20	52			
24	21	53			
25	22	48			
26	23	48			
27	24	55			
28	25	56			
29	26	55			
30	27	49			
31	28	53			
32	29	52			
33	30	52			
34	31	50			
35	32	51			

	A	B	C	D	E	F	G
1	<b>Задание 1 Вариант 1</b>						
2	<b>Исходные данные:</b>						
3	<b>№</b>	<b>Рост, см</b>	<b>Вычисление статистических величин</b>				
4	1	49	Среднее по совокупности	51,59375			
5	2	52	Стандартное отклонение	2,16808783			
6	3	54	мода	52			
7	4	49	медiana	52			
8	5	52	коэффициент вариации	4%			
9	6	54					
10	7	50					
11	8	49	Т.к. коэффициент вариации меньше 10%, то можно сказать, что в рассматриваемом случае <b>характер разнообразия признака</b> (роста новорожденных мальчиков) - <b>слабый</b>				
12	9	53					
13	10	52					
14	11	54					
15	12	50					
16	13	50					
17	14	54					
18	15	49					
19	16	51					
20	17	51					
21	18	53					
22	19	51					
23	20	52					
24	21	53					
25	22	48					
26	23	48					
27	24	55					
28	25	56					
29	26	55					
30	27	49					
31	28	53					
32	29	52					
33	30	52					
34	31	50					
35	32	51					

## Порядок выполнения работы.

1. Изучение теоретического материала.
2. Выполнение вариантов заданий с помощью рассмотренных инструментов, средств, приемов и технологий.
3. Составление отчета о проделанной работе. Отчет должен содержать следующие разделы:
  - наименование работы;
  - цель работы;
  - пошаговое последовательное описание процесса выполнения варианта задания по видам выполняемых действий.
4. Результат выполнения варианта задания должен быть сохранен под именем ФИО \_ Работа №\_Вариант№ (например, «ИвановНН\_Работа1\_Вариант1.xls») на жесткий диск в папку «Мои документы\ИТ в медицине» и на дискету – в двух копиях (две копии одной и той же информации в разных папках на дискете).
5. Представление результатов выполнения работы (отчета и файлов на дискете) для проверки преподавателю.
6. Защита выполненной работы: ответ на контрольные вопросы к теоретическому материалу занятия и ответ на замечания преподавателя по выполненной работе.
7. Оценка преподавателем выполненной работы.

С целью изучения распространенности дизентерии и др. острых кишечных инфекций в одной из районов области в 1992-93 гг. проанализированы "экстренные извещения об инфекционном заболевании" (учетная форма № 58).

За эти два года зарегистрировано, дизентерии: 1992 г. - 2349 случаев. 1993 г. - 1205 случаев; колиэнтериты: 1992 г. - 306, 1993 г. - 282; диспепсия простая: 1992 г. - 15, 1993 г. - 12 случаев; диспепсия токсическая. 1992 г. - 11, 1993 г. - 14 случаев.

Население района составляет 56756 человек, из них детей до 15 лет- 14190 чел, 15-19 лет - 6250 чел, 20 лет и старше - 36316 чел. Из общего числа детей посещает детские учреждения 12144, не посещают детские учреждения 2046 чел.

Контрольные вопросы	Ответы
1. Какая исследована совокупность: генеральная или выборочная?	1. Генеральная совокупность: Все население района - 56756 чел (все заболевшие дизентерией или другими острыми кишечными инфекциями в зависимости от цели исследования).
2. Каков объем исследованной совокупности	$\Pi = 56756$ человек
3. Перечислите признаки,	Каждый случай инфекционного

характеризующие исследуемую совокупность. Назовите единицу наблюдения.

заболевания дизентерией или другими острыми кишечными инфекциями в 1992 и 1993 гг.

4. Укажите признаки, по которым различаются элементы статистической совокупности

Заболевшие различными инфекционными заболеваниями (по нозологическим формам), возрастные группы, дети, посещающие детские учреждения

5. Сколько учетных признаков в данной совокупности? Назовите признаки количественные и атрибутивные.

А) по характеру: атрибутивных – два. 1. Дети, посещающие и не посещающие детские учреждения. 2. Донозологические формы количественных - два 1. Годы наблюдения 2. Возрастные группы б) по роли в совокупности: факторных – три 1. Годы наблюдения 2. Возрастные группы 3. посещение детских учреждений результативных – один Заболеваемость инфекционными заболеваниями

6. Репрезентативна ли по качеству изучаемая совокупность?

Здесь имеет место генеральная совокупность, а не выборочная, поэтому вопрос о репрезентативности отпадает

7. Назовите объект исследования.

Население одного из районов области.

8. Назовите вид наблюдения.

Текущее

9. Назовите метод проведения наблюдения

Сплошное наблюдение.

#### **Задание для самостоятельного решения.**

В целях совершенствования организации обслуживания больных дизентерией в 1993 г. в поликлинических отделениях Владикавказа совместно с городским ЦГСЭН проведено изучение влияния сроков постановки диагнозов дизентерии на длительность стационарного лечения.

Результаты исследования показали, что у 100 больных дизентерией диагноз был поставлен в первые три дня заболевания. Длительность пребывания в стационаре этой группы больных колебалась в пределах 15-20 дней. У 30 больных диагноз дизентерии

был поставлен после 3-х дней от начала заболевания. Длительность пребывания их в стационаре колебалась в пределах от 21 до 35 дней.

Опишите статистическую совокупность согласно модели.

**4. Перечень вопросов для проверки исходного уровня знаний:**

- a. План и программа исследования.
- b. Статистическое наблюдение (объект и единица наблюдения).
- c. Статистическая группировка. Вычисление статистических показателей.
- d. Статистический анализ данных.

**5. Перечень вопросов для проверки конечного уровня знаний:**

1. Перечислите этапы медико-статистического исследования.
2. Дайте краткую характеристику первого этапа исследования.
3. Какие задачи решаются на втором этапе исследования.
4. Опишите общие правила уточнения и формулирования признаков на третьем этапе исследования.
5. Охарактеризуйте задачи четвертого этапа исследования.
6. Дайте определение статистической совокупности. Какие бывают виды статистической совокупности.
7. Что является целью изучения статистической совокупности.
8. Дайте определение вариационного ряда.
9. Перечислите характеристики вариационного ряда.
10. Что представляют собой простой и сгруппированный вариационный ряд?
11. Какие статистические величины используются для оценки изучаемых явлений?
12. Каковы критерии разнообразия признака?
13. Что такое дисперсия?
14. Дать определение стандартного отклонения.
15. Что характеризует коэффициент вариации?

**6. Хронокарта учебного занятия:**

1. Организационный момент – 10 мин.
2. Разбор темы – 40 мин.
3. Текущий контроль (тестирование, практическая работа) - 90 мин.
4. Подведение итогов занятия – 10 мин.

**7. Самостоятельная работа аспиранта.**

Освоить медико-статистические приемы исследования. Составить план исследования в зависимости от специализации аспиранта.

**8. Перечень учебной литературы к занятию:**

1. Есауленко И.Э., Семенов С.Н. Основы практической информатики в медицине; Воронеж, 2005.

## **ТЕМА 3: Графический анализ данных. Изучение распределения случайных величин, подчиняющихся нормальному закону распределения Гаусса**

### **1. Научно-методическое обоснование темы:**

В научно-исследовательской работе приходится использовать все формы анализа. Графический анализ является первоначальной формой, которая может подсказать исследователю, какие методы работы нужно применять для получения более точной информации об изучаемых явлениях.

### **2. Краткая теория:**

Цели, которые преследуются при построении графиков, следующие:

1. Представить наглядно сущность и характер изучаемых явлений.
2. Популяризовать результаты статистических исследований.
3. Оказать помощь при анализе изучаемых явлений.

Графики позволяют наглядно представить статистические показатели, полученные при анализе результатов проведенного исследования. Они облегчают сравнение показателей, дают представление о характере связи между явлениями и указывают на тенденции их изменения во времени. Графическое изображение статистических данных в сравнении с табличным позволяет быстро и легко заметить существующие закономерности. Эти последние ярче выражены и подчеркнуты, усваиваются легче и быстрее запоминаются. Вместе с этим связь между статистическими показателями заметна полнее и нагляднее, а скрытые закономерности становятся явственнее. Это создает условия для углубленного исследования и способствует аналитическому мышлению.

В научно-исследовательской работе необходимо хорошо владеть методами графического анализа.

При составлении графиков исследователь должен придерживаться некоторых основных правил. Он должен предварительно внимательно изучить данные, которые следует представить графически. Во-вторых, исследователь должен хорошо знать статистические методы анализа, с помощью которых получены данные. Комбинируя эти два требования, он сможет выбрать наиболее подходящее для данного случая графическое изображение. При таком выборе исследователь принимает во внимание также следующее:

**1. Характер данных.** Нужно отметить, что некоторые данные не поддаются графическому изображению. Для других подходят графики только определенного типа. Так, например, если ставится задача изобразить структуру данного явления, то наиболее подходящими будут секторные диаграммы; для изображения динамики явлений во времени наиболее подходят линейные диаграммы; если изображается сезонность, то наиболее подходят линейно-круговые диаграммы и т. д.

**2. Назначение графиков.** Они могут быть использованы для различных целей (для репродукций в книгах, для лекций, выставок, диапозитивов, кино или телевидения). В зависимости от того, для чего предназначен график, подбирают его величину, характер линий, используемые штрихи и краски, шрифт, величину букв и т. д.

**3. Цель графиков.** Очень часто цель графических изображений – наглядно представить результаты проведенного исследования. В других случаях – подчеркнуть известные закономерности, иллюстрировать новые открытые факты, выдвинуть и обосновать новые гипотезы.

**4. Аудитория, для которой предназначены графические изображения.** Нужно учитывать уровень знаний аудитории, ее практические и научные интересы и пр. Так, например, графики, предназначенные для широкой аудитории, не должны быть сложными



в техническом отношении. Для такой аудитории можно использовать фигурные диаграммы. Графики же, предназначенные для высококвалифицированных специалистов, могут быть технически сложнее.

*Графики можно условно разделить на следующие группы:*

1. Линейные диаграммы.
2. Плоскостные диаграммы.
3. Фигурные диаграммы.
4. Объемные диаграммы.
5. Картограммы.
6. Картодиаграммы.

Если значения интересующего нас признака у большинства объектов близки к их среднему и с равной вероятностью отклоняются от него в большую или в меньшую сторону, то есть на изменение признака оказывают незначительное влияние внешние факторы, то такое распределение учетного признака называется **нормальным (гауссовым)**, описывается формулой:

$$f(X_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X_i - M}{\sigma}\right)^2} \quad (5)$$

и полностью определяется следующими статистическими параметрами:

- ♦ среднее значение  $M$  (см. зан.1, формула 1);
- ♦ стандартное отклонение  $\sigma$  (см. зан.1, формула 3);

Если значения признака распределены несимметрично относительно среднего, то распределение не является нормальным и совокупность лучше описать с помощью:

- ♦ медианы  $Me$ ;
- ♦ процентилей.

Займемся исследованием количественного признака, например роста.

Проведем исследование на планетах Марс и Венера. Отметим, что все жители одной планеты образуют генеральную статистическую совокупность. Для исследования отберем 200 марсиан и 150 венерианцев, где каждая из двух групп инопланетян является выборочной статистической совокупностью, а каждый инопланетянин – единицей наблюдения. Измерим рост каждой единицы наблюдения и запишем в простой вариационный ряд.

В результате получили два простых вариационных ряда: рост марсиан и рост венерианцев. В Microsoft Excel следует набрать получившиеся вариационные ряды. Для наглядности картины распределения роста преобразуем простые вариационные ряды в сгруппированные и расположим значение исследуемого признака (роста) по возрастанию (в ранговом порядке). Сгруппированные ряды представлены ниже:

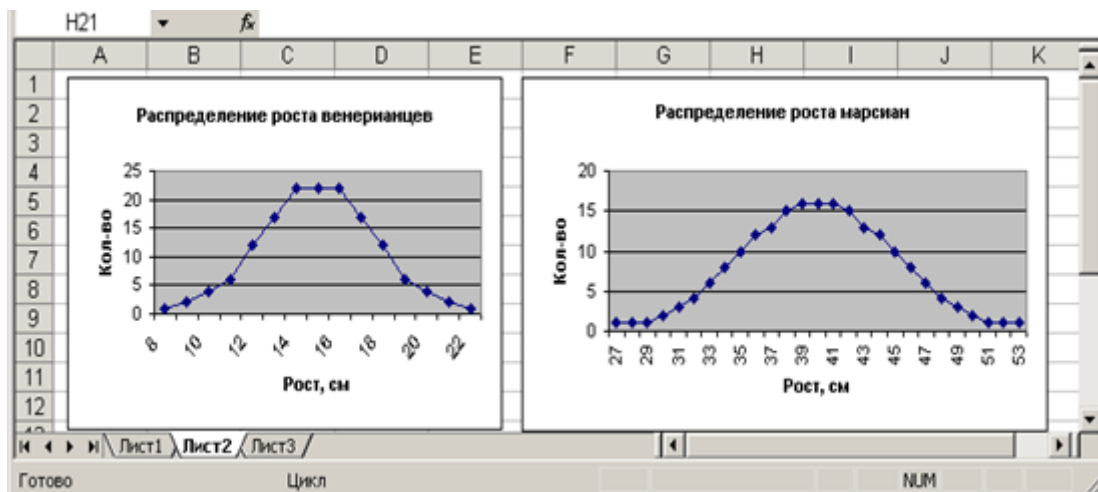
	A	B	C	D	E	F	G
1		<b>МАРС</b>				<b>ВЕНЕРА</b>	
2	рост, X	кол-во, n			рост, X	кол-во, n	
3	27	1			8	1	
4	28	1			9	2	
5	29	1			10	4	
6	30	2			11	6	
7	31	3			12	12	
8	32	4			13	17	
9	33	6			14	22	
10	34	8			15	22	
11	35	10			16	22	
12	36	12			17	17	
13	37	13			18	12	
14	38	15			19	6	
15	39	16			20	4	
16	40	16			21	2	
17	41	16			22	1	
18	42	15					
19	43	13					
20	44	12					
21	45	10					
22	46	8					
23	47	6					
24	48	4					
25	49	3					
26	50	2					
27	51	1					
28	52	1					
29	53	1					

Используя встроенные функции Excel (см. разд.3, зан.2), вычислим среднее по совокупности и стандартное отклонение, как показано на следующем рисунке:

	A	B	C	D	E	F	G
1		<b>МАРС</b>				<b>ВЕНЕРА</b>	
2	рост, X	кол-во, n	$(X-M)^2 \cdot n$		рост, X	кол-во, n	$(X-M)^2 \cdot n$
3	27	1	169		8	1	49
4	28	1	144		9	2	72
5	29	1	121		10	4	100
6	30	2	200		11	6	96
7	31	3	243		12	12	108
8	32	4	256		13	17	68
9	33	6	294		14	22	22
10	34	8	288		15	22	0
11	35	10	250		16	22	22
12	36	12	192		17	17	68
13	37	13	117		18	12	108
14	38	15	60		19	6	96
15	39	16	16		20	4	100
16	40	16	0		21	2	72
17	41	16	16		22	1	49
18	42	15	60			150	1030
19	43	13	117				
20	44	12	192				
21	45	10	250		M=	15	
22	46	8	288		σ=	2,6204325	
23	47	6	294				
24	48	4	256				
25	49	3	243				
26	50	2	200				
27	51	1	121				
28	52	1	144				
29	53	1	169				
30		200	4700				
32	M=	40					
33	σ=	4,84767986					

Получим средний рост марсиан 40 см, а венерианцев – 15 см. Стандартное отклонение роста у марсиан составляет 4,8 см, у венерианцев – 2,6 см.

Для наглядности картины распределения роста у марсиан и венерианцев воспользуемся графическим изображением в виде графика. С помощью мастера диаграмм (см. разд.3, зан.4) построим графики распределения роста у марсиан и венерианцев:



На графике «Распределение роста марсиан» мы видим, что Марсиан среднего роста больше всего, высокорослых столько же, сколько коротышек (распределение симметрично). А на графике «Распределение роста венерианцев» - венерианцы ниже марсиан, разброс значений меньше. Однако по форме распределения венерианцы и марсиане схожи друг с другом.

Составим таблицу, которая сжато представляет то, что мы узнали о марсианах и венерианцах. Из нее можно узнать об объеме совокупности, о среднем росте, и о том, насколько велик разброс относительно среднего.

Параметры распределения марсиан и венерианцев по росту.

	Объем совокупности $N$	Среднее по совокупности $M$ , см	Стандартное отклонение $\sigma$ , см
<b>Марсиане</b>	200	40	4,8
<b>Венерианцы</b>	150	15	2,6

Из рассмотренного примера мы видим, что данное распределение является нормальным и полностью определяется средним по совокупности  $M$  и стандартным отклонением  $\sigma$ , поэтому значения в вышеприведенной таблице – это не просто удачное представление данных, но также и полное их описание.

В научно-исследовательской работе приходится использовать все формы анализа. Графический анализ является первоначальной формой, которая может подсказать исследователю, какие методы работы нужно применять для получения более точной информации об изучаемых явлениях.

### 3. Цель деятельности аспирантов на занятии:

**Аспирант должен знать:**

6. Назначение пакета анализа.
7. Способы проведения графического анализа данных.
8. Закон Гаусса.

**Аспирант должен уметь:**

1. Иметь навыки работы с пакетом анализа в программе MS EXCEL.
2. Уметь проводить графический анализ данных в программе MS EXCEL.

### Содержание обучения:

#### Теоретическая часть:

1. Работа с пакетом анализа.
2. Назначение нормального закона распределения случайной величины (закон Гаусса).
3. Классификация встроенных функций

#### Практическая часть:

Пример выполнения задания.

**Цель работы:** Изучение статистических методов обработки опытных данных подчиняющихся нормальному закону распределения случайных величин.

#### Необходимо:

1. В Excel набрать заданный ряд статистических данных.
2. Расположить значения признака в ранговом порядке.
3. Вычислить среднее по совокупности и стандартное отклонение с помощью встроенных статистических функций.
4. Получить сгруппированный вариационный ряд.
5. Вычислить значения функции  $F(x)$  для каждого значения сгруппированного вариационного ряда по формуле (5), используя значения вычисленные п.2.
6. По полученным значениям функции  $F(x)$  построить график распределения заданного признака. С помощью полученного графика проанализировать динамику распределения.

Данные: Результаты измерения температуры ( $^{\circ}\text{C}$ ) у 14 новорожденных: 36,7; 37,1; 37,0; 37,2; 37,2; 36,8; 36,9; 36,7; 36,5; 37,1; 36,8; 36,8; 36,9; 37,2.

1. Набрать статистические данные и расположить их в ранговом порядке:

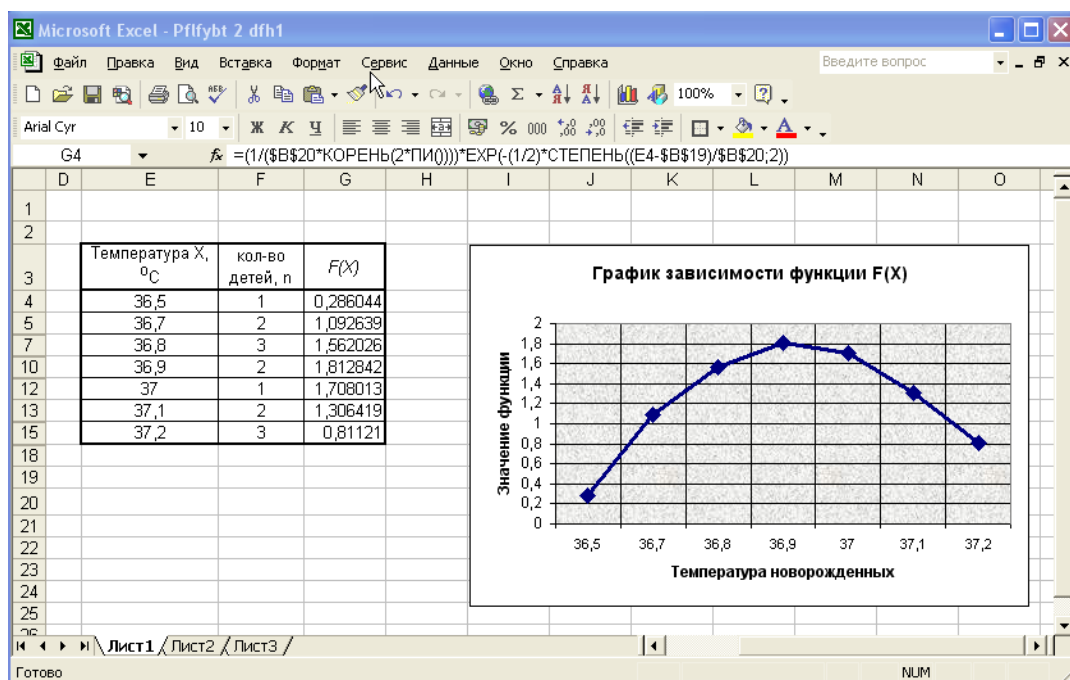
	А	В	С
1	<b>Задание 2 Вариант 1</b>		
2	<b>Исходные данные:</b>		
3	№	Температура, $^{\circ}\text{C}$	
4	1	36,7	
5	2	37,1	
6	3	37	
7	4	37,2	
8	5	37,2	
9	6	36,8	
10	7	36,9	
11	8	36,7	
12	9	36,5	
13	10	37,1	
14	11	36,8	
15	12	36,8	
16	13	36,9	
17	14	37,2	

	А	В	С
1	<b>Задание 2 Вариант 1</b>		
2	<b>Исходные данные:</b>		
3	№	Температура, $^{\circ}\text{C}$	
4	9	36,5	
5	1	36,7	
6	8	36,7	
7	6	36,8	
8	11	36,8	
9	12	36,8	
10	7	36,9	
11	13	36,9	
12	3	37	
13	2	37,1	
14	10	37,1	
15	4	37,2	
16	5	37,2	
17	14	37,2	

2. С помощью встроенных статистических функций Excel вычислить среднее по совокупности и стандартное отклонение:

	A	B	C
1	<b>Задание 2 Вариант 1</b>		
2	<b>Исходные данные:</b>		
3	№	Температура X, °C	
4	9	36,5	
5	1	36,7	
6	8	36,7	
7	6	36,8	
8	11	36,8	
9	12	36,8	
10	7	36,9	
11	13	36,9	
12	3	37	
13	2	37,1	
14	10	37,1	
15	4	37,2	
16	5	37,2	
17	14	37,2	
18			
19	<b>M = 36,9</b>		
20	<b><math>\sigma = 0,22</math></b>		
21			

3. Построить сгруппированный вариационный ряд и вычислить для каждого значения признака (температуры тела новорожденных) значение функции Гаусса, используя формулу (5). По полученным значениям функции Гаусса построить график распределения температуры тела новорожденных:



4. **Вывод:** Данное распределение признака (температуры тела новорожденных) является нормальным, все значения распределены симметрично относительно среднего на одно и два стандартных отклонения. Можно сказать, что новорожденных с низкой и высокой температурой намного меньше, чем с нормальной, и в основном у новорожденных нормальная температура.

Порядок выполнения работы:

1. Изучение теоретического материала.
  2. Выполнение вариантов заданий с помощью рассмотренных инструментов, средств, приемов и технологий
  3. Составление отчета о проделанной работе. Отчет должен содержать следующие разделы:
    - наименование работы;
    - цель работы;
    - пошаговое последовательное описание процесса выполнения варианта задания по видам выполняемых действий.
  4. Результат выполнения варианта задания должен быть сохранен под именем ФИО\_Работа№\_Вариант№ (например, «ИвановНН\_Работа1\_Вариант1.xls») на жесткий диск в папку «Мои документы\ИТ в медицине» и на дискету – в двух копиях (две копии одной и той же информации в разных папках на дискете).
  5. Представление результатов выполнения работы (отчета и файлов на дискете) для проверки преподавателю.
  6. Защита выполненной работы: ответ на контрольные вопросы к теоретическому материалу занятия и ответ на замечания преподавателя по выполненной работе.
  7. Оценка преподавателем выполненной работы.
- Задания для практической части:

#### Вариант 1

25 - 30 мин.

Найдите среднее и стандартное отклонение для следующих данных:  
частота пульса (число ударов в минуту) у 55 студентов-медиков перед экзаменом: 64, 66, 60, 62, 64, 68, 70, 66, 70, 68, 62, 68, 70, 72, 60, 60, 70, 74, 62, 70, 72, 72, 64, 70, 72, 66, 76, 68, 70, 58, 76, 74, 76, 76, 82, 76, 72, 76, 74, 79, 78, 74, 78, 74, 78, 74, 74, 78, 76, 78, 76, 80, 80, 80, 78, 78.

Можно ли считать, что это – выборка из совокупности с нормальным распределением?

#### Вариант 2

25 - 30 мин.

Найдите среднее и стандартное отклонение для следующих данных:  
длительность лечения в стационаре 45 больных пневмонией (в днях): 25, 11, 12, 13, 24, 23, 23, 24, 21, 22, 21, 23, 22, 21, 14, 14, 22, 20, 20, 15, 15, 16, 20, 20, 16, 16, 20, 17, 17, 19, 19, 19, 18, 18, 18, 18, 19, 19, 17, 17, 18, 18, 19, 26.

Можно ли считать, что это – выборка из совокупности с нормальным распределением?

#### Вариант 3

25 - 30 мин.

Найдите среднее и стандартное отклонение для следующих данных:

частота дыхания (число дыхательных движений в минуту) у 47 мужчин в возрасте от 40 до 45 лет: 12, 14, 13, 15, 16, 16, 16, 19, 19, 20, 20, 20, 19, 13, 15, 12, 15, 13, 15, 12, 17, 12, 17, 16, 17, 13, 16, 17, 18, 14, 15, 16, 18, 14, 15, 14, 17, 18, 14, 18, 20, 17, 18, 19, 20, 21, 22.

Можно ли считать, что это – выборка из совокупности с нормальным распределением?

#### Вариант 4

25 - 30 мин.

Найдите среднее и стандартное отклонение для следующих данных:  
число состоящих на диспансерном учете больных у 33 невропатологов поликлиник крупного города: 85, 87, 90, 91, 89, 91, 90, 93, 94, 90, 93, 88, 98, 92, 94, 88, 96, 90, 92, 95, 87, 90, 91, 86, 92, 89, 94, 89, 99, 100, 82, 93, 88.

Можно ли считать, что это – выборка из совокупности с нормальным распределением?

#### Вариант 5

25 - 30 мин.

По данным Всемирной организации здравоохранения (ВОЗ) на конец 1994 года в мире зарегистрировано 17 миллионов ВИЧ-инфицированных. Причем 66% из них находится в Африке на территориях южнее Сахары (11.2 миллиона человек). Ниже в табл. 1 приводятся некоторые конкретные количественные показатели распространенности СПИДа на нашей планете по годам.

Таблица 1.

#### НОВЫЕ СЛУЧАИ СПИДа

ГОД	Африка	Америка	Азия	Европа	Океания	ВСЕГО
1979	0	2	0	0	0	2
1980	0	185	1	17	0	203
1981	0	322	1	20	0	343
1982	2	1156	1	80	91	1330
1983	17	3352	8	295	6	3678
1984	187	6680	8	570	76	7521
1985	521	12682	27	1475	142	14847
1986	5438	21322	86	2395	252	29493
1987	16854	34562	150	9640	324	61530
1988	28212	47697	176	10811	598	87494
1989	41295	56202	288	14355	699	112839
1990	54528	65041	478	17311	770	138128
1991	72756	78579	838	18937	897	172007
1992	73631	99881	2039	20697	866	197114
1993	67124	100731	7368	22053	879	198155
1994	65684	83475	11707	23541	906	185313
1995	16486	47793	5454	11906	174	81813
<b>ВСЕГО</b>	<b>442735</b>	<b>659662</b>	<b>28630</b>	<b>154103</b>	<b>6680</b>	<b>1291810</b>

Произведите статистический анализ данных, представленных в таблице, в отдельности для каждого государства, используя пакет анализа программы Microsoft Excel.

Построить круговые и столбчатые диаграммы, отображающие изменение уровня заболеваемости СПИДом в каждом государстве по годам.

**4. Перечень вопросов для проверки исходного уровня знаний:**

1. Понятие случайной величины. Виды случайных величин.
2. Статистическое исследование.
3. Вычисление статистических величин.

**5. Перечень вопросов для проверки конечного уровня знаний:**

1. Какие цели преследуются при построении графиков?
2. Что принимается во внимание при выборе графического изображения?
3. На какие группы условно можно разделить графики?
4. Какое распределение называется нормальным?
5. Какие средние величины полностью описывают нормальное распределение.

**6. Хронокарта учебного занятия:**

5. Организационный момент – 10 мин.
6. Разбор темы – 40 мин.
7. Текущий контроль (тестирование, практическая работа)- 90 мин.
8. Подведение итогов занятия – 10 мин.

**7. Самостоятельная работа аспиранта.**

Изучите распределение случайных величин, подчиняющихся ненормальному закону с использованием MS Excel 2007.

**8. Перечень учебной литературы к занятию:**

1. Есауленко И.Э., Семенов С.Н. Основы практической информатики в медицине; Воронеж, 2005.



## **ТЕМА 4: Статистические гипотезы. Компьютеры в медико-биологической статистике»**

### **1. Научно-методическое обоснование темы:**

В настоящее время все больше внимания уделяется статистической обработке данных медико-биологических экспериментов, цель которых состоит в выявлении влияний внешних факторов на биологические объекты, механизм которых неизвестен.

Применяя методы статистики, медик-исследователь по-всегда в полной мере осознает их значения в медицинском эксперименте. Статистика иногда представляется ему в виде набора правил и формул, освобождающих исследователя от трудностей и сомнений при решении какой-либо сложной медицинской проблемы.

Роль математической статистики в медицине далеко не ограничиваясь расчетами по определенным формулам при обработке материалов исследования. Наиболее существенную и важную для медицины сторону математической статистики составляет логику статистического анализа, критическое отношение к эмпирическим данным, порядок и система научной аргументации при выборе необходимого статистического метода.

Существует ряд факторов, которые приводят к ряду негативных последствий: некорректное определение генеральной совокупности исследуемых объектов; получение недостоверных результатов из-за недостаточных объемов выборки; несоответствие или отсутствие статистически обоснованного плана эксперимента с поставленными задачами, и как следствие невозможность их корректного решения; неумение осуществить полномасштабный статистический анализ полученных данных;

получение неверных выводов, основанных на статистически незначимых результатах проверки тех или иных гипотез, которые не всегда четко сформулированы.

Статистические методы используются в различных областях биологии и медицины. В частности, они применяются для оценки эффективности методов лечения и применения лекарственных препаратов, диагностики заболеваний, при прогнозировании развития той или иной болезни у конкретного больного или патологического процесса для группы больных. Знакомство с статистическими методами необходимо также для понимания и критической оценки сообщений в медицинских и биологических журналах.

Любое исследование начинается с формирования статистической гипотезы.

Оговариваются особенности статистических закономерностей при получении результатов эксперимента, анализируются допустимость распространения отдельных выборок на всю генеральную совокупность.

### **2. Краткая теория:**

Процедура сопоставления высказанного предположения (гипотезы) с выборочными данными называется проверкой гипотез.

Задачи статистической проверки гипотез:

Относительно некоторой генеральной совокупности высказывается та или иная гипотеза  $H_0$ .

Из этой генеральной совокупности извлекается выборка.

Требуется указать правило, при помощи которого можно было бы по выборке решить вопрос о том, следует ли отклонить гипотезу  $H_0$  или принять ее.

Статистическая гипотеза- это предположение о виде распределения или о величинах неизвестных параметров генеральной совокупности, которая может быть проверена на основании выборочных показателей. В медицинской статистике различают два вида гипотез:

$H_0$  -нулевая, гипотеза отсутствия различий, изменений, эффектов воздействия на совокупность;

$H_1$  – альтернативная, гипотеза о наличии различий, изменений, эффектов при воздействии на совокупность.

Эти дихотомические гипотезы наиболее часто составляют суть медицинских и биологических исследований. Редко гипотеза может включать и более двух возможных вариантов решения.

Примеры статистических гипотез:

Генеральная совокупность распределена по нормальному закону Гаусса.

Дисперсии двух нормальных совокупностей равны между собой.

Если исследование правильно спланировано, то результат практически всегда безупречен.

Валидность- главное определяющее серьезности исследования.

Валидность- способность, применяемого метода отражать именно те качества, на выявление которых данный метод и был направлен. И если условия опыта не меняют кардинально, то выбранный метод будет давать идентичный результат и на других совокупностях.

Статистические гипотезы

↓                      ↓  
Параметрические    Непараметрические

Если исследователь серьезно относится к результату своего труда, он до проведения статистической обработки данных и даже до начала проведения эксперимента должен продумать будет ли анализируемая им совокупность данных отвечать требованиям нормальности, соответствовать закону Гаусса. Дело в том, что математическая статистика и теория проверки статистических гипотез ориентированы на специфику нормального закона распределения. И для корректного применения параметрических методов действительно обязательно выполнение ряда условий, которыми начинающие аналитики и статистики пренебрегают. Если планируется ответственная исследовательская работа и столь же серьезная обработка ее данных, то алгоритм действий исследователя должен содержать обязательные ответы на такие вопросы:

- 1) как близки распределения экспериментальных данных к нормальному закону;
- 2) какая шкала измерений наиболее применима в его исследованиях, как минимум это должна быть интервальная шкала;
- 3) каковы ограничения на минимальный и (или) максимальный объем выборки или согласованность объемов нескольких исследуемых выборок.

Когда требования нормальности распределения и интервальности используемой шкалы не выполняются или их трудно осуществить, то стоит использовать непараметрические методы проверки гипотез.

При получении результата работы можно допустить:

- 1) принятие верной нулевой гипотезы;
- 2) отклонения верной нулевой гипотезы;
- 3) принятие ложной нулевой гипотезы;
- 4) отклонение ложной нулевой гипотезы.

Когда первый и четвертый варианты решения правильны, а второй и третий – ошибочны, то возникает риск ошибки первого и второго рода.

Нулевой гипотезой  $H_0$  называется основная гипотеза, которая проверяется.

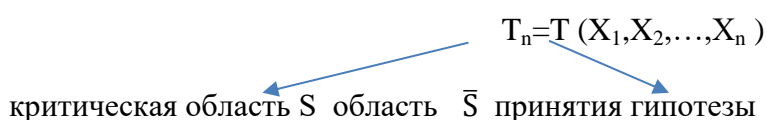
Альтернативной гипотезой  $H_1$ , называется гипотеза, конкурирующая с нулевой, то есть противоречащая ей.

Простой называют гипотезу, содержащую только одно предположение.  $a=a_0$

Сложной называют гипотезу, которая состоит из конечного или бесконечного числа простых гипотез.

Статистическим критерием проверки гипотезы  $H_0$  называется правило, по которому принимается решение принять или отклонить гипотезу  $H_0$ . Проверку гипотез

осуществляют на основании результатов выборки  $X_1, X_2, \dots, X_n$ , из которых формируют функцию выборки  $T_n = T(X_1, X_2, \dots, X_n)$ , называемой статистикой критерия.




---

Возможные ошибки при проверке гипотез

Первого рода

Второго рода

Гипотеза  $H_0$  Отвергается

Принимается

Верна

Ошибка 1-го рода    Нет ошибки

Неверна

Нет ошибки    Ошибка 2-го рода

---

Уровнем значимости критерия ( $\alpha$ ) называется вероятность допустить ошибку 1-го рода.

Вероятность ошибки 2-го рода обозначается через  $\beta$ .

Мощностью критерия называется вероятность недопущения ошибки 2-го рода  $(1 - \beta)$ .

$\alpha = P(\text{отвергнуть } H_0 / H_0 \text{ верна})$  или  $\alpha = P(H_1 / H_0)$

$\beta = P(\text{принять } H_0 / H_0 \text{ неверна})$  или  $\beta = P(H_0 / H_1)$

$1 - \beta = P(\text{принять } H_1 / H_1 \text{ верна})$

Чем больше мощность критерия, тем вероятность ошибки 2-го рода меньше.

Разумное соотношение между  $\alpha$  и  $\beta$  находят, исходя из тяжести последствий каждой из ошибок.

Методика проверки гипотез:

1. Формирование нулевой  $H_0$  и альтернативной  $H_1$  гипотез исходя из выборки

$X_1, X_2, \dots, X_n$ .

2. Подбор статистики критерия  $T_n = T(X_1, X_2, \dots, X_n)$

3. По статистике критерия  $T_n$  и уровню значимости  $\alpha$  определяют критическую точку  $t_{кр}$ , то есть границу, отделяющую область  $\bar{S}$  от  $S$ .

4. Для полученной реализации выборки  $X = (X_1, X_2, \dots, X_n)$  подсчитывают значение критерия, то есть  $T_{набл} = T(X_1, X_2, \dots, X_n) = t$

5. Если  $t \in S$  (например,  $t > t_{кр}$  для правосторонней области  $S$ ), то нулевую гипотезу  $H_0$  отвергают; если же  $t \in \bar{S}$  ( $t < t_{кр}$ ), то нет оснований, чтобы отвергнуть гипотезу  $H_0$ .

В современных медицинских учреждениях применение компьютерных технологий стало обычным делом обработки и анализа данных. Используются статистические пакеты: пакет анализа Excel, SPSS, STATISTICA.

#### 4. Цель деятельности аспирантов на занятии:

**Аспирант должен знать:**

9. Генеральная совокупность. Выборка
10. Статистическая гипотеза. Виды.
11. Статистический критерий проверки гипотезы.
12. Ошибка 1-го и 2-го рода.

**Аспирант должен уметь:**

1. Четко формулировать статистическую гипотезу (нулевую и альтернативную).

### **Содержание обучения:**

#### **Теоретическая часть:**

1. Формирование статистической гипотезы.
2. Использование параметрических и непараметрических методов проверки статистических гипотез.

#### **Практическая часть:**

**Задача №1.** Процент положительных исходов оперативных вмешательств на позвоночнике по поводу остеохондроза люмбального отдела в двух хирургических отделениях.

*Выдвинуть альтернативные гипотезы.*

**Задача №2.** Определить, имеются ли изменения вибрационной чувствительности у подростков, осваивающих массовую рабочую профессию сборщика изделий из мелких деталей, до и после работы.

*Выдвинуть гипотезы.*

#### **5. Перечень вопросов для проверки исходного уровня знаний:**

1. Понятие генеральной совокупности, выборки.
2. Дихотомические гипотезы.
3. Контроль-опыт. Валидность

#### **6. Перечень вопросов для проверки конечного уровня знаний:**

1. Статистическая гипотеза. Виды.
2. Параметрические и непараметрические методы проверки статистических гипотез.
3. Выбор метода проверки гипотезы.
4. Ошибки 1-го и 2-го рода.

#### **7. Хронокарта учебного занятия:**

**9.** Организационный момент – 10 мин.

**10.** Разбор темы – 40 мин.

**11.** Текущий контроль (тестирование, практическая работа) - 90 мин.

**12.** Подведение итогов занятия – 10 мин.

#### **8. Самостоятельная работа аспиранта.**

Статистическая обработка результатов клинических исследований.

#### **9. Перечень учебной литературы к занятию:**

1. Есауленко И.Э., Семенов С.Н. Основы практической информатики в медицине; Воронеж, 2005.
2. Жижин К. С. Медицинская статистика; Ростов н/Д, 2007.

## **ТЕМА 5: Критерии Стьюдента для двух несвязанных выборок, F- критерии Фишера, U-критерии Манна-Уитни, критерии Краскела-Уолиллиса для выявления различий в уровне признака**

### **1. Научно-методическое обоснование темы:**

Задача оценки различий признаков – основа клинико-диагностического и профилактического процессов в медицине. Для выявления таких различий в статистике разработаны высокоэффективные критерии: параметрические (Стьюдента, Фишера и др.) и непараметрические. Параметрические критерии требуют выполнения условия нормальности, что для реальных эмпирических данных часто не выполняется. Поэтому чаще врачи-практики внимание уделяют непараметрическим критериям, которые не предполагают соответствия эмпирических данных какому-либо теоретическому закону распределения.

В случае нормального распределения эмпирических данных параметрические критерии являются более мощными по сравнению с непараметрическими. Поэтому в общем случае исследователь должен сначала выполнить проверку на нормальность распределения и лишь затем, в зависимости от ее результатов, принимать решение о выборе статистического критерия.

Все статистические критерии выявления различий в уровне исследуемого признака (параметрические и непараметрические) можно разделить на две основные группы:

- ✓ для двух выборок;
- ✓ для трех и более выборок.

Наиболее популярным параметрическим критерием для сравнения двух выборок является t- критерий Стьюдента для независимых выборок. Вариант критерия, используемый в SPSS и STATISTICA , предназначенный для сравнения средних величин выборок, ориентирован на проверку гипотезы однородности о том, что выборки извлечены из одной и той же генеральной совокупности.

При этом предполагается, что обе выборки извлечены из генеральных совокупностей, имеющих нормальные распределения. На практике получается, что критерий Стьюдента при больших объемах выборок устойчив к отклонениям от нормальности.

В том же случае, когда выборки взяты из иных совокупностей, истинные значения признаков должны оцениваться с помощью специальных приемов. Исходя из этого критерий Стьюдента требует нормальности распределения выборок.

Не менее популярен другой параметрический критерий – F-критерий Фишера. Этот прием обработки статистической информации используют при проведении дисперсионного анализа при отыскании причинно-следственных связей между анализируемыми признаками.

Из непараметрических критериев для сравнения двух выборок популярен критерий Манна-Уитни. Эта группа методик в медико-биологических исследованиях используется слабо. Данный критерий практически не имеет ограничений на объемы выборок, он позволяет сравнивать выборки разного объема.

### **2. Краткая теория:**

### **t-критерий Стьюдента.**

Для того, чтобы **рассчитать t-критерий Стьюдента (для зависимых и для независимых выборок) в Excel** необходимо сделать следующие шаги:

1. Вносим значения для двух переменных в таблицу (Например, *Переменная 1* и *Переменная 2*).
2. Ставим курсор в пустую ячейку
3. В строке формул выбрать кнопку ***fx*** (*вставить формулу*), или *Формулы-Вставить функцию*.
4. В открывшемся окне «*Мастер функций*» в поле «Категории» выбираем **Полный алфавитный перечень**
5. Затем в поле «*Выберите функцию*» находим функцию **ТТЕСТ**, которая возвращает вероятность, соответствующую критерию Стьюдента.
- 5.1. Нажимаем **Ок**
6. В открывшемся окне «*Аргументы функции*» в поле Массив1 вносим **номера ячеек**, содержащие значения Переменной 1, в поле Массив2 вносим **номера ячеек**, содержащие значения Переменной2.
7. В поле «*Хвосты*» пишем **2** (критерий будет рассчитываться используя **двустороннее распределение**, как и в SPSS); либо **1** (критерий будет рассчитываться используя **одностороннее распределение**).
- Важно!** 8. В поле «Тип» пишем **1** (рассчитывается, если **выборки зависимые**); либо **2** или **3** (если **выборки независимые**).
9. Нажимаем **Ок**
10. Смотрим получившийся результат

#### **ВЫВОД:**

- ✓ Критерий Стьюдента может быть использован для проверки гипотезы о различии средних только для двух групп.
- ✓ Критерий Стьюдента применяется в случае малых выборок, что характерно для медико- биологических экспериментов.
- ✓ Если схема эксперимента предполагает большее число групп, воспользуйтесь дисперсионным анализом.
- ✓ Если критерий Стьюдента был использован для проверки различий между несколькими группами, то истинный уровень значимости можно получить, умножив уровень значимости, на число возможных сравнений.

## Ф-критерий Фишера.

Критерии различия называют непараметрическими, если он не базируется на предположении о типе распределения генеральной совокупности и не использует параметры этой совокупности.

*Применение непараметрических методов целесообразно:*

на этапе разведочного анализа;

при малом числе наблюдений (до 30);

когда нет уверенности в соответствии данных закону нормального распределения.

*Непараметрические критерии представлены основными группами:*

- ✓ критерии различия между группами
- ✓ независимых выборок;
- ✓ критерии различия между группами
- ✓ зависимых выборок.

*Назначение.* Проверка гипотезы о принадлежности двух дисперсий одной генеральной совокупности и следовательно — их равенстве.

*Нулевая гипотеза.*  $S_1^2 = S_2^2$

*Альтернативная гипотеза.* Существуют следующие варианты  $H_A$  в зависимости от которых различаются критические области:

1.  $S_1^2 > S_2^2$ . Наиболее часто используемый вариант  $H_A$ . Критическая область — верхний хвост F-распределения.
2.  $S_1^2 < S_2^2$ . Критическая область — нижний хвост F-распределения. Ввиду частого отсутствия нижнего хвоста, в таблицах критическую область обычно сводят к варианту 1, меняя местами дисперсии.
3. Двухсторонняя  $S_1^2 \neq S_2^2$ . Комбинация первых двух.

*Предпосылки.* Данные независимы и распределены по нормальному закону. Гипотеза о равенстве дисперсий двух нормальных генеральных совокупностей принимается, если отношение большей дисперсии к меньшей меньше критического значения распределения Фишера.

$$F_p = S_1^2 / S_2^2$$

Примечание. При описываемом способе проверки значение  $F_{расч}$  обязательно должно быть больше единицы. Критерий чувствителен к нарушению предположения о нормальности.

Для двухсторонней альтернативы  $S_1^2 \neq S_2^2$  нулевая гипотеза принимается при выполнении условия:

$$F_{1-\alpha/2} < F_{расч} < F_{\alpha/2}$$

## Пример

Комплексным теплометрическим методом определяли теплофизические характеристики (ТФХ) зеленого солода. Для приготовления образцов брали воздушно-сухой (средняя влажность  $W=19\%$ ) и влажный солод четырехсуточного ращения ( $W=45\%$ ) в соответствии новой технологией приготовления карамельного солода. Опыты показали, что теплопроводность  $\lambda$  влажного солода примерно в 2,5

раза больше, чем сухого, а объемная теплоемкость не имеет четкой зависимости от влажности солода. Поэтому с помощью F-критерия проверили возможность обобщить данные по средним значениям без учета влажности

Расчетные данные сведены в таблицу 5.1

Таблица 5.1. Данные к расчету F-критерия

W, %	19					45				
t, °C	30,6	34,2	38,9	44,2	49,1	29,1	36,5	41,5	47,2	54,3
y, МДж/(м³·К)	0,92	1,32	1,31	1,62	1,06	1,28	1,71	1,72	1,53	1,25
$\bar{y}$	1,27	1,31	1,36	1,41	1,46	1,26	1,34	1,39	1,44	1,51
$ y - \bar{y} $	0,35	0,01	0,05	0,21	0,40	0,02	0,37	0,33	0,09	0,26
$(y - \bar{y})^2$	0,12	0,00	0,00	0,04	0,16	0,00	0,13	0,11	0,01	0,06
S²	0,080					0,108				

Большее значение дисперсии получено для W=45%, т.е.  $S_{45}^2 = S_1^2$ ,  $S_{19}^2 = S_2^2$ , и  $F_p = S_1^2/S_2^2 = 1,35$ . Из таблицы 5.2 для степени свободы  $f_1 = N_1 - 1 = 5$   $f_2 = N_2 - 1 = 4$  при  $\gamma = 0,95$  определяем  $F_{кр} = 6,2$ . Нуль гипотеза сформулированная как «В диапазоне влажности зеленого солода от 19 до 45% ее влиянием на объемную теплоемкость можно пренебречь» или « $S_{45}^2 = S_{19}^2$ » с доверительной вероятностью 95% подтвердилась, поскольку  $F_p < F_{кр}$ .

### Пример проверки гипотезы о принадлежности двух дисперсий одной генеральной совокупности по критерию Фишера с помощью Excel

Приведены данные по двум независимым выборкам (табл. 5.2) степени водопоглощения зерна пшеницы. Было проведено исследование воздействия магнитными полями низкой частоты.

Таблица 5.2. Результаты исследований

Номер опыта	Номер выборки	
	0,027	0,075
	0,036	0,4
	0,1	0,08
	0,12	0,105
	0,32	0,075
	0,45	0,12



	0,049	0,06
	0,105	0,075

Прежде, чем мы будем проверять гипотезу о равенстве средних этих выборок, необходимо проверить гипотезу о равенстве дисперсий, чтобы знать какой из критериев выбрать для ее проверки.

На рис. 5.1 приведен пример проверки гипотезы о принадлежности двух дисперсий одной генеральной совокупности по критерию Фишера используя программный продукт Microsoft Excel.

F15		=FPACП(F13;7;7)				
	A	B	C	D	E	F
1		№ опыта	Номер выборки			
2			1	2		
3		1	0,027	0,075		
4		2	0,036	0,4		
5		3	0,1	0,08		
6		4	0,12	0,105		
7		5	0,32	0,075		
8		6	0,45	0,12		
9		7	0,049	0,06		
10		8	0,105	0,075		
11	Дисперсии		0,02323	0,01283		
12						
13	Расчетное критерия Фишера					1,81144
14	Критическое значение для критерия Фишера					3,78705
15	Расчетный уровень значимости					0,22566
16						
17						

Рисунок 5.1 Пример проверки принадлежности двух дисперсий одной генеральной совокупности по критерию Фишера

Исходные данные размещены в ячейках, находящихся на пересечении столбцов C и D со строками 3-10. Выполним следующие действия.

1. Определим, можно ли считать закон распределения первой и второй выборок нормальным (столбцы C и D соответственно). Если нет (хотя бы для одной выборки), то необходимо использовать непараметрический критерий, если да – продолжаем.
2. Рассчитаем дисперсии для первого и второго столбца. Для этого в ячейках C11 и D11 поместим функции =ДИСП(C3:C10) и =ДИСП(D3:D10) соответственно. Результатом работы этих функций является рассчитанное значение дисперсии для каждого столбца соответственно.
3. Находим расчетное значение для критерия Фишера. Для этого нужно большую дисперсию разделить на меньшую. В ячейку F13 помещаем формулу =C11/D11, которая и выполняет эту операцию.

4. Определяем, можно ли принять гипотезу о равенстве дисперсий. Существует два способа, которые представлены в примере. По первому способу, задавшись уровнем значимости, например 0,05, вычисляют критическое значение распределения Фишера для этого значения и соответствующего числа степеней свободы. В ячейку F14 вводится функция =ФРАСПОВР(0,05;7;7) (где 0,05 - заданный уровень значимости; 7 — число степеней свободы числителя, а 7 (второе) — число степеней свободы знаменателя). Число степеней свободы равно числу экспериментов минус единица. Результат — 3,787051. Поскольку это значение больше расчетного 1,81144, мы должны принять нулевую гипотезу о равенстве дисперсий.

о второму варианту рассчитывают для полученного расчетного значения критерия Фишера соответствующую вероятность. Для этого в ячейку F15 вводится функция =ФРАСП(F13;7;7). Поскольку полученное значение 0,22566 больше, чем 0,05, то принимается гипотеза о равенстве дисперсий.

Это может быть выполнено специальной функцией. Выберите на ленте вкладку *Данные, Анализ данных*. Появится окно следующего вида (рис. 5.2).

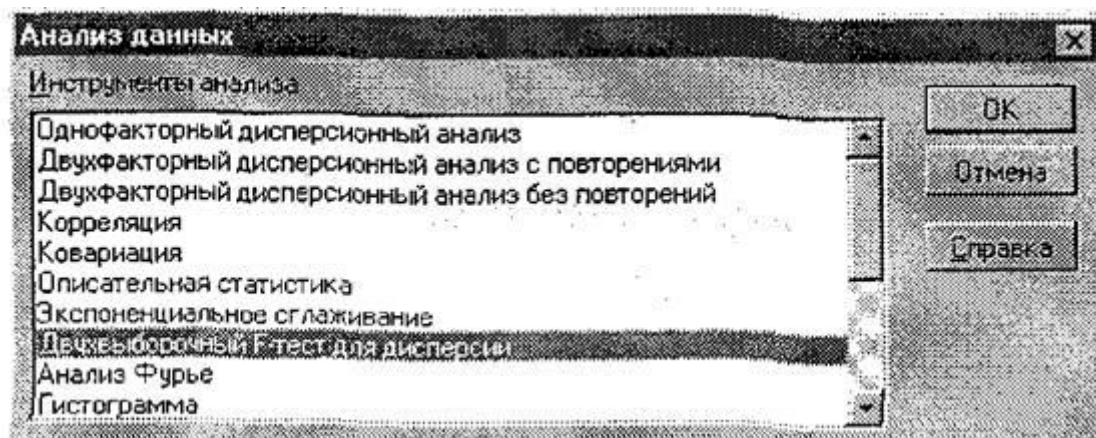


Рисунок 5.2 Окно выбора метода обработки

В этом окне выбираете «*Двухвыборочный F-тест для дисперсий*». В результате появится окно вида, показанного на рис. 5.3. Здесь задаются интервалы (номера ячеек) первой и второй переменной, уровень значимости (альфа) и место, где будет находиться результат.

Задавайте все необходимые параметры и нажимайте ОК. Результат работы приведен на рис. 5.4

Следует отметить, что функция проверяет односторонний критерий и делает это правильно. Для случая, когда критериальное значение больше 1, вычисляется верхнее критическое значение.

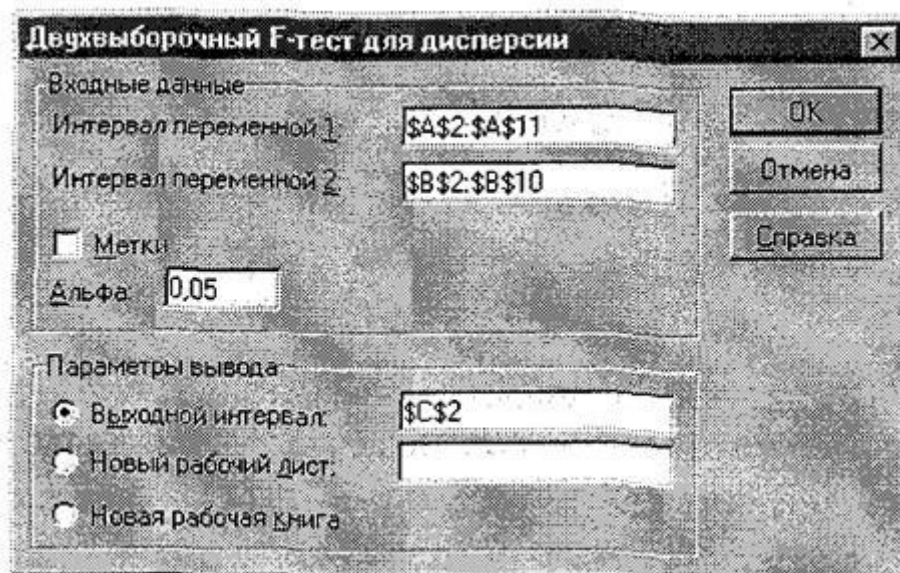


Рисунок 5.3 Окно задания параметров

Когда критериальное значение меньше 1, то вычисляется нижнее критическое.

Напоминаем, что гипотеза о равенстве дисперсий отвергается, если критериальное значение больше верхнего критического или меньше нижнего.

	A	B	C	D	E	F	G	H	I	J
1	Группа 1	Группа 3								
2	1,85	2,27	Двухвыборочный F-тест для дисперсии							
3	1,87	2,09								
4	1,87	2,09		Переменная 1	Переменная 2					
5	2,3	2,41	Среднее	2,001	2,162222222					
6	2,52	2,31	Дисперсия	0,083254444	0,019469444					
7	1,89	2,17	Наблюдения	10	9					
8	2,37	2	df	9	8					
9	1,7	2,1	F	4,276159224						
10	1,7	2,02	P(F<=f) одностороннее	0,026382941						
11	1,94		F критическое одностороннее	3,388123559						
12										
13										
14										

Рисунок 5.4 Проверка равенства дисперсий

Различия между независимыми группами:

- ✓ U критерий Манна-Уитни
- ✓ двухвыборочный критерий Колмогорова – Смирнова.

## U-КРИТЕРИЙ МАННА-УИТНИ

U-критерий Манна-Уитни – непараметрический статистический критерий, используемый для сравнения двух независимых выборок по уровню какого-либо признака, измеренного количественно. Метод основан на определении того, достаточно ли мала зона перекрещивающихся значений между двумя вариационными рядами (ранжированным рядом значений параметра в первой выборке и таким же во второй выборке). Чем меньше значение критерия, тем вероятнее, что различия между значениями параметра в выборках достоверны.

U-критерий Манна-Уитни используется для оценки различий между двумя независимыми выборками по уровню какого-либо количественного признака.

U-критерий Манна-Уитни является непараметрическим критерием, поэтому, в отличие от t-критерия Стьюдента, не требует наличия нормального распределения сравниваемых совокупностей.

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.

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

Аналогом U-критерия Манна-Уитни для сравнения более двух групп является Критерий Краскала-Уоллиса.

Как рассчитать U-критерий Манна-Уитни?

Сначала из обеих сравниваемых выборок составляется единый ранжированный ряд, путем расставления единиц наблюдения по степени возрастания признака и присвоения меньшему значению меньшего ранга. В случае равных значений признака у нескольких единиц каждой из них присваивается среднее арифметическое последовательных значений рангов.

Например, две единицы, занимающие в едином ранжированном ряду 2 и 3 место (ранг), имеют одинаковые значения. Следовательно, каждой из них присваивается ранг равный  $(2 + 3) / 2 = 2,5$ .

В составленном едином ранжированном ряду общее количество рангов получится равным:

$$N = n_1 + n_2$$

где  $n_1$  - количество элементов в первой выборке, а  $n_2$  - количество элементов во второй выборке.

Далее вновь разделяем единый ранжированный ряд на два, состоящие соответственно из единиц первой и второй выборок, запоминая при этом значения рангов для каждой единицы. Подсчитываем отдельно сумму рангов, пришедшихся на долю элементов первой выборки, и отдельно - на долю элементов второй выборки. Определяем большую из двух ранговых сумм ( $T_x$ ) соответствующую выборке с  $n_x$  элементами.

Наконец, находим значение U-критерия Манна-Уитни по формуле:

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

Как интерпретировать значение U-критерия Манна-Уитни?

Полученное значение U-критерия сравниваем по таблице для избранного уровня статистической значимости ( $p=0,05$  или  $p=0,01$ ) с критическим значением U при заданной численности сопоставляемых выборок:

Если полученное значение U меньше табличного или равно ему, то признается статистическая значимость различий между уровнями признака в рассматриваемых выборках (принимается альтернативная гипотеза). Достоверность различий тем выше, чем меньше значение U.

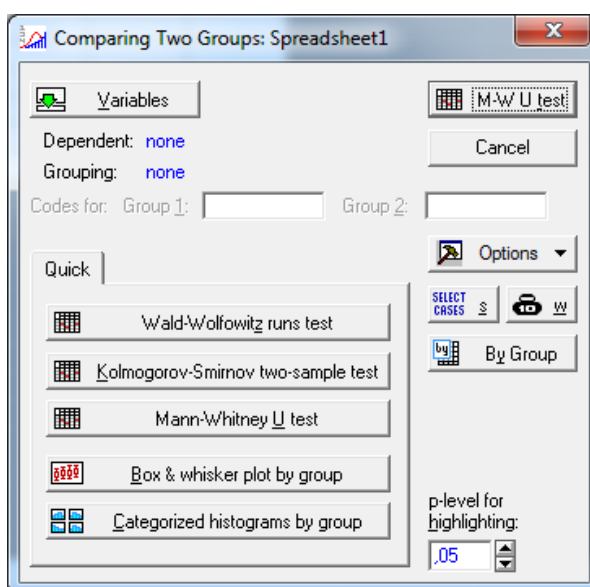
Если же полученное значение U больше табличного, принимается нулевая гипотеза.

### Реализация в STATISTICA

Для сравнения двух независимых выборок с помощью U-критерия необходимо использовать следующую последовательность команд:

Statistics (Статистики) – Nonparametrics (Непараметрические) – Comparing two independent samples (groups) (Сравнение двух независимых выборок)

В результате чего, откроется диалоговое окно (рис.12), в котором необходимо указать зависимую и группирующую переменную, а затем, в поле Codes for (Коды для) указать



коды групп.

Рис.12. Диалоговое окно Comparing Two Groups

После нажатия кнопки Mann – Whitey U test на экран будет выведена таблица результатов теста.

Пример сравнения индекса массы тела до программы похудения для мужчин и женщин представлен на рис.13.

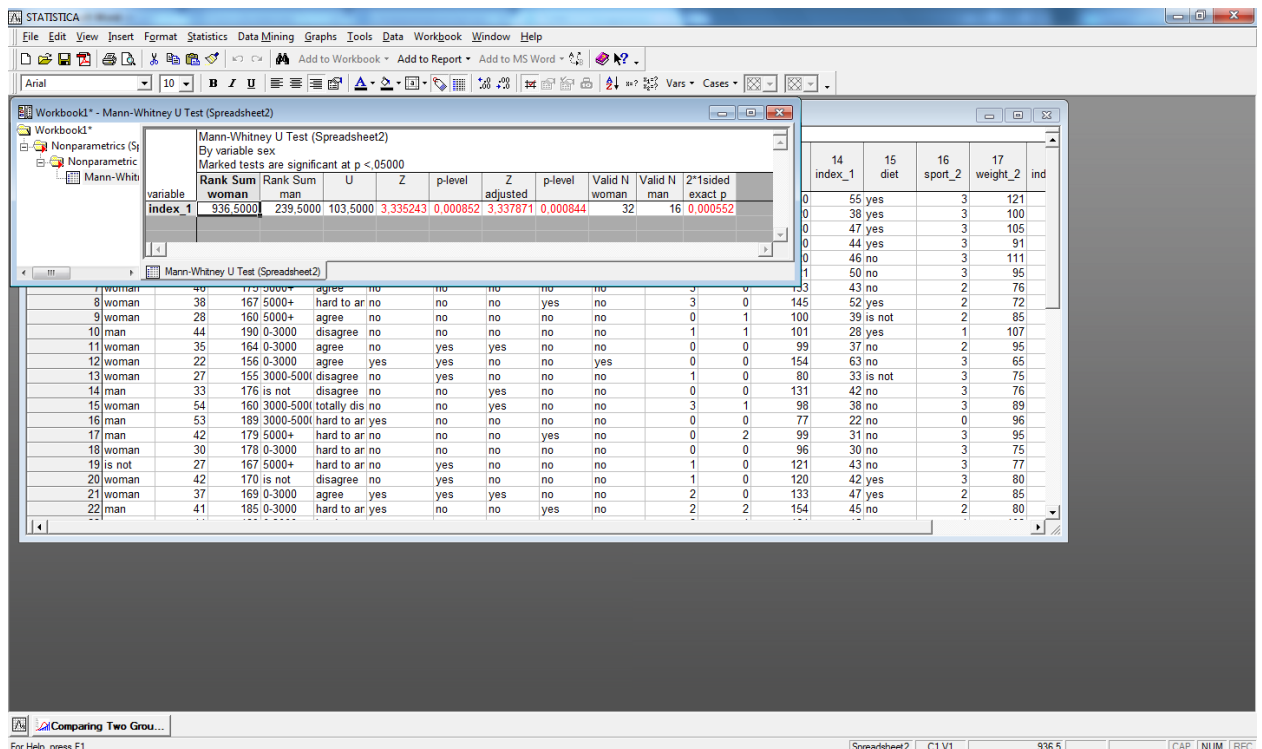


Рис.13.Пример расчета U-критерия Манна-Уитни

Тут:

Rank Sum Group 1 и Rank Sum Group 2 –суммы рангов (по возрастанию) для первой и второй групп соответственно, по которым можно определить в какой из групп выше уровень признака;

U –значение статистики U-критерия Манна – Уитни;

(ближайший) p-level – уровень значимости критерия в данном случае (односторонний критерий);

и т.д., в том числе и количество наблюдений в первой и второй группах (выборках): Valid N Group 1 и Valid N Group 2.

Результаты теста говорят о значимом различии между индексами массы тела для мужчин и женщин ( $p < 0.05$ ).

### Реализация в SPSS

Для сравнения двух независимых выборок с помощью U-критерия необходимо использовать следующую последовательность команд:

Analyze (Анализ) –Nonparametric Tests(Непараметрические критерии) –Legacy Dialogs(Устаревшие диалоги) - 2Independent Samples(Две независимые группы)

В результате на экране появится диалоговое окно (рис.1), в котором необходимо задать в поле Test Variable List переменные, которые подлежат проверке, а в поле Grouping Variables– группирующую переменную.

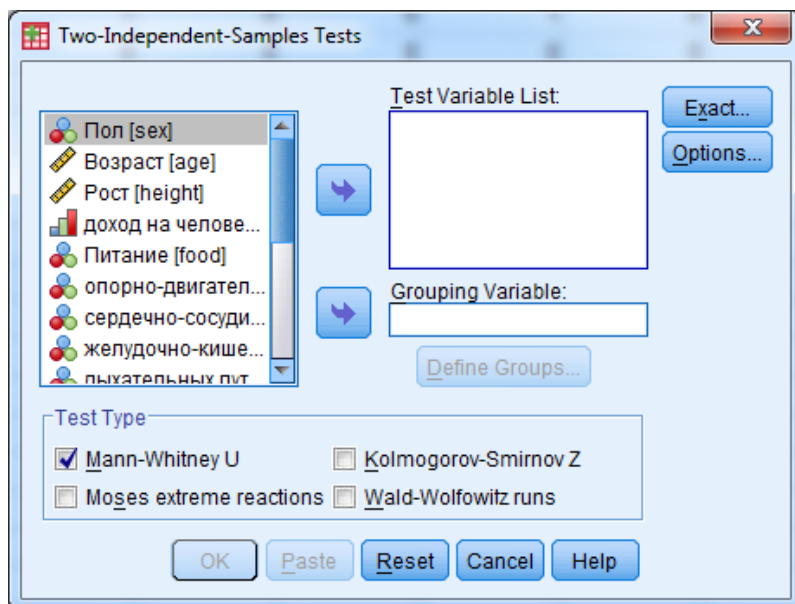


Рис.1. Диалоговое окно Two Independent Samples Tests

Чтобы рассчитать критерий Манна – Уитни для двух независимых выборок необходимо установить флажок в поле Mann-Whitney U. После щелчка на кнопке ОК на экран будет выведен результат теста.

На рис.2. представлено сравнение средних значений индекса массы тела до программы похудения для мужчин и женщин

SPSS Output window showing the results of the Mann-Whitney U test. The output includes the NPar Tests section, the Mann-Whitney test results, and a table of ranks.

**NPar Tests**

[DataSet1] D:\STAT\метод\1b2.sav

**Mann-Whitney**

Ranks			
Пол	N	Mean Rank	Sum of Ranks
Индекс массы тела до программы похудения	32	29,28	937,00
man	16	14,94	239,00
Total	48		

**Test Statistics<sup>a</sup>**

	Индекс массы тела до программы похудения
Mann-Whitney U	103,000
Wilcoxon W	239,000
Z	-3,346
Asymp. Sig. (2-tailed)	,001

a. Grouping Variable: Пол

Рис.2. Пример расчета U-критерия Манна-Уитни.

Результаты теста говорят о значимом различии между индексами массы тела для мужчин и женщин ( $p=0,001 < 0,05$ ).

### 3. Цель деятельности аспирантов на занятии:

#### Аспирант должен знать:

13. Генеральная совокупность. Выборка
14. Статистическая гипотеза. Виды.
15. Ранжированный ряд.
16. Зависимая и независимая выборки.
17. Критерии проверки гипотез.

#### Аспирант должен уметь:

1. Рассчитывать критерии Стьюдента, Фишера, Манна-Уитни.

#### Содержание обучения:

##### Теоретическая часть:

1. Использование параметрических и непараметрических методов проверки статистических гипотез.
2. Расчет при помощи пакета анализа EXCEL, STATISTICA, SPSS.

##### Практическая часть:

Задача №1. Рассчитать критерий Стьюдента для выявления различий в уровне исследуемого признака для несвязанных выборок.

Процент положительных исходов оперативного вмешательства на позвоночнике по поводу остеохондроза в двух хирургических отделениях.

№	Отделение №1	Отделение №2
1	40	44,2
2	35,8	37
3	41,2	38,8
4	44	44,2
5	42,8	43,4
6	47,6	49,6
7	42,8	43,2
8	39,6	40,6
9	36,8	37,4
10	45	46,2



Задача №2. Применение критерия Краскела-Уоллиса для выявления различий в уровне признака.

*Условие:* установить степень достоверности отличий числа допущенных ошибок по корректурному тесту Анфимова в трех исследуемых группах студентов перед началом лонгитудинального эксперимента по изучению умственной работоспособности.

<b>№</b>	<b>I</b>	<b>II</b>	<b>III</b>
<b>1</b>	3	4	4
<b>2</b>	4	4	5
<b>3</b>	5	5	6
<b>4</b>	2	3	5
<b>5</b>	7	5	6
<b>6</b>	8	6	3
<b>7</b>	3	2	3
<b>8</b>	3	4	5
<b>9</b>	4	4	6
<b>10</b>	4	4	7
<b>11</b>	4	5	4
<b>12</b>	5	6	3
<b>13</b>	3	3	2
<b>14</b>	6	3	0
<b>15</b>	0	5	5
<b>16</b>	0	4	7
<b>17</b>	4	5	8
<b>18</b>	3	3	9
<b>19</b>	2	3	2
<b>20</b>	4	4	3
<b>21</b>	4	2	4

**4. Перечень вопросов для проверки исходного уровня знаний:**

1. Параметрические критерии различий в уровне признака.
2. Непараметрические критерии в уровне признака.
3. Две группы выявления различий исследуемого признака.

**5. Перечень вопросов для проверки конечного уровня знаний:**

1. Критерий Стьюдента. Использование данного критерия.
2. Критерий Фишера. Использование данного критерия.
3. Критерий Манна-Уитни. Использование данного критерия.
4. Критерий Краскела-Уоллиса. Использование данного критерия.

**6. Хронокарта учебного занятия:**

13. Организационный момент – 10 мин.
14. Разбор темы – 40 мин.
15. Текущий контроль (тестирование, практическая работа) - 90 мин.
16. Подведение итогов занятия – 10 мин.

**7. Самостоятельная работа аспиранта.**

Проверка нормальности выборочных распределений согласно критерию Шапиро-Уилкса.

**8. Перечень учебной литературы к занятию:**

1. Есауленко И.Э., Семенов С.Н. Основы практической информатики в медицине; Воронеж, 2005.
2. Жижин К. С. Медицинская статистика; Ростов н/Д, 2007.

## **ТЕМА 6: Параметрические коэффициенты корреляции. Применение критериев Стьюдента, Вилкоксона, Фридмана.**

### **1. Научно-методическое обоснование темы:**

Задача оценки различий признаков – основа клинико-диагностического и профилактического процессов в медицине. Для выявления таких различий в статистике разработаны высокоэффективные критерии: параметрические (Стьюдента, Фишера и др.) и непараметрические (Т-критерий Вилкоксона). Параметрические критерии требуют выполнения условия нормальности, что для реальных эмпирических данных часто не выполняется. Поэтому чаще врачи-практики внимание уделяют непараметрическим критериям, которые не предполагают соответствия эмпирических данных какому-либо теоретическому закону распределения.

В случае нормального распределения эмпирических данных параметрические критерии являются более мощными по сравнению с непараметрическими. Поэтому в общем случае исследователь должен сначала выполнить проверку на нормальность распределения и лишь затем, в зависимости от ее результатов, принимать решение о выборе статистического критерия.

Оценка достоверности сдвига в изучаемых совокупностях для связанных выборок, понятие сдвига в исследуемом признаке, а также оценка их разновидностей для медико-биологических исследований – все это в подавляющем большинстве случаев основные определяющие при установлении научной истины.

### **2. Краткая теория:**

Заключение о случайности или неслучайности различий между выборочными совокупностями при использовании параметрических критериев осуществляется на основании сравнения параметров распределений, т.е. сводных числовых характеристик. Каждый из параметров компактно, в виде одного единственного числа, отражает некие характерные свойства распределения данной случайной величины. Они являются количественными мерами этих свойств. На практике, как правило, рассматривают лишь два параметра – среднее значение, являющееся «мерой положения математического центра» полученного вариационного ряда, и дисперсию, но чаще всего корень из нее – стандартное отклонение, являющиеся мерой вариации. Для этих параметров разработаны два наиболее популярных параметрических критерия: критерий Стьюдента и критерий Фишера.

Критерий Стьюдента (t-критерий) – критерий, основанный на сравнении средних значений выборок. Критерий Стьюдента является наиболее известным. С одной стороны, анализ средних значений сравнительно прост для вычислений. С другой стороны, средние величины наиболее наглядны и понятны.

Наиболее часто t-критерий используется в двух вариантах. В первом случае его применяют для проверки гипотезы о равенстве генеральных средних двух независимых, несвязанных выборок (так называемый двухвыборочный t-критерий). В этом случае есть контрольная группа и опытная группа, состоящая из разных пациентов, количество которых в группах может быть различно. Во втором же случае используется так называемый парный t-критерий, когда одна и та же группа объектов порождает числовой материал для проверки гипотез о средних. Поэтому эти выборки называют зависимыми, связанными. Например, измеряется содержание лейкоцитов у здоровых животных, а затем у тех же самых животных после облучения определенной дозой излучения. В обоих случаях должно выполняться требование нормальности распределения исследуемого признака в каждой из сравниваемых групп.

Для того, чтобы определить, является ли нормальным исследуемое распределение, используются критерии Шапиро-Уилка и Колмогорова-Смирнова.

### *Критерий Стьюдента*

Под выборочным методом в статистике понимается такой метод наблюдения, при котором для отыскания типичных черт характеристик какой-либо совокупности изучаются не все единицы этой совокупности, а лишь часть их. Как бы тщательно ни производилась выборка, какой репрезентативной ни была бы выборочная совокупность (отобранная часть наблюдений), она неизбежно будет отличаться от всей генеральной (общей) совокупности. Таким образом, полного тождества достичь не удастся, и некоторая неточность встречается неизбежно. Однако имеются методы установления степени различий числовых характеристик обеих совокупностей и пределов возможных колебаний показателей при данном числе наблюдений. Число наблюдений играет значительную роль - чем больше число наблюдений, тем точнее отображается генеральная совокупность и тем меньше размеры ошибки.

Так называемые средние ошибки являются мерой точности и достоверности любых статистических величин. ***Под достоверностью статистических показателей*** (синонимы: существенность, значимость, надежность) ***понимают доказательность, то есть право на обобщение явления, правомерность распространения выводов и на другие аналогичные явления. Или - степень их соответствия отображаемой ими действительности.*** Достоверными результатами считаются те, которые не искажают и правильно отражают объективную реальность.

Оценить достоверность результатов исследования означает определить, с какой вероятностью возможно перенести результаты, полученные на выборочной совокупности, на всю генеральную совокупность.

В большинстве медицинских исследований врачу приходится, как правило, иметь дело с частью изучаемого явления, а выводы по результатам такого исследования переносить на все явление в целом - на генеральную совокупность.

Оценка достоверности результатов исследования предусматривает определение:

- 1) ошибок репрезентативности (средних ошибок средних арифметических и относительных величин) - ***m***;
- 2) доверительных границ средних (или относительных) величин;
- 3) достоверности разности средних (или относительных) величин (по критерию ***t*** - Стьюдента).

#### ***1. Определение средней ошибки средней (или относительной) величины (ошибка репрезентативности – m).***

Теория выборочного метода, наряду с обеспечением репрезентативности, практически сводится к оценке расхождений между числовыми характеристиками генеральной и выборочной совокупности, т. е. к определению средних ошибок и так называемых доверительных границ или интервалов. ***Средняя ошибка позволяет установить тот интервал, в котором заключено действительное значение производной величины при данном числе наблюдений, т. е. средняя ошибка всегда является конкретной.***

Ошибка репрезентативности является важнейшей статистической величиной, необходимой для оценки достоверности результатов исследования. Эта ошибка возникает в тех случаях, когда требуется по части охарактеризовать явление в целом. Эти ошибки неизбежны. Они «вытекают» из сущности выборочного исследования. Генеральная совокупность может быть охарактеризована по выборочной совокупности только с некоторой погрешностью, измеряемой ошибкой репрезентативности.

Ошибки репрезентативности не тождественны обычным представлением об ошибках: методических, точности измерения, арифметических и др.

***По величине ошибки репрезентативности определяют, насколько результаты,***

**полученные при выборочном исследовании, отличаются от результатов, которые могли бы быть получены при проведении сплошного исследования без исключения всех элементов генеральной совокупности.**

Это единственный вид ошибок, учитываемых статистическими методами, которые не могут быть устранены, если не проведено сплошное исследование.

Ошибки репрезентативности можно свести к достаточно малой величине, т.е. к величине допустимой погрешности. Делается это путем увеличения числа наблюдений ( $n$ ).

Каждая средняя величина -  $M$  (средняя длительность лечения, средний рост, средняя масса тела и др.), а также относительная величина -  $P$  (уровень летальности, заболеваемости и др.) должны быть представлены со своей средней ошибкой -  $m$ .

Средняя арифметическая величина выборочной совокупности ( $M$ ) имеет ошибку репрезентативности, которая называется **средней ошибкой средней арифметической** ( $m_M$ ) и определяется по формуле:

$$m_M = \pm \frac{\sigma}{\sqrt{n}}$$

Как видно из этой формулы, между размерами сигмы (отражающей разнообразие явления) и размерами средней ошибки существует прямая связь. Между числом наблюдений и размерами средней ошибки существует обратная связь (пропорциональная не числу наблюдений, а квадратному корню из этого числа). Следовательно, уменьшение величины этой ошибки при определении степени разнообразия ( $\sigma$ ) возможно путем увеличения числа наблюдений. При числе наблюдений менее 30 в знаменателе следует взять ( $n - 1$ ).

$$m_M = \pm \frac{\sigma}{\sqrt{n-1}}$$

На этом принципе основан метод определения достаточного числа наблюдений для выборочного исследования.

Относительные величины ( $P$ ), полученные при выборочном исследовании, также имеют свою ошибку репрезентативности, которая называется **средней ошибкой относительной величины** и обозначается  $m_P$ .

Для определения средней ошибки относительной величины ( $P$ ) используется следующая формула:

$$m_P = \pm \sqrt{\frac{P \cdot q}{n}}$$

Где:  $P$  - относительная величина.;

$q$  – разность между основанием, на которое рассчитана относительная величина и самой относительной величиной. Если показатель выражен в процентах, то  $q = 100 - P$ ; если  $P$  - в промиллях, то  $q = 1000 - P$ , если  $P$  - в процепцимиллях, то  $q = 10.000 - P$ , и т.д.;

$n$  - число наблюдений. При числе наблюдений менее 30 в знаменатель следует взять ( $n - 1$ ).

$$m_P = \pm \sqrt{\frac{P \cdot q}{n-1}}$$

Каждая средняя арифметическая или относительная величина, полученная на выборочной совокупности, должна быть представлена со своей средней ошибкой. Это дает возможность рассчитать доверительные границы средних и относительных величин, а также определить достоверность разности сравниваемых показателей (результатов исследования).

## **2. Определение доверительных границ.**

Определяя для средней арифметической (или относительной) величины два крайних значения: минимально возможное и максимально возможное, находят пределы, в которых может быть искомая величина **генерального параметра**. Эти пределы называют доверительными границами.

**Доверительные границы - границы средних (или относительных) величин, выход за пределы которых вследствие случайных колебаний имеет незначительную вероятность.**

Вероятность попадания средней или относительной величины в доверительный интервал называется **доверительной вероятностью**.

Доверительные границы **средней арифметической генеральной совокупности** определяют по формуле:

$$M_{ген} = M_{выб} \pm t \cdot m_M$$

Доверительные границы относительной величины в генеральной совокупности определяют по следующей формуле:

$$P_{ген} = P_{выб} \pm t \cdot m_p$$

Где:  $M_{ген}$  и  $P_{ген}$  - значения средней и относительной величин, полученных для генеральной совокупности;

$M_{выб}$  и  $P_{выб}$  - значения средней и относительной величин, полученных для выборочной совокупности;

$m_M$  и  $m_p$  - ошибки репрезентативности выборочных величин;

$t$  - доверительный критерий, который зависит от величины безошибочного прогноза, устанавливаемого при планировании исследования.

Произведение  $t \cdot m$  ( $\Delta$ ) - предельная ошибка показателя, полученного при данном выборочном исследовании.

Размеры предельной ошибки зависят от коэффициента  $t$ , который избирает сам исследователь, исходя из заданной вероятности безошибочного прогноза.

Величина критерия  $t$  связана с вероятностью безошибочного прогноза ( $P$ ) и числом наблюдений в выборочной совокупности (табл. 4.1).

Таблица 4.1

Зависимость доверительного критерия  $t$  от степени вероятности безошибочного прогноза  $P$  (при  $n > 30$ )

Степень вероятности безошибочного прогноза ( $P$ %)	Доверительный критерий $t$
95,0	2
99,0	2,6
99,9	3,3

Для большинства медико-биологических и социальных исследований достоверными считаются доверительные границы, установленные с вероятностью безошибочного прогноза = 95% и более.

Чтобы найти критерий  $t$  при числе наблюдений ( $n$ ) < 30, необходимо пользоваться специальной таблицей Н.А.Плохинского (табл. 4.2), в которой слева показано число наблюдений - единица ( $n - 1$ ), а сверху ( $P$ ) - степень вероятности безошибочного прогноза.

При определении доверительных границ сначала надо решить вопрос о том, с какой степенью вероятности безошибочного прогноза необходимо представить доверительные границы средней или относительной величины. Избрав определенную степень вероятности, соответственно этому находят величину доверительного критерия  $t$  при данном числе наблюдений. Таким образом, доверительный критерий устанавливается заранее, при планировании исследования.

Таблица 4.2  
Значение критерия t для трех степеней вероятности (по Н.А. Плохинскому)

n = n-1	P	95%	99%	99,9%
1		12,7	63,7	37,0
2		4,3	9,9	31,6
3		3,2	5,8	12,9
4		2,8	4,6	8,6
5		2,6	4,0	6,9
6		2,4	3,7	6,0
7		2,4	3,5	5,3
8		2,3	3,4	5,0
9		2,3	3,3	4,8
10		2,2	3,2	4,6
11		2,2	3,1	4,4
12		2,2	3,1	4,3
13		2,3	3,0	4,1
14-15		2,1	3,0	4,1
16-17		2,1	2,9	4,0
18-20		2,1	2,9	3,9
21-24		2,1	2,8	3,8
25-29		2,0	2,8	3,7

Любой параметр (средняя или относительная величина) может оцениваться с учетом доверительных границ, полученных при расчете.

**Например:** требуется определить доверительные границы среднего уровня пепсина у больных гипертертиозом с 95% вероятностью безошибочного прогноза. Если известно, что:

$$\begin{aligned} n &= 49; \\ M_{\text{выб}} &= 1\text{г}\%; \\ m_M &= \pm 0,05\text{г}\% \end{aligned}$$

1.Определение доверительных границ средней величины в генеральной совокупности:

$$M_{\text{ген}} = M_{\text{выб}} \pm t \cdot m_M = 1\text{г}\% \pm 2 \cdot 0,05\text{г}\%$$

$$1\text{г}\% + 0,1\text{г}\% = 1,1\text{ г}\%$$

$$M_{\text{ген}} =$$

$$1\text{г}\% - 0,1\text{г}\% = 0,9\text{ г}\%$$

**Заключение:** установлено с вероятностью безошибочного прогноза 95%, что средний уровень пепсина в генеральной совокупности у больных гипертертиозом находится в пределах от 1,1 г% до 0,9 г%.

Как видно, доверительные границы зависят от размера доверительного интервала.

Анализ доверительных интервалов указывает, что при заданных степенях вероятности и  $n > 30$  - t имеет неизменную величину и при этом доверительный интервал зависит от величины ошибки репрезентативности.

С уменьшением величины ошибки суживаются доверительные границы средних и относительных величин, полученных на выборочной совокупности, т.е. уточняются

результаты исследования, которые приближаются к соответствующим величинам генеральной совокупности. Если ошибка большая, то получают для выборочной величины большие доверительные границы, которые могут противоречить логической оценке искомой величины в генеральной совокупности. В подобном случае надо искать резервы сокращения размаха доверительных границ в размере величины ошибки репрезентативности.

Доверительные границы  $M_{выб}$  и  $P_{выб}$  зависят не только от средних ошибок этих величин, но и от избранной исследователем степени вероятности безошибочного прогноза. При большой степени вероятности размах доверительных границ увеличивается.

### **3. Определение достоверности разности средних (или относительных) величин (по критерию $t$ - Стьюдента).**

В медицине и здравоохранении по разности параметров оценивают средние и относительные величины, полученные для разных групп населения по полу, возрасту, а также групп больных и здоровых и т.д. Во всех случаях при сопоставлении двух сравниваемых величин возникает необходимость не только определить их разность, но и оценить ее достоверность.

Достоверность разности величин, полученных при выборочных исследованиях, означает, что вывод об их различии может быть перенесен на соответствующие генеральные совокупности.

Достоверность разности выборочной совокупности измеряется доверительным критерием, который рассчитывается по специальным формулам для средних и относительных величин.

Формула оценки достоверности разности сравниваемых средних величин:

$$t = \frac{M_1 - M_2}{\sqrt{m_1^2 + m_2^2}}$$

Для относительных величин:

$$t = \frac{P_1 - P_2}{\sqrt{m_1^2 + m_2^2}}$$

Где:  $M_1$ ;  $M_2$ ;  $P_1$ ;  $P_2$  - параметры, полученные при выборочных исследованиях;  $m_1$ ;  $m_2$  - их средние ошибки;  $t$  - критерий достоверности (Стьюдента).

Разность статистически достоверна при  $t \geq 2$ , что соответствует вероятности безошибочного прогноза, равной 95% и более.

Для большинства исследований, проводимых в медицине и здравоохранении, такая степень вероятности является вполне достаточной.

При величине критерия достоверности  $t < 2$  степень вероятности безошибочного прогноза составляет  $P < 95\%$ . При такой степени вероятности нельзя утверждать, что полученная разность показателей достоверна с достаточной степенью вероятности. В этом случае необходимо получить дополнительные данные, увеличив число наблюдений.

Иногда при увеличении численности выборки разность продолжает оставаться не достоверной. Если при повторных исследованиях разность остается недостоверной, можно считать доказанным, что между сравниваемыми совокупностями не обнаружено различий по изучаемому признаку.



Рассмотрим выборку объемом  $n$  – пусть среднее значение этой выборки равно  $M_1$ , среднеквадратичное отклонение. И выборку объемом  $n$  со средним  $M_2$ , среднеквадратичным отклонением. При этом  $M_1 \neq M_2$ , а выборки подчиняются нормальному закону распределения. Обозначим разницу средних значений выборок.

Нулевая гипотеза в данном случае гласит: «Наблюдаемая разница между выборочными средними была получена случайным образом. не выходит за пределы своих собственных случайных колебаний». Как говорилось выше, нулевая гипотеза не может быть отвергнута, если ее вероятность превысит некоторый порог, называемый уровнем значимости.

Альтернативная гипотеза утверждает противоположное: «Наблюдаемая разница между выборочными средними не могла быть получена случайным образом. Наблюдаемая разница средних выходит за пределы возможных случайных колебаний». Альтернативная гипотеза может быть принята, если ее вероятность сравнивается с некоторым порогом или превысит его.

Проверка гипотез производится при помощи критерия Стьюдента, обозначаемого символом  $t$ :

$$t = \frac{M_1 - M_2}{\sqrt{m_1^2 + m_2^2}}$$

где  $m_1$  и  $m_2$  стандартные ошибки или меры отклонения наблюдаемой разницы выборочных средних от теоретически возможной, «генеральной». Формально величина  $t$  показывает, во сколько раз разница выборочных средних превышает свою собственную случайную вариацию.

При этом, как для первой, так и для второй выборки стандартная ошибка  $m$  рассчитывается по формуле:

$$m = \pm \frac{\sigma}{\sqrt{n}}$$

Полученное значение критерия  $t$  сравнивают со стандартным табличным значением  $t_{кр}$  критерия Стьюдента для выбранного уровня значимости и числа степеней свободы.

Если, нулевая гипотеза не может быть отвергнута, и различие выборочных средних считается «статистически незначимым» (при этом обязательно указывается при каком уровне значимости это имеет место).

Например, при сравнении двух групп критерий  $t_{кр}$  равен 2, и, если полученное значение  $t$  больше 2, то различие статистически значимо и это можно утверждать с вероятностью безошибочного прогноза, равной 95% (при  $t_{кр} = 3$  и более – с вероятностью безошибочного прогноза – 99%). Величина критерия менее 2 свидетельствует об отсутствии статистической значимости различий сравниваемых показателей.

Пример.

Имеется две группы пациентов численностью 247 и 116 человек. Средний возраст пациентов первой группы наблюдения составил  $(32,06 \pm 9,62)$  лет ( $M \pm \sigma$ ), средний возраст пациентов второй группы –  $(39,22 \pm 6,39)$  лет. Сравним 2 группы пациентов по возрасту при условии, что возраста в обеих группах были распределены нормально.

Вначале рассчитаем стандартные ошибки для возрастов в каждой группе.

Поскольку полученная величина  $t$  больше  $t_{кр} = 3$ , то нулевая гипотеза отвергается, и различия между группами по возрасту можно считать статистически значимыми ( $p < 0,01$ ).

Параметрические критерии обладают высокой информативностью, поскольку позволяют не только обнаружить достоверность различий, но и точно, конкретно демонстрируют их характер и степень. Однако, при всех несомненных достоинствах параметрические критерии обладают и рядом существенных недостатков – **ограничениями их применимости**. Самый серьезный из них - допущение о нормальности распределения сравниваемых величин. Второе ограничение - непригодность таких критериев к выборкам малого объема ( $< 10-15$  измерений). На таких выборках параметры распределения (средние, дисперсии) могут резко измениться от добавления или убавления даже одного единственного числа. Третье – высокая чувствительность к артефактам, которые оказывают сильное влияние на параметры распределения, вызывая сдвиг средних значений в ту или иную сторону. В результате может «всплыть» различие, которого на самом деле нет или наоборот – оказаться «зашумленной» действительная разница. Влияние артефактов особенно велико на малых выборках. Специфика же медицинской работы состоит в том, что из-за сложности исследуемых процессов и явлений они, как правило, имеют дело именно с выборками малого объема, имеющими неизвестный закон распределения, часто полученными в результате достаточно грубых измерений, «нашпигованными» артефактами.

Для извлечения содержательной информации из числовых массивов такого рода были разработаны **непараметрические критерии**. Это критерии, применение которых не требует пересчета массивов исходных данных в компактно заменяющие их параметры распределения - средние значения, дисперсии или стандартные отклонения и т.д. – и их последующее сравнение.

Как следствие, не только теряет силу требование «нормальности» генеральной совокупности, но и, более того, закон распределения сравниваемых величин **вообще не играет никакой роли**. Особые, достаточно простые, способы преобразования исходных данных делают эту группу критериев еще и практически нечувствительными к артефактам. В результате, непараметрические критерии успешно работают даже на чрезвычайно малых выборках при наличии грубых измерений и грубых ошибок.

### **Рассмотрим критерий Вилкоксона.**

Критерий Вилкоксона – **критерий ранговый**, т.е. основанный на сравнении сумм рангов, полученных тем или иным образом из сравниваемых выборочных распределений. В данном конкретном случае **рангом называется порядковый номер числа в ранжированном (расставленном в порядке возрастания) массиве данных** – чем больше число, тем выше его ранг. При этом, если числа не повторяются, то их ранги в точности соответствуют их порядковым номерам. Если же некое число повторяется несколько раз, то всем им приписывается *средний ранг*. Продемонстрируем, как все это происходит и выглядит. Допустим, мы получили следующий вариационный ряд данных  $x$ :

5.6 11.7 -3.5 6.3 8 7.4 0.5 8 3 3.1 15.2 3.1 8 6.7 111 4.4

Здесь числа представлены в том порядке, как они были получены.

Расставим их в порядке возрастания и припишем порядковые номера, а также ранги  $R$ :

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
x	-3.5	0.5	3	3.1	3.1	4.4	5.6	6.3	6.7	7.4	8	8	8	11.7	15.2	111
R	1	2	3	4.5	4.5	6	7	8	9	10	12	12	12	14	15	16

Из приведенного примера хорошо видно, что при ранжировании происходит «линеаризация данных» - сглаживание их резких колебаний за счет того, что ранг числа не зависит от его абсолютной величины и разницы с соседними вариантами. Например, последнее число 111 чуть ли не на порядок превышает ближайшее к нему 15.2. Тем не менее, ранг его всего на 1 выше, чем у предпоследнего числа.

Ранговые критерии для сравнения выборочных совокупностей делятся на две группы – для независимых и зависимых выборок.

### **Критерий парных сравнений Вилкоксона – ранговый критерий для сравнения зависимых выборок.**

*Рассмотрим его на примере.* У 10 здоровых взрослых людей измеряли кровяное давление после введения кофеина и плацебо. Получены следующие данные для «верхнего», систолического давления СД:

x(кофеин)	126	145	137	116	137	157	125	139	143	163
y(плацебо)	121	143	115	120	135	157	115	130	153	160

Возникает вопрос, можно ли на основании этих данных полагать, что кофеин оказывает физиологическое действие.

Вначале значения одного ряда строго попарно вычитают из значений другого с учетом знака разницы  $d$ . Вычтем нижний ряд из верхнего:

x(кофеин)	126	145	137	116	137	157	125	139	143	163
y(плацебо)	121	143	115	120	135	157	115	130	153	160
$d$	5	2	22	-4	2	0	10	9	-10	3

Далее разницы ранжируют по известным правилам, но при этом не учитывают знак разницы (т.е. ранжируют по модулю). Нулевую пару отбрасывают.

$d$	2	2	3	-4	5	9	10	-10	22
$R$	1.5	1.5	3	4	5	6	7.5	7.5	9

Отдельно суммируют ранги для положительных и отрицательных разниц. В нашем случае получаем: , . В качестве значения критерия  $T_z$  берут меньшую сумму независимо от знака, т.е.  $T_z=11,5$ . Сравниваем это значение с «критическим» из специальной таблицы, входом в которую является число сравниваемых пар, но лишь тех, которые **не дают нулевые** разницы. В нашем случае таковых 9. Тогда  $T_{кр} = 6$  для и  $T_{кр} = 2$  для. Поскольку даже для первого уровня значимости, различий уровней СД нулевую гипотезу отвергнуть нельзя и различия не являются статистически значимыми ( $p < 0,05$ ). Иными словами, у нас нет пока оснований утверждать, что действие кофеина носит исключительно физиологический характер.

Смысл теста состоит в следующем. Если бы мы имели бесконечно большой ряд случайных разниц, то число и величина положительных разниц равнялись бы числу отрицательных и, соответственно, суммы их рангов были бы равны. На конечном и ограниченном числовом массиве опять же чисто случайно может иметь место «перекос» в сторону преимущественно положительных или отрицательных разниц. Это обстоятельство и учитывается в критических значениях критерия.

$T_{кр}$  – это граница между практически возможными и **практически** невозможными значениями **критерия**. Соответственно, если, то полученная нами сумма рангов с достаточно высокой вероятностью могла возникнуть чисто случайно и о сдвиге одного числового ряда относительно другого ничего определенного сказать нельзя. Это недостоверное различие. Если же, то наблюдаемое различие положительных и отрицательных разниц **не могло быть получено случайным образом**. Это означает, что смещение значений в сопоставляемых числовых рядах объясняется действием какой-то систематически действующей, **неслучайной** причины, т.е. носит статистически достоверный (устойчивый и прогнозируемый) характер.

Как было показано выше, пары, имеющие одинаковые числовые значения и, соответственно, дающие нулевые разницы, **исключаются из рассмотрения**. И если таких случаев много, то «жесткость» критерия нарастает, поскольку  $T_{кр}$  тем меньше, чем меньше сравниваемых пар. Соответственно, увеличивается число ситуаций, когда нулевую гипотезу отвергнуть невозможно, и различие будет считаться незначимым. Более того, если число пар окажется меньше 6, то критерий Вилкоксона вообще перестанет «работать»: **6 - минимальное число пар**, для которого еще существует  $T_{кр}$ . Для меньшего числа его просто невозможно рассчитать. А подобные ситуации в медико-биологической практике возникают довольно часто, поскольку многие измерения неизбежно приходится выполнять с достаточно высокой степенью грубости, и вероятность появления совпадающих значений здесь все еще весьма высока.

Для того, чтобы **рассчитать t-критерий Стьюдента (для зависимых и для независимых выборок) в Excel** необходимо сделать следующие шаги:

1. Вносим значения для двух переменных в таблицу (Например, *Переменная 1* и *Переменная 2*).
2. Ставим курсор в пустую ячейку.
3. В строке формул (**fx**) (*вставить функцию*) или на ленте вкладка **ФОРМУЛЫ - ВСТАВИТЬ ФУНКЦИЮ (fx)**.
4. В открывшемся окне «*Мастер функций*» в поле «Категории» выбираем **Полный алфавитный перечень**.

5. Затем в поле «*Выберите функцию*» находим функцию **ТТЕСТ**, которая возвращает вероятность, соответствующую критерию Стьюдента.

5.1. Нажимаем **Ок**.

6. В открывшемся окне «*Аргументы функции*» в поле Массив1 вносим **номера ячеек**, содержащие значения Переменной 1, в поле Массив2 вносим **номера ячеек**, содержащие значения Переменной2.

7. В поле «*Хвосты*» пишем **2** (критерий будет рассчитываться используя **двустороннее распределение**, как и в SPSS); либо **1** (критерий будет рассчитываться используя **одностороннее распределение**).

**Важно!** 8. В поле «Тип» пишем **1** (рассчитывается, если **выборки зависимые**); либо **2** или **3** (если **выборки независимые**).

9. Нажимаем **Ок**.

10. Смотрим получившийся результат.

### ***Пример расчета Т-критерия Вилкоксона***

Допустим мы сравниваем между собой уровень тревожности подростков до и после тренинга уверенности в себе.

Шаг 1. Запишем значения в таблицу.

Шаг 2. Рассчитаем разность значений. Для данного случае типичным сдвигом будет считаться сдвиг в отрицательную сторону (7 значений, красный цвет заливки), а нетипичным в положительную сторону (3 значения, зеленый цвет заливки)

Шаг 3. Найдем значения шага 2 по модулю

Шаг 4. Проранжируем значения по модулю.

Все четыре шага приведены в таблице.

№	Уровень тревожности (до тренинга)	Уровень тревожности (после тренинга)	Шаг 2: Разность (после-до)	Шаг 3: Значение разности по модулю	Шаг 4: Ранг разности
1	15	14	-1	1	3
2	14	11	-3	3	8
3	16	17	1	1	3
4	18	19	1	1	3
5	21	20	-1	1	3
6	21	18	-3	3	8

7	20	15	-5	5	10
8	15	17	2	2	6
9	17	14	-3	3	8
10	13	12	-1	1	3

Шаг 5. Найдем  $T$  эмпирическое вычислив сумму рангов в нетипичном направлении (зеленый цвет заливки).

$$T_{emp} = 3 + 3 + 6 = 12$$

Шаг 6. Используя [таблицу критических значений Т-критерия Вилкоксона](#) определяем Т-критическое

6.1. Находим количество человек в выборке.  $n=10$

6.2. Определяем Т-критическое справа от значения количества человек в выборке. для  $p<0,05$   $T=10$ ; для  $p<0,01$   $T=5$

Шаг 7. Сравниваем Т-критическое и Т-эмпирическое.

$$T_{emp} = 12 \geq T_{kr} = 10$$

Шаг 8. Делаем выводы.

### Критерий хи-квадрат для таблиц сопряженности в Excel

Поскольку в Excel можно строить таблицы сопряженности для категориальных переменных при помощи сводных таблиц, то легко вычислить и критерий независимости переменных хи-квадрат, а также меры связи, основанные на нем. Пусть у нас имеется такая таблица сопряженности:

A	B	C	D	E
	да	затр.	нет	всего
да	12	23	35	70
затр.	25	15	30	70
нет	40	14	24	78
всего	77	52	89	218

Прежде всего следует вычислить ожидаемые частоты для всех ячеек, используя такое **правило**:

*ожидаемая частота (в предположении независимости переменных) в ячейке равна сумме по столбцу, умноженной на сумму по строке, деленной на общую сумму.*

В Excel такие вычисления удобнее всего провести при помощи формул-массивов. Скопируем таблицу на другой лист и удалим значения всех ячеек, кроме итоговых. Теперь выделим прямоугольный диапазон с пустыми ячейками внутри будущей таблицы ожидаемых частот.

Нажмите знак равенства и начинайте вводить формулу:

1. выделите строку итогов, нажмите \* (знак умножения),
2. затем выделите столбец итогов, нажмите / (знак деления),
3. потом щелкните по ячейке с общим итогом в таблице и нажмите сочетание клавиш Shift+Control+Enter.

Эта последовательность действий создаст особый тип формул в данном диапазоне (вы увидите, что формула заключена в фигурные скобки):

24,7	16,7	28,6	70
24,7	16,7	28,6	70
24,6	18,6	31,8	78
77	52	89	218

Вот теперь вычисление хи-квадрат не представляет трудностей. По обычной формуле нужно вычислить вклады каждой ячейки в критерий хи-квадрат:  $(\text{наблюдаемая} - \text{ожидаемая})^2 / \text{ожидаемая}$ , потом все сложить для получения хи-квадрат для таблицы в целом и получить вероятность наблюдать такое или большее значение критерия

#### 4. Цель деятельности аспирантов на занятии:

##### Аспирант должен знать:

18. Генеральная совокупность. Выборка.
19. Корреляция.
20. Параметрические и непараметрические коэффициенты.
21. Оценка достоверности сдвига.

##### Аспирант должен уметь:

1. Применять критерии Стьюдента, Вилкоксона, Фридмана для выявления достоверности сдвига исследуемого признака.

##### Содержание обучения:

##### Теоретическая часть:

3. Использование параметрических и непараметрических критериев для выявления достоверности сдвига исследуемого признака.
4. Алгоритм получения значений критериев Стьюдента, Вилкоксона, Фридмана. Формулы для вычисления этих критериев.
5. Расчет при помощи пакета анализа EXCEL, STATISTICA, SPSS.

##### Практическая часть:

Задача №1. Применение критериев Стьюдента и Вилкоксона для выявления достоверности сдвига исследуемого признака.

Условие: найти, вызывает ли выбранная тактика лечения изменения в длительности сердечного цикла у одного и того же человека до и после купирования острой сердечной недостаточности.

№	До	После
1	0,91	0,92
2	0,71	0,74
3	0,73	0,71
4	0,82	0,83
5	0,67	0,92
6	0,89	0,89
7	0,9	0,93
8	0,77	0,86
9	0,78	0,85

Задача №2. Применение критерия Фридмана для определения достоверности сдвига исследуемого признака.

*Условие:* исследовалась реакция переключения внимания по таблице Шульте-Платонова у студентов четырех темпераментных групп по Д. Кейрси: SP, SJ, NF, NT.

*Найти:* имеется ли достоверный сдвиг в показателях скорости(сек) на отыскание 25 чисел и цифр?

№	Типы темперамента			
	SP	SJ	NF	NT
	Время, сек.			
<b>1</b>	42	56	58	45
<b>2</b>	42	44	44	61
<b>3</b>	79	70	62	63
<b>4</b>	69	65	67	56
<b>5</b>	50	50	64	49
<b>6</b>	45	66	56	66
<b>7</b>	42	43	55	70
<b>8</b>	45	44	55	66
<b>9</b>	46	51	45	67
<b>10</b>	40	45	60	66

**9. Перечень вопросов для проверки исходного уровня знаний:**

4. Параметрические критерии различий в уровне признака.
5. Непараметрические критерии в уровне признака.
6. Корреляция.
7. Достоверность сдвига исследуемого признака.

**10. Перечень вопросов для проверки конечного уровня знаний:**

5. Применение критерия Стьюдента.
6. Алгоритм получения критерия Вилкоксона. Применение данного критерия.
7. Применение критерия Фридмана.

**11. Хронокарта учебного занятия:**

17. Организационный момент – 10 мин.
18. Разбор темы – 40 мин.
19. Текущий контроль (тестирование, практическая работа) - 90 мин.
20. Подведение итогов занятия – 10 мин.

**12. Самостоятельная работа аспиранта.**

Использование пакета SPSS и Statistica для расчета критериев Стьюдента, Вилкоксона и Фридмана.

**13. Перечень учебной литературы к занятию:**

2. Есауленко И.Э., Семенов С.Н. Основы практической информатики в медицине; Воронеж, 2005.
3. Жижин К. С. Медицинская статистика; Ростов н/Д, 2007.



## ТЕМА 7: Применение линейной корреляции. Пирсона. Коэффициент Спирмена

### 1. Научно-методическое обоснование темы:

В медико-биологических исследованиях чаще всего встречаются связи второго типа, поэтому в качестве мер связи наиболее часто используются либо линейный коэффициент корреляции Пирсона, либо ранговый коэффициент корреляции Спирмена

Корреляционный анализ позволяет выявить наличие, отсутствие и степень связи между двумя и более рядами экспериментальных данных. В корреляционном анализе все переменные являются зависимыми, т.е. выявление того, что является причиной, а что является следствием связано с качественным анализом.

### 2. Краткая теория:

Согласованность изменений признаков исследуется с помощью различных мер связи, которые традиционно разделяются на функциональные (точные) и корреляционные (вероятностные или стохастические). В медико-биологических исследованиях чаще всего встречаются связи второго типа, поэтому в качестве мер связи наиболее часто используются либо линейный коэффициент корреляции Пирсона, либо ранговый коэффициент корреляции Спирмена (значения этих коэффициентов: от -1 до +1).

Корреляционный анализ позволяет выявить наличие, отсутствие и степень связи между двумя и более рядами экспериментальных данных. В корреляционном анализе все переменные являются зависимыми, т.е. выявление того, что является причиной, а что является следствием связано с качественным анализом. Виды корреляционного анализа:

1. **линейный:** позволяет выявить связи между абсолютными значениями переменной.
2. **ранговый:** позволяет выявить связи между рангами значений переменной.
3. **парный:** позволяет выявить связи между парами переменных.
4. **множественный:** позволяет выявить связи между многими переменными одновременно

**Методы изучения** корреляционной связи: параметрические (применимы для распределений, близких к нормальным: метод линейной корреляции Пирсона, дихотомический коэффициент корреляции) и непараметрические (для любых распределений: Спирмена, Кендела).

**Количественной мерой** корреляционной связи являются коэффициенты корреляции.

В зависимости от знака при коэффициенте различают положительные и отрицательные корреляционные связи. Нулевое значение коэффициента означает отсутствие связи; чем ближе абсолютная величина коэффициента к 1, тем корреляционная связь сильнее (и ближе к функциональной зависимости).

**Т.О., СИЛА КОРРЕЛЯЦИОННОЙ СВЯЗИ ОПРЕДЕЛЯЕТСЯ ЗНАЧЕНИЕМ АБСОЛЮТНОЙ ВЕЛИЧИНЫ КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ.**

Общая корреляция		
№ п/п	Тип связи	Сила связи(г)
1	Сильная, или тесная, связь	Более 0,70
2	Средняя связь	От 0,5 до 0,69
3	Умеренная связь	От 0,30 до 0,50
4	Слабая связь	От 0,20 до 0,29
5	Очень слабая связь	Менее 0,20
Частная корреляция		
№ п/п	Тип связи	Уровень статистической значимости связи
1	Высокая значимая корреляция	$p \leq 0,01$

2	Значимая корреляция	0,01 до 0,05
3	Тенденция достоверной связи	$0,05 < p \leq 0,10$
4	Незначимая корреляция	$0,10 < p$

Общая классификация характеризует абсолютную величину коэффициента корреляции (силу корреляции), а частная классификация выделяет уровень статистической значимости – величину коэффициента корреляции при заданном объеме выборки. **В результате, для малых выборок даже сильная корреляционная связь может оказаться НЕДОСТОВЕРНОЙ; напротив, для больших выборок даже слабая связь может оказаться ДОСТОВЕРНОЙ.** В медицине и биологии в первую очередь ориентируются на частную корреляцию и лишь потом применяют для их ранжирования общую корреляцию.

Наиболее распространен в исследованиях линейный коэффициент корреляции Пирсона, выборочный коэффициент корреляции, коэффициент корреляции Бравейса-Пирсона. Он измеряет силу линейной корреляционной связи количественных признаков.

**Коэффициент корреляции Спирмена:** Позволяет исследовать связь между двумя и более рядами экспериментальных данных. При этом связь интерпретируется как ранговая, т.е. фактически оценивается соответствие порядка следования данных. Коэффициент ранговой корреляции Спирмена – это непараметрический метод, который используется с целью статистического изучения связи между явлениями. В этом случае определяется фактическая степень параллелизма между двумя количественными рядами изучаемых признаков и дается оценка тесноты установленной связи с помощью количественно выраженного коэффициента.

Коэффициент ранговой корреляции Спирмена используется для выявления и оценки тесноты связи между двумя рядами сопоставляемых количественных показателей. В том случае, если ранги показателей, упорядоченных по степени возрастания или убывания, в большинстве случаев совпадают (большему значению одного показателя соответствует большее значение другого показателя - например, при сопоставлении роста пациента и его массы тела), делается вывод о наличии прямой корреляционной связи. Если ранги показателей имеют противоположную направленность (большему значению одного показателя соответствует меньшее значение другого - например, при сопоставлении возраста и частоты сердечных сокращений), то говорят об обратной связи между показателями.

**Коэффициент корреляции Спирмена обладает следующими свойствами:**

- Коэффициент корреляции может принимать значения от минус единицы до единицы, причем при  $r_s=1$  имеет место строго прямая связь, а при  $r_s= -1$  – строго обратная связь.
- Если коэффициент корреляции отрицательный, то имеет место обратная связь, если положительный, то – прямая связь.
- Если коэффициент корреляции равен нулю, то связь между величинами практически отсутствует.
- Чем ближе модуль коэффициента корреляции к единице, тем более сильной является связь между измеряемыми величинами.

В связи с тем, что коэффициент является методом непараметрического анализа, проверка на нормальность распределения не требуется.

Сопоставляемые показатели могут быть измерены как в непрерывной шкале (например, число эритроцитов в 1 мкл крови), так и в порядковой (например, баллы экспертной оценки от 1 до 5).

Эффективность и качество оценки методом Спирмена снижается, если разница между различными значениями какой-либо из измеряемых величин достаточно велика. *Не рекомендуется использовать коэффициент Спирмена, если имеет место неравномерное распределение значений измеряемой величины.*

Расчет коэффициента ранговой корреляции Спирмена включает следующие этапы:

- Сопоставить каждому из признаков их порядковый номер (ранг) по возрастанию или убыванию.
- Определить разности рангов каждой пары сопоставляемых значений (d).
- Возвести в квадрат каждую разность и суммировать полученные результаты.
- Вычислить коэффициент корреляции рангов по формуле:

$$r = 1 - \frac{6 \cdot \sum d^2}{n(n^2 - 1)}$$

Определить статистическую значимость коэффициента при помощи t-критерия, рассчитанного по следующей формуле:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

При использовании коэффициента ранговой корреляции условно оценивают тесноту связи между признаками, считая значения коэффициента равные 0,3 и менее - показателями слабой тесноты связи; значения более 0,4, но менее 0,7 - показателями умеренной тесноты связи, а значения 0,7 и более - показателями высокой тесноты связи.

Статистическая значимость полученного коэффициента оценивается при помощи t-критерия Стьюдента. Если рассчитанное значение t-критерия меньше табличного при заданном числе степеней свободы, статистическая значимость наблюдаемой взаимосвязи - отсутствует. Если больше, то корреляционная связь считается статистически значимой.

Для того, чтобы **рассчитать коэффициент корреляции в Excel** необходимо сделать следующие шаги:

1. Вносим значения для двух переменных в таблицу (Например *Переменная 1* и *Переменная 2*).
2. Ставим курсор в пустую ячейку.
- 3 В строке формул нажимаем кнопку ***fx*** или **на ленте вкладки *Формулы*- Вставить функцию (fx)**.
4. В открывшемся окне «*Мастер функций*» в поле «Категории» выбираем **Полный алфавитный перечень**.
5. Затем в поле «*Выберите функцию*» находим функцию **КОРЕЛЛ**.
- 5.1. Нажимаем **Ок**.
6. В открывшемся окне «*Аргументы функции*» в поле Массив1 вносим **номера ячеек**, содержащие значения Переменной 1, в поле Массив2 вносим **номера ячеек**, содержащие значения Переменной2.
7. Нажимаем **Ок**.
8. Смотрим получившийся результат.

### **Коэффициент корреляции Пирсона:**

*Критерий корреляции Пирсона* – это метод параметрической статистики, позволяющий определить наличие или отсутствие линейной связи между двумя количественными показателями, а также оценить ее тесноту и статистическую значимость. Другими словами, критерий корреляции Пирсона позволяет определить, есть

ли линейная связь между изменениями значений двух переменных. В статистических расчетах и выводах коэффициент корреляции обычно обозначается как  $r_{xy}$  или  $R_{xy}$ .

Критерий корреляции Пирсона позволяет определить, какова теснота (или сила) корреляционной связи между двумя показателями, измеренными в количественной шкале. При помощи дополнительных расчетов можно также определить, насколько статистически значима выявленная связь.

Например, при помощи критерия корреляции Пирсона можно ответить на вопрос о наличии связи между температурой тела и содержанием лейкоцитов в крови при острых респираторных инфекциях, между ростом и весом пациента, между содержанием в питьевой воде фтора и заболеваемостью населения кариесом.

### Условия и ограничения применения критерия хи-квадрат Пирсона

1. Сопоставляемые показатели должны быть измерены в *количественной шкале* (например, частота сердечных сокращений, температура тела, содержание лейкоцитов в 1 мл крови, систолическое артериальное давление).
2. Посредством критерия корреляции Пирсона можно определить лишь *наличие и силу линейной взаимосвязи* между величинами. Прочие характеристики связи, в том числе направление (прямая или обратная), характер изменений (прямолинейный или криволинейный), а также наличие зависимости одной переменной от другой - определяются при помощи регрессионного анализа.
3. Количество сопоставляемых величин должно быть равно двум. В случае анализ взаимосвязи трех и более параметров следует воспользоваться методом *факторного анализа*.
4. Критерий корреляции Пирсона является *параметрическим*, в связи с чем условием его применения служит *нормальное распределение* сопоставляемых переменных. В случае необходимости корреляционного анализа показателей, распределение которых отличается от нормального, в том числе измеренных в порядковой шкале, следует использовать коэффициент ранговой корреляции Спирмена.
5. Следует четко различать понятия зависимости и корреляции. Зависимость величин обуславливает наличие корреляционной связи между ними, но не наоборот.

Например, рост ребенка зависит от его возраста, то есть чем старше ребенок, тем он выше. Если мы возьмем двух детей разного возраста, то с высокой долей вероятности рост старшего ребенка будет больше, чем у младшего. Данное явление и называется *зависимостью*, подразумевающей причинно-следственную связь между показателями. Разумеется, между ними имеется и *корреляционная связь*, означающая, что изменения одного показателя сопровождаются изменениями другого показателя. В другой ситуации рассмотрим связь роста ребенка и частоты сердечных сокращений (ЧСС). Как известно, обе эти величины напрямую зависят от возраста, поэтому в большинстве случаев дети большего роста (а значит и более старшего возраста) будут иметь меньшие значения ЧСС. То есть, *корреляционная связь* будет наблюдаться и может иметь достаточно высокую тесноту. Однако, если мы возьмем детей *одного возраста*, но *разного роста*, то, скорее всего, ЧСС у них будет различаться незначительно, в связи с чем можно сделать вывод о *независимости* ЧСС от роста. Приведенный пример показывает, как важно различать фундаментальные в статистике понятия *связи* и *зависимости* показателей для построения верных выводов.

Расчет коэффициента корреляции Пирсона производится по следующей формуле:

$$r_{xy} = \frac{\sum (d_x \times d_y)}{\sqrt{(\sum d_x^2 \times \sum d_y^2)}}$$

## Как интерпретировать значение коэффициента корреляции Пирсона?

Значения коэффициента корреляции Пирсона интерпретируются исходя из его абсолютных значений. Возможные значения коэффициента корреляции варьируют от 0 до  $\pm 1$ . Чем больше абсолютное значение  $r_{xy}$  – тем выше теснота связи между двумя величинами.  $r_{xy} = 0$  говорит о полном отсутствии связи.  $r_{xy} = 1$  – свидетельствует о наличии абсолютной (функциональной) связи. Если значение критерия корреляции Пирсона оказалось больше 1 или меньше -1 – в расчетах допущена ошибка.

Для оценки тесноты, или силы, корреляционной связи обычно используют общепринятые критерии, согласно которым абсолютные значения  $r_{xy} < 0.3$  свидетельствуют о *слабой* связи, значения  $r_{xy}$  от 0.3 до 0.7 - о связи *средней* тесноты, значения  $r_{xy} > 0.7$  - о *сильной* связи.

Оценка *статистической значимости* коэффициента корреляции  $r_{xy}$  осуществляется при помощи t-критерия, рассчитываемого по следующей формуле:

$$t_r = \frac{r_{xy} \sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$$

Полученное значение  $t_r$  сравнивается с критическим значением при определенном уровне значимости и числе степеней свободы  $n-2$ . Если  $t_r$  превышает  $t_{\text{крит}}$ , то делается вывод о статистической значимости выявленной корреляционной связи.

## Планирование корреляционного исследования:

Выделяют различные планы корреляционного исследования, схематически они могут быть охарактеризованы: испытуемые \* тестирование \* временные этапы.

**1.** Одномерное исследование группы испытуемых через временные этапы. T1(самооценка) T2(самооценка) : r12.

**2.** Косвенное корреляционное исследование. Исследовать связь между полом и интеллектом или полом и эмоц.устойчивостью, пол и интровертность, пол и нейротизм, возраст и чувствительность, возраст и уровень достижений, среда и IQ, семья и социальный интеллект. Среда и IQ: связь в данном плане исследования док-ся через изучение различий в уровне развития свойств у разных групп испытуемых.

**3.** Структурное корреляционное исследование. **3.1.** Позволяет исследовать различия в уровне связей между разными характеристиками у одних и тех же испытуемых. **3.2.** Исследование различий в уровне связей в одной и той же характеристике у разных групп испытуемых.

**План:** Анализ статьи проводится в соответствии с этапами экспериментального исследования:

1. постановка проблемы, выбор первичных гипотез.
2. выбор литературы по проблеме исследования, выбор авторов, на которых опирается исследователь (в конце статьи), описание позиции автора.
3. постановка цели исследования, формулировка гипотезы исследования (чёткая), частные задачи исследования.
4. выбор методов и методик, авторы, шкалы.
5. планирование и этапы исследования: а) берём выборки, б) изучение того ..., др..., возраст.
6. проведение исследования и получение количественных данных.
7. математическая обработка, интерпретация и качественный анализ данных. Отбор и проверка статистических гипотез (гипотеза о различиях, о связях, о причинно-следственных отношениях, о динамике связи, гипотезы о факторах).
8. анализ соответствия гипотез и результатов исследования.

## Формулы, используемые при корреляционном анализе

Формула для вычислений	Функция или инструмент Анализа данных в Excel	Результат вычислений/ Примечания
Среднее значение $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$	СРЗНАЧ (число1; число2;...)	Возвращает среднее значение (среднее арифметическое) аргументов
Дисперсия $S_x^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$	ДИСП (число1; число2;...)	Оценивает дисперсию по выборке
Стандартное отклонение $S_x = \sqrt{S_x^2}$	СТАНДОТКЛОН (число1; число2;...)	Оценивает стандартное отклонение по выборке. Стандартное отклонение – это мера того, насколько широко разбросаны точки данных относительно их среднего
Сумма квадратов отклонений $\sum_{i=1}^n (x_i - \bar{x})^2$	КВАДРОТКЛ(число1;число2;...)	Возвращает сумму квадратов отклонений точек данных от их среднего
Коэффициент корреляции $r_{x,y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$	КОРРЕЛ(массив1;массив2)	Возвращает коэффициент корреляции между интервалами ячеек массив1 и массив2
t-критерий Стьюдента для проверки значимости коэффициента корреляции $t_{\text{набл}} = \sqrt{\frac{r_{y,x}^2}{1 - r_{y,x}^2}} (n - 2)$	СТЮДРАСПОБР (вероятность; степени_свободы)	Вычисленное по этой формуле значение t набл сравнивается с критическим значением t-критерия, которое берется из таблицы значений t-распределения Стьюдента с учетом заданного уровня значимости и числа степеней свободы (n – 2) или определяется с помощью функции СТЮДРАСПОБР()
Матрица коэффициентов парной корреляции	Обращение к средствам анализа данных. Для вычисления матрицы коэффициентов парной корреляции R следует воспользоваться	Инструмент Корреляция применяется, если имеется более двух переменных измерений

$R = \begin{matrix} & \begin{matrix} y_1 \\ y_2 \\ \dots \\ y_n \end{matrix} \\ \begin{matrix} x_1 \\ x_2 \\ \dots \\ x_n \end{matrix} & \begin{matrix} r_{y_1} & r_{y_2} & \dots & r_{y_n} \\ 1 & r_{x_2} & \dots & r_{x_n} \\ r_{x_2} & 1 & \dots & r_{x_n} \\ \dots & \dots & \dots & \dots \\ r_{x_n} & r_{x_n} & r_{x_n} & 1 \end{matrix} \end{matrix}$	инструментом Корреляция из пакета Анализ данных	для каждого объекта. В результате выдается таблица – корреляционная матрица, показывающая значение функции КОРРЕЛ( ) для каждой возможной пары переменных измерений. Любое значение коэффициента корреляции должно находиться в диапазоне от –1 до +1 включительно
--	---	--

## 5. Цель деятельности аспирантов на занятии:

### Аспирант должен знать:

22. Корреляция. Виды корреляции.
23. Типы связи. Сила связи.
24. Формулы расчета коэффициентов Спирмена и Пирсона.

### Аспирант должен уметь:

1. Применять критерий Пирсона для определения корреляционной силы связи между двумя показателями, измеренными в количественной шкале и насколько статистически значима выявленная связь.
2. Применять коэффициент ранговой корреляции Спирмена для выявления и оценки тесноты связи между двумя рядами сопоставляемых количественных показателей.

### Содержание обучения:

#### Теоретическая часть:

6. Формулы расчетов критериев Пирсона и Спирмена.
7. Расчет при помощи пакета анализа EXCEL.

**Практическая часть:**  
**Пример расчета коэффициента корреляции Пирсона.**

Целью исследования явилось выявление, определение тесноты и статистической значимости корреляционной связи между двумя количественными показателями: уровнем тестостерона в крови (X) и процентом мышечной массы в теле (Y). Исходные данные для выборки, состоящей из 5 исследуемых ( $n = 5$ ), сведены в таблице:

N	Содержание тестостерона в крови, нг/дл (X)	Процент мышечной массы, % (Y)
1.	951	83
2.	874	76
3.	957	84
4.	1084	89
5.	903	79

- Вычислим суммы анализируемых значений X и Y:  
 $\Sigma(X) = 951 + 874 + 957 + 1084 + 903 = 4769$   
 $\Sigma(Y) = 83 + 76 + 84 + 89 + 79 = 441$
- Найдем средние арифметические для X и Y:  
 $M_x = \Sigma(X) / n = 4769 / 5 = 953,8$   
 $M_y = \Sigma(Y) / n = 441 / 5 = 88,2$
- Рассчитаем для каждого значения сопоставляемых показателей величину отклонения от среднего арифметического  $d_x = X - M_x$  и  $d_y = Y - M_y$ :

N	Содержание тестостерона в крови, нг/дл (X)	Процент мышечной массы, % (Y)	Отклонение содержания тестостерона от среднего значения ( $d_x$ )	Отклонение % мышечной массы от среднего значения ( $d_y$ )
1.	951	83	-2,8	0,8
2.	874	76	-79,8	-6,2
3.	957	84	3,2	1,8
4.	1084	89	130,2	6,8
5.	903	79	-50,8	-3,2

- Возведем в квадрат каждое значение отклонения  $d_x$  и  $d_y$ :

N	Содержание тестостерона в крови, нг/дл (X)	Процент мышечной массы, % (Y)	Отклонение содержания тестостерона от среднего значения ( $d_x$ )	Отклонение % мышечной массы от среднего значения ( $d_y$ )	$d_x^2$	$d_y^2$
1.	951	83	-2,8	0,8	7,84	0,64



N	Содержание тестостерона в крови, нг/дл (X)	Процент мышечной массы, % (Y)	Отклонение содержания тестостерона от среднего значения ( $d_x$ )	Отклонение % мышечной массы от среднего значения ( $d_y$ )	$d_x^2$	$d_y^2$
2.	874	76	-79,8	-6,2	6368,04	38,44
3.	957	84	3,2	1,8	10,24	3,24
4.	1084	89	130,2	6,8	16952,04	46,24
5.	903	79	-50,8	-3,2	2580,64	10,24

5. Рассчитаем для каждой пары анализируемых значений произведение отклонений  $d_x \times d_y$ :

N	Содержание тестостерона в крови, нг/дл (X)	Процент мышечной массы, % (Y)	Отклонение содержания тестостерона от среднего значения ( $d_x$ )	Отклонение % мышечной массы от среднего значения ( $d_y$ )	$d_x^2$	$d_y^2$	$d_x \times d_y$
1.	951	83	-2,8	0,8	7,84	0,64	-2,24
2.	874	76	-79,8	-6,2	6368,04	38,44	494,76
3.	957	84	3,2	1,8	10,24	3,24	5,76
4.	1084	89	130,2	6,8	16952,04	46,24	885,36
5.	903	79	-50,8	-3,2	2580,64	10,24	162,56

6. Определим значения суммы квадратов отклонений  $\Sigma(d_x^2)$  и  $\Sigma(d_y^2)$ :

$$\Sigma(d_x^2) = 25918,8$$

$$\Sigma(d_y^2) = 98,8$$

7. Найдем значение суммы произведений отклонений  $\Sigma(d_x \times d_y)$ :

$$\Sigma(d_x \times d_y) = 1546,2$$

8. Рассчитаем значение коэффициента корреляции Пирсона  $r_{xy}$  по приведенной выше формуле:

$$r_{xy} = \frac{\Sigma(d_x \times d_y)}{\sqrt{(\Sigma d_x^2 \times \Sigma d_y^2)}} = \frac{1546.2}{\sqrt{(25918.8 \times 98.8)}} = 0.966$$

9. Найдем значение  $t$ -критерия для оценки статистической значимости корреляционной связи:

$$t_r = \frac{r_{xy} \sqrt{n-2}}{\sqrt{1-r_{xy}^2}} = \frac{0.97 \sqrt{5-2}}{\sqrt{1-0.97^2}} = 7.0$$

Критическое значение  $t$ -критерия найдем по таблице, где при числе степеней свободы  $f = n-2 = 3$  и уровне значимости  $p = 0,01$  значение  $t_{крит} = 5,84$ . Рассчитанное значение  $t_r (7,0)$  больше  $t_{крит} (5.84)$ , следовательно, связь является статистически значимой.

10. Сделаем *статистический вывод*:

Значение коэффициента корреляции Пирсона составило 0,97, что соответствует весьма высокой тесноте связи между уровнем тестостерона в крови и процентом мышечной массы. Данная корреляционная связь является статистически значимой ( $p < 0,01$ ).

**14. Перечень вопросов для проверки исходного уровня знаний:**

8. Корреляция. Виды корреляции.
9. Связи. Типы связи. Сила связи.
10. В чем разница между общей и частной классификацией корреляции?

**15. Перечень вопросов для проверки конечного уровня знаний:**

8. Применение критерия Пирсона.
9. Применение критерия Спирмена.
10. Условия и ограничения применения критерия хи-квадрат Пирсона.
11. В каких случаях не рекомендуется применять критерий Спирмена?

**16. Хронокарта учебного занятия:**

21. Организационный момент – 10 мин.
22. Разбор темы – 40 мин.
23. Текущий контроль (тестирование, практическая работа) - 90 мин.
24. Подведение итогов занятия – 10 мин.

**17. Самостоятельная работа аспиранта.**

Применение линейной корреляции Пирсона для выявления связи между переменными (пакет SPSS и Statistica).

**18. Перечень учебной литературы к занятию:**

3. Есауленко И.Э., Семенов С.Н. Основы практической информатики в медицине; Воронеж, 2005.
4. Жижин К. С. Медицинская статистика; Ростов н/Д, 2007.

## **ТЕМА 8: Регрессионные коэффициенты t-критерию Стьюдента. Коэффициент множественной детерминации. Стандартный и пошаговый метод. Регрессионный анализ с помощью метода ввода.**

### **1. Научно-методическое обоснование темы:**

В зависимости от поставленной задачи, объема и характера материала, вида данных и их связей находится выбор методов математической обработки на этапах как предварительного (для оценки характера распределения в исследуемой выборке), так и окончательного анализа в соответствии с целями исследования. Крайне важным аспектом является проверка однородности выбранных групп наблюдения, в том числе контрольных, что может быть проведено или экспертным путем, или методами многомерной статистики (например, с помощью кластерного анализа). Первым этапом является составление вопросника, в котором предусматривается стандартизованное описание признаков. В особенности при проведении эпидемиологических исследований, где необходимо единство в понимании и описании одних и тех же симптомов разными врачами, включая учет диапазонов их изменений. В случае существенности различий в регистрации исходных данных (субъективная оценка характера патологических проявлений различными специалистами) и невозможности их приведения к единому виду на этапе сбора информации используются многомерный дисперсионный и регрессионный анализы, которые предполагают нормализацию переменных, т.е. устранение ненормальностей показателей в матрице данных.

### **2. Краткая теория:**

Методы регрессионного анализа позволяют по имеющимся данным предсказывать новые результаты, т.е. ориентированы на планирование и прогнозирование. Цель регрессионного анализа заключается в том, чтобы статистически адекватно связать «выходные», зависимые варианты с «входными» -независимыми.

Независимые переменные иногда называют предикторами, регрессорами, факторами, а зависимые – откликами.

Регрессия бывает линейной или нелинейной, простой, когда связаны не более двух признаков, или сложной (множественной), когда число связываемых анализом признаков больше, чем два.

Общий вид модели линейной множественной регрессии может быть задан следующим образом: предположим, что в выборке испытуемых есть независимые и зависимые переменные. Чтобы не усложнять обозначения, в модели линейной множественной регрессии предполагается, что значение отклика, принимаемые им на рассматриваемом множестве объектов, связаны со значениями предикторов на этих объектах с помощью системы линейных уравнений. В обобщенном виде этот процесс можно представить в виде одного-единственного уравнения регрессии, в котором подразумевается, что отклик и предикторы могут принимать значения на любом из рассматриваемых объектов: исследователя обычно интересует, насколько точны прогнозы, получаемые по построенной регрессии.

Стандартный вид уравнений регрессии получается в том случае, если и отклики, и предикторы представлены в стандартизованных «z-значениях», т.е. в значениях, находящихся в диапазоне от 0 до 1. При практической реализации регрессионного анализа, в том числе в статистических пакетах программ, понять, что уравнение регрессии записано в стандартизованном виде, можно, во -первых по наличию обозначений «бета» для коэффициентов регрессии, во-вторых, нередко используют «смешанную» форму уравнения: когда предикторы представлены z-значениями, а отклик – исходными, ненормированными значениями. В-третьих, часто под уравнением регрессии понимают прогностическое уравнение, т.е. уравнение, используемое для предсказания значений

отклика по известным значениям предикторов. Другими словами, на практике уравнение регрессии может быть записано в одной из множества форм. Такая ситуация требует от исследователя внимательности и эрудиции, чтобы по контексту определить, о какой именно форме уравнения регрессии идет речь в конкретном случае.

Независимо от конкретной формы используемого регрессионного уравнения результат регрессионного анализа оценивается по:

- 1) суммарному уровню взаимосвязи предикторов и отклика,
- 2) существенности вклада каждого предиктора в оценку отклика,
- 3) точности предсказания значений отклика и вероятностных ошибок их оценки.

Суммарный уровень взаимосвязи оценивается по величине коэффициентов множественной корреляции  $R$  или множественной детерминации  $R^2$ .

Коэффициент множественной детерминации является одним из основных показателей качества регрессии. Он принимает значения в диапазоне от 0 до 1, при этом, чем ближе его значение к 1, тем выше качество регрессии.

Коэффициент множественной корреляции равен квадратному корню из коэффициента множественной детерминации. Он также принимает значения в диапазоне от нуля до единицы, и чем ближе к 1, тем выше качество регрессии.

И чем ближе эти два показателя по своим абсолютным значениям, тем ближе линия регрессии к прямо пропорциональной или линейной зависимости между анализируемыми переменными, чем больше разница – тем более вероятна между ними криволинейная зависимость.

Обычно при оценивании качества регрессии с помощью F-критерия Фишера выполняется оценка уровня статистической значимости коэффициента множественной корреляции.

Т.о., один из основных критериев оценивания качества регрессии связан с суммарной величиной остатков: чем эта величина меньше, тем лучше регрессия описывает имеющиеся данные. При этом используется сумма квадратов остатков, так как сами по себе остатки могут иметь разные знаки и в силу этого взаимно «погашать» друг друга. В силу этого в регрессионном анализе часто применяют следующие вспомогательные показатели:

- сумма квадратов отклонений от среднего точных(измеренных) значений откликов;
- сумма квадратов отклонений, предсказанных (вычисленных с помощью регрессионного уравнения) значений откликов от среднего по всем предсказанным значениям;
- сумма квадратов остатков, т.е. разностей между точными и предсказанными значениями откликов.

Определение существенности вклада каждого предиктора в оценку отклика проводится с помощью регрессионных коэффициентов по t-критерию Стьюдента.

Мерилом точности предсказания значений отклика и вероятностных ошибок их оценки является значение коэффициента множественной детерминации.

Для корректного вывода при использовании регрессионного анализа требуется выполнение ряда условий:

- использование только количественных – интервальных шкал;
- распределение предикторов, отклика и остатков должно соответствовать нормальному закону;
- не должно быть взаимной коррелированности предикторов.

Регрессионный анализ включает в себя множество разнообразных методов, из которых на практике распространены стандартный и пошаговый.

Пошаговый метод применяется в одном из следующих двух вариантов:

Прямой – до максимально возможного количества предикторов, обеспечивающих статистически значимый коэффициент множественной корреляции;

Обратный – до минимального количества предикторов, также обеспечивающих статистически значимый коэффициент множественной корреляции.

Математическое моделирование, связанное с инструментальным (через меню) выполнением оптимизационных расчетов, например, при определении коэффициентов полинома в задаче прогнозирования графиков нагрузки), имеет тот недостаток, что при изменении исходных данных (ошибка, опечатка и др.) требуется повторная процедура оптимизации. Такие повторы не требуются, если имеется функция, ставящая в соответствие заданному набору данных требуемые результирующие величины. Задача прогнозирования нагрузки как случайной величины на основе предшествующих наблюдений относится к классу задач корреляции и регрессии, где производится оценка корреляционных и регрессионных характеристик по выборке. Для случайного вектора  $(t, P)$  выборка (объема  $n$ ) дает  $n$  пар значений признака  $(t_1, P_1), (t_2, P_2), \dots, (t_n, P_n)$ . В этом случае говорят о связанной выборке. Аппарат теорий корреляций и регрессий позволяет оценить характеристики этой связи. В *Excel* имеется набор статистических функций (**ЛИНЕЙН()**, **ПРЕДСКАЗ()**, **ТЕНДЕНЦИЯ()**, **НАКЛОН()**, **ЛГРФПРИБЛ()** и др. ), которые могут выполнить их расчет.

### Функция **ЛИНЕЙН**( $y$ ; $x$ ; конст; статистика)

Эта функция использует метод наименьших квадратов, чтобы вычислить параметры линейной зависимости, которая наилучшим образом аппроксимирует имеющиеся данные. Функция возвращает массив параметров, который описывает данную линейную зависимость. Уравнение для линейной функции имеет следующий вид:

$$y = mx + b,$$

где зависимое значение  $y$  является функцией независимого значения  $x$ .

При большом числе факторов  $y = m_1x_1 + m_2x_2 + \dots + b$  функция **ЛИНЕЙН()** возвращает массив  $\{m_n; m_{n-1}; \dots; m_1; b\}$ . Функция **ЛИНЕЙН()** может также возвращать дополнительную регрессионную статистику. В этом случае синтаксис функции - **ЛИНЕЙН**( $Y$ ;  $X$ ; конст; статистика),

где  $Y$ - это заданное множество значений  $y$ . Если массив  $Y$  имеет один столбец (одну строку), то каждый столбец (каждая строка) массива  $X$  интерпретируется как отдельная переменная;

$X$ - это необязательное множество значений параметра  $x$  (по умолчанию ряд  $\{x=1, 2, \dots\}$ ), которые соотносятся с  $y = mx + b$ . Массив  $X$  может содержать одно или несколько множеств (строк, столбцов) переменных. Массивы  $Y$  и  $X$  должны иметь одинаковую размерность;

$m_n$	$m_{n-1} \dots$	$m_1$	$b$
$se_n$	$se_{n-1} \dots$	$se_1$	$se_b$
$r^2$	$se_y$		
F	df		
$ss_{reg}$	$ss_{resid}$		

**конст**- это логическое значение, которое указывает, требуется ли, чтобы константа  $b$  была равна 0. Если **конст** имеет значение **ИСТИНА** или *опущено*, то  $b$  вычисляется обычным образом. Если **конст** имеет значение **ЛОЖЬ**, то  $b$  полагается равным 0 и значения  $m$  подбираются так, чтобы выполнялось соотношение  $y = mx$ ;

## Рис. Статистики

**статистика** - это логическое значение, которое указывает, требуется ли вернуть **дополнительную статистику по регрессии**. Если статистика имеет значение **ЛОЖЬ** или опущена, то функция **ЛИНЕЙН()** возвращает только коэффициенты  $m$  и постоянную  $b$ . Если статистика имеет значение **ИСТИНА**, то функция **ЛИНЕЙН()** возвращает дополнительную регрессионную статистику, так что возвращаемый массив будет иметь вид рис.3.21.

### Дополнительная регрессионная статистика

Параметры  $se_1, se_2, \dots, se_n, se_b$  - стандартные значения ошибок для коэффициентов  $m_1, m_2, \dots, m_n$  и для постоянной  $b$  ( $se_b = \#Н/Д$ , если параметр **конст** имеет значение **ЛОЖЬ**);

$r^2$  - коэффициент детерминированности (квадрат коэффициента корреляции). Сравниваются фактические значения  $y$  и значения, получаемые из уравнения прямой; по результатам сравнения вычисляется коэффициент детерминированности, нормированный от 0 до 1. Если он равен 1, то имеет место полная корреляция с моделью, т.е. нет различия между фактическим и оценочным значениями  $y$ ; если коэффициент детерминированности равен 0, то уравнение регрессии неудачно для предсказания значений  $y$ ;

$se_y$  - стандартная ошибка для оценки  $y$ ;

$F$ - $F$ -статистика, или  $F$ -наблюдаемое значение.  $F$ - статистика используется для определения того, является ли наблюдаемая взаимосвязь между зависимой и независимой переменными случайной или нет;

$df$  - степени свободы. Степени свободы полезны для нахождения  $F$ -критических значений (распределение Фишера) в статистической таблице. Для определения уровня надежности модели нужно сравнить значения в таблице с  $F$ -статистикой, возвращаемой функцией **ЛИНЕЙН()**;

$ss_{reg}$  - регрессионная сумма квадратов;

$ss_{resid}$  - остаточная сумма квадратов.

**Замечание.** Если имеется только одна независимая переменная  $x$ , можно получить крутизну и  $y$ -пересечение непосредственно, используя функции **НАКЛОН()** и **ИНДЕКС (ЛИНЕЙН( $Y;X$ ); 2)**

Точность аппроксимации с помощью прямой, вычисленной функцией **ЛИНЕЙН()**, зависит от степени разброса данных. Чем ближе данные к прямой, тем более точной является модель, используемая функцией **ЛИНЕЙН()**. Когда имеется только одна независимая переменная  $x$ ,  $m$  и  $b$  вычисляются по следующим формулам:

$$m = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}; \quad b = \frac{(\sum y)(\sum x)^2 - (\sum x)(\sum xy)}{n \sum x^2 - (\sum x)^2}.$$

Функции аппроксимации **ЛИНЕЙН()** и **ЛГРФПРИБЛ()** могут вычислить прямую или экспоненциальную кривую, наилучшим образом описывающую статистические данные. Однако вам самим предстоит решать, какой из двух результатов лучше. Можно также

вычислить функцию **ТЕНДЕНЦИЯ()** для прямой или функцию **РОСТ()** для экспоненциальной кривой. Эти функции, если не задавать аргумент  $X$ , возвращают массив вычисленных значений  $y$  для фактических значений  $x$  в соответствии с прямой или кривой. Теперь можно сравнить вычисленные значения с фактическими. Можно также построить диаграммы для визуального сравнения.

Проводя регрессионный анализ, Excel вычисляет для каждой точки квадрат разности между прогнозируемым и фактическим значениями  $y$ . Сумма квадратов этих разностей называется остаточной суммой квадратов. Затем Excel подсчитывает сумму квадратов разностей между фактическими и средним значениями  $y$ , которая называется общей суммой квадратов (регрессионная сумма квадратов + остаточная сумма квадратов). Чем меньше остаточная сумма квадратов по сравнению с общей суммой квадратов, тем больше значение коэффициента детерминированности  $r^2$ , который показывает, насколько хорошо уравнение, полученное с помощью регрессионного анализа, описывает взаимосвязи между переменными.

Следует заметить, что значения  $y$ , предсказанные с помощью уравнения регрессии, будут иметь меньшую точность, если они располагаются вне интервала значений  $y$ , который использовался для определения коэффициентов регрессии.

### Простая линейная регрессия

**Пример 1.** Предположим, что фирма по продаже электрооборудования за первые шесть месяцев отчетного года имела доход на сумму 3100 руб., 4500 руб., 4400 руб., 5400 руб., 7500 руб. и 8100 руб. Пусть эти значения находятся в интервале ячеек B2:B7. Тогда можно использовать следующую простую линейную регрессионную модель для оценки объема продаж *в девятом месяце*.

СУММ(ЛИНЕЙН(B2:B7)\*{9;1}) равняется СУММ({1000;2000}\*{9;1}) равняется 11 000 руб.

**Пример 2.** Рассмотрим представленный в табл.3.21 пример интерполяции временного ряда  $\{y_j\}$  функцией  $Y(x) = a_0 + a_1x$ . Парные наблюдения  $(x, y)$  записаны соответственно в столбцах A и B. В блоке G2:H6 получены результирующие параметры функции ЛИНЕЙН(B2:B13;A2:A13;;1)

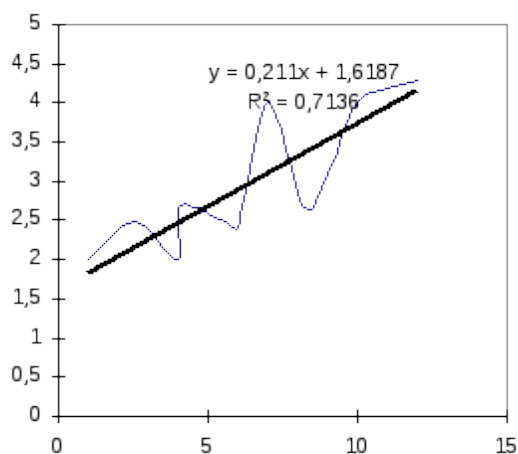


Рис. 3.70. Экспериментальная кривая и тренд

В результате  $Y(x) = 1,616 + 0,211x$ . Эта функция для разных  $x$  представлена в столбце С. В соседнем столбце вычислена поэлементная разность теоретических и экспериментальных значений  $y$ . В ячейке D14 вычисляется сумма квадратов отклонений. Нетрудно видеть, что она полностью совпадает с результирующим параметром  $S_{\text{res}}$  (H6). Стандартные ошибки для полученных коэффициентов позволяют оценить разброс  $y(x)$ , т.е. получить, например, по критерию  $2\sigma$  максимальную и минимальную оценку  $y(x)$ - столбцы I,J.

На рис. 3.22 представлена экспериментальная функциональная зависимость и средствами графики получено уравнение тренда. Нетрудно видеть, что коэффициенты тренда практически совпадают с параметрами  $m, b$  (этого следовало ожидать, поскольку использовалась одна и та же функция).

### Самостоятельная работа

- Повторить предложенные расчеты.
- Представить диаграмму, где были бы отражены минимальная и максимальная оценки линейной зависимости.
- Изменением в столбце В снизить максимальное отклонение эксперимента от тренда. Оценить изменение коэффициента детерминированности.

### Пример 3. Множественная линейная регрессия

Рассмотрим представленный в табл. 3.22. пример интерполяции временного ряда  $\{y_j\}$  функцией  $Y(t) = a_0 + a_1t + a_2t^2 + a_3 \cos(2\pi t/10)$ .

Таблица 3.31

	A	B	C	D	E	F	G	H	I	J
1	x	Y	Y_лин	Y-Y_лин					Y_min	Y_max
2	1	2	1,830	0,170		m,b	0,211	1,619	1,128	2,531
3	2,5	2,5	2,146	0,354		St.er	0,042	0,316	1,321	2,972
4	4	2	2,463	-0,463		R <sup>2</sup> ,se <sub>Y</sub>	0,714	0,484	1,513	3,413
5	4,1	2,7	2,484	0,216		F, df	24,919	10,000	1,526	3,442
6	5,5	2,5	2,779	-0,279		S, S <sub>res</sub>	5,843	2,345	1,705	3,853
7	6	2,4	2,885	-0,485					1,769	4,000
13	12	4,3	4,151	0,149					2,538	5,764
14			Суммкв()	2,345						

Таблица 3.32

	A	B	C	D	E	F	G	H	I
1									
2		x1	x2	x3		Из оптимизационных расчетов			



3	Yэ	t	t^2	cos(pi/5*t)		a <sub>3</sub>	a <sub>2</sub>	a <sub>1</sub>	a <sub>0</sub>
4	5	1	1	0,81		548,6	68,07	74,8	-582
5	5	1,1	1,21	0,77		<b>корреляционный анализ</b>			
6	4,8	1,2	1,44	0,73		554,81	68,86	75,53	-588,61
7	6,2	1,3	1,69	0,68		297,18	35,78	43,09	318,96
8	6,4	1,4	1,96	0,64		0,78	0,53	#Н/Д	#Н/Д
9	5,8	1,5	2,25	0,59		7,17	6	#Н/Д	#Н/Д
10	5,6	1,6	2,56	0,54		6,20	1,73	#Н/Д	#Н/Д
11	6	1,7	2,89	0,48					
12	6,4	1,8	3,24	0,43					
13	8	1,9	3,61	0,37					

Решение получим с помощью функции ЛИНЕЙН(). Для этого представим интерполирующую функцию в виде многопараметрической зависимости

$$Y(t) = a_0 + a_1x_1(t) + a_2x_2(t) + a_3x_3(t)$$

В ячейках F6:I10 записана формула =ЛИНЕЙН(A4:A13;B4:D13;;1).

### 3.8.5.Использование F-статистики

В предыдущем примере коэффициент детерминированности  $r^2$  равен 0,78 (ячейка F8 в результатах функции ЛИНЕЙН()), что указывает на некоторую, не очень сильную зависимость между независимыми переменными и функционалом. Можно использовать F-статистику, чтобы определить, является ли этот результат (с таким значением  $r^2$ ) случайным.

Предположим, что на самом деле нет взаимосвязи между переменными, а просто были выбраны те случайные данные, для которых статистический анализ вывел показанную взаимозависимость. Допустимую вероятность (уровень значимости) ошибки гипотезы о том, что имеется значимая взаимозависимость, принято обозначать величиной  $\alpha$ .

Для оценки гипотезы используется так называемый F-критерий, который служит для проверки гипотезы о равенстве дисперсий ( $\sigma_x = \sigma_y$ ), при условии, что X и Y распределены нормально. В общем случае из каждой генеральной совокупности производятся выборки объемом  $n_1$  и  $n_2$ . В качестве контрольной величины используется отношение эмпирических дисперсий  $F = S_x^2 / S_y^2$ . Величина F удовлетворяет F-распределению (распределение Фишера) с  $v_1$  и  $v_2$  степенями свободы ( $v_1 = n_1 - 1, v_2 = n_2 - 1$ ). В рассматриваемом случае в качестве сопоставляемых величин являются функционал и вектор параметров.

Если F-наблюдаемое (ячейка F9) больше, чем F-критическое, то взаимосвязь между переменными и функционалом (трендом) является значимой. Величину F-критическое можно получить из таблицы F-критических значений в любом справочнике по математической статистике. Для того чтобы найти это значение, необходимо иметь уровень значимости  $\alpha = 0,05$  и значения степеней свободы  $v_1$  и  $v_2$ , где  $v_1 = k$  - это число переменных,  $v_2 = n - (k + 1)$ , а n - число статистических данных. В нашем случае  $v_1 = 3, v_2 = 10 - (3 + 1) = 6$  (в ячейке G9).

Из таблицы справочника  $F$ -критическое равно 4,76. Наблюдаемое  $F$ -значение равно 7,17, что больше 4,76. Следовательно, гипотеза о взаимосвязи линейной корреляции функционала и переменных не отвергается, и полученное регрессионное уравнение может быть использовано для прогнозирования нагрузки.

Было бы неразумно окружать свое рабочее место толстыми математическими справочниками. Не поможет ли нам Excel получить  $F$ -критическое? Нет проблем. Для этого имеется встроенная функция **FRASПОБР()**.

**Функция FRASПОБР( $\alpha$ ;  $\nu_1$ ;  $\nu_2$ )**

возвращает обратное значение для  $F$ -распределения вероятностей (функция **FRASП( $x$ ;...)**). Если  $\alpha = \text{FRASП}(x;...)$ , то  $x = \text{FRASПОБР}(\alpha;...)$ . Параметры  $\nu_1$ ,  $\nu_2$  - это числа степеней свободы.

#### Замечания

- Если число степеней свободы не целое число, то оно усекается до целой.
- Функция возвращает значение ошибки **#ЧИСЛО!**, если вероятность  $\alpha < 0$  или  $\alpha > 1$ , или  $\nu_1, \nu_2 < 1$ , или  $\nu_1, \nu_2 > 10^{10}$ .

**Пример.** **FRASПОБР(0,05;3;6)** равняется 4,757.

#### 3.8.6. Вычисление $t$ -статистики

Другой гипотетический эксперимент определит, полезен ли каждый коэффициент наклона для оценки тренда мощности (см. табл. 3.22). Например, для проверки того, имеет ли статистическую значимость циклическая составляющая, разделим 554,81 (коэффициент пропорциональности при  $\cos(\pi/5 \cdot t)$ ) на 297,18 (оценка стандартной ошибки для коэффициента времени эксплуатации):  $t = m_3/se_3 = 554,81/297,18 = 1,846$ . Эта величина сопоставляется с  $t$ -критерием (распределение Стьюдента), который служит для сравнения двух средних значений из нормально распределенных генеральных совокупностей случайных величин в предположении, что равны их дисперсии.

Если посмотреть в таблицу справочника по математической статистике, то окажется, что  $t$ -критическое с шестью степенями свободы и  $\alpha = 0,1$  равно 1,94. Поскольку абсолютная величина  $t$ , равная 1,846, меньше, чем 1,94, то можно сделать вывод о том, что циклическая составляющая - это незначимая переменная для оценки тренда мощности. Аналогичным образом можно протестировать на статистическую значимость все другие переменные.

Как и для  $F$ -распределения, Excel имеет возможность вычислить  $t$ -критическое с помощью встроенной функции **СТЮДРАСПОБР()**.

**Функция СТЮДРАСПОБР(вероятность;  $df$ )** возвращает обратное распределение Стьюдента для заданных вероятности, соответствующей двустороннему распределению Стьюдента, и числа степеней свободы  $df$ .

#### Замечания

- Если  $df$  не целое, то оно усекается до целой.

- Если *вероятность* меньше нуля или больше единицы, или число степеней свободы меньше единицы, то функция **СТЮДРАСПОБР()** возвращает значение ошибки #ЧИСЛО!.
- Рассматриваемая функция использует итерационный метод для вычисления возвращаемого значения. Если итерационный процесс не сходится за 100 итераций, то функция возвращает значение ошибки #Н/Д.

**Пример.** **СТЮДРАСПОБР(0,1;6)** равняется 1,94.

*Назначение F- критерия Фишера.* Проверка гипотезы о принадлежности двух дисперсий одной генеральной совокупности и следовательно — их равенстве.

*Нулевая гипотеза.*  $S_2^2 = S_1^2$

*Альтернативная гипотеза.* Существуют следующие варианты  $H_A$  в зависимости от которых различаются критические области:

1.  $S_1^2 > S_2^2$ . Наиболее часто используемый вариант  $H_A$ . Критическая область — верхний хвост F-распределения.
2.  $S_1^2 < S_2^2$ . Критическая область — нижний хвост F-распределения. Ввиду частого отсутствия нижнего хвоста, в таблицах критическую область обычно сводят к варианту 1, меняя местами дисперсии.
3. Двухсторонняя  $S_1^2 \neq S_2^2$ . Комбинация первых двух.

*Предпосылки.* Данные независимы и распределены по нормальному закону. Гипотеза о равенстве дисперсий двух нормальных генеральных совокупностей принимается, если отношение большей дисперсии к меньшей меньше критического значения распределения Фишера.

$$F_p = S_1^2 / S_2^2$$

*Примечание.* При описываемом способе проверки значение  $F_{расч}$  обязательно должно быть больше единицы. Критерий чувствителен к нарушению предположения о нормальности.

Для двухсторонней альтернативы  $S_1^2 \neq S_2^2$  нулевая гипотеза принимается при выполнении условия:

$$F_{1-\alpha/2} < F_{расч} < F_{\alpha/2}$$

**Пример.**

Комплексным теплотметрическим методом определяли теплофизические. характеристики (ТФХ) зеленого солода. Для приготовления образцов брали воздушно-сухой (средняя влажность  $W=19\%$ ) и влажный солод четырехсуточного ращения ( $W=45\%$ ) в соответствии новой технологией приготовления карамельного солода. Опыты показали, что теплопроводность  $\lambda$  влажного солода примерно в 2,5 раза больше, чем сухого, а объемная теплоемкость не имеет четкой зависимости от влажности солода. Поэтому с помощью F-критерия проверили возможность обобщить данные по средним значениям без учета влажности

Расчетные данные сведены в таблицу 5.1

Таблица 5.1

Данные к расчету F-критерия

W, %	19					45					
t, °C	30,6	34,2	38,9	44,2	49,1	29,1	36,5	41,5	47,2	54,3	61,8
y, МДж/(м³·К)	0,92	1,32	1,31	1,62	1,06	1,28	1,71	1,72	1,53	1,25	2,07
$\bar{y}$	1,27	1,31	1,36	1,41	1,46	1,26	1,34	1,39	1,44	1,51	1,59
$ y - \bar{y} $	0,35	0,01	0,05	0,21	0,40	0,02	0,37	0,33	0,09	0,26	0,48
$(y - \bar{y})^2$	0,12	0,00	0,00	0,04	0,16	0,00	0,13	0,11	0,01	0,06	0,23
$S^2$	0,080					0,108					

Большее значение дисперсии получено для W=45%, т.е.  $S^2_{45} = S^2_{19} = S^2_2$ , и  $F_p = S^2_1/S^2_2 = 1,35$ . Из таблицы 5.2 для степени свободы  $f_1 = N_1 - 1 = 5$   $f_2 = N_2 - 1 = 4$  при  $\gamma = 0,95$  определяем  $F_{кр} = 6,2$ . Нуль гипотеза сформулированная как «В диапазоне влажности зеленого солода от 19 до 45% ее влиянием на объемную теплоемкость можно пренебречь» или « $S^2_{45} = S^2_{19}$ » с доверительной вероятностью 95% подтвердилась, поскольку  $F_p < F_{кр}$ .

### Пример проверки гипотезы о принадлежности двух дисперсий одной генеральной совокупности по критерию Фишера с помощью Excel

Приведены данные по двум независимым выборкам (табл. 5.2) степени водопоглощения зерна пшеницы. Было проведено исследование воздействия магнитными полями низкой частоты.

Таблица 5.2

Результаты исследований

Номер опыта	Номер выборки	
	1	2
1	0,027	0,075
2	0,036	0,4
3	0,1	0,08
4	0,12	0,105
5	0,32	0,075
6	0,45	0,12
7	0,049	0,06
8	0,105	0,075

Прежде, чем мы будем проверять гипотезу о равенстве средних этих выборок, необходимо проверить гипотезу о равенстве дисперсий, чтобы знать какой из критериев выбрать для ее проверки.

На рис. 5.1 приведен пример проверки гипотезы о принадлежности двух дисперсий одной генеральной совокупности по критерию Фишера используя программный продукт Microsoft Excel.

	F15		=FPACП(F13;7;7)			
	A	B	C	D	E	F
1		№ опыта	Номер выборки			
2			1	2		
3		1	0,027	0,075		
4		2	0,036	0,4		
5		3	0,1	0,08		
6		4	0,12	0,105		
7		5	0,32	0,075		
8		6	0,45	0,12		
9		7	0,049	0,06		
10		8	0,105	0,075		
11	Дисперсии		0,02323	0,01283		
12						
13	Расчетное критерия Фишера				1,81144	
14	Критическое значение для критерия Фишера				3,78705	
15	Расчетный уровень значимости				0,22566	
16						
17						

Рисунок 5.1 Пример проверки принадлежности двух дисперсий одной генеральной совокупности по критерию Фишера

Исходные данные размещены в ячейках, находящихся на пересечении столбцов C и D со строками 3-10. Выполним следующие действия.

1. Определим, можно ли считать закон распределения первой и второй выборок нормальным (столбцы C и D соответственно). Если нет (хотя бы для одной выборки), то необходимо использовать непараметрический критерий, если да – продолжаем.
2. Рассчитаем дисперсии для первого и второго столбца. Для этого в ячейках C11 и D11 поместим функции =ДИСП(C3:C10) и =ДИСП(D3:D10) соответственно. Результатом работы этих функций является рассчитанное значение дисперсии для каждого столбца соответственно.
3. Находим расчетное значение для критерия Фишера. Для этого нужно большую дисперсию разделить на меньшую. В ячейку F13 помещаем формулу =C11/D11, которая и выполняет эту операцию.
4. Определяем, можно ли принять гипотезу о равенстве дисперсий. Существует два способа, которые представлены в примере. По первому способу, задавшись уровнем значимости, например 0,05, вычисляют критическое значение распределения Фишера для этого значения и соответствующего числа степеней свободы. В ячейку F14 вводится функция =FPACПОВР(0,05;7;7) (где 0,05 - заданный уровень значимости; 7 — число степеней свободы числителя, а 7 (второе) — число степеней свободы знаменателя). Число степеней свободы равно числу экспериментов минус единица. Результат — 3,787051. Поскольку это значение больше расчетного 1,81144, мы должны принять нулевую гипотезу о равенстве дисперсий.

По второму варианту рассчитывают для полученного расчетного значения критерия Фишера соответствующую вероятность. Для этого в ячейку F15 вводится функция =FPACП(F13;7;7). Поскольку полученное значение 0,22566 больше, чем 0,05, то принимается гипотеза о равенстве дисперсий.

Это может быть выполнено специальной функцией. Выберите на ленте вкладку *Данные-Анализ данных*. Появится окно следующего вида (рис. 5.2).

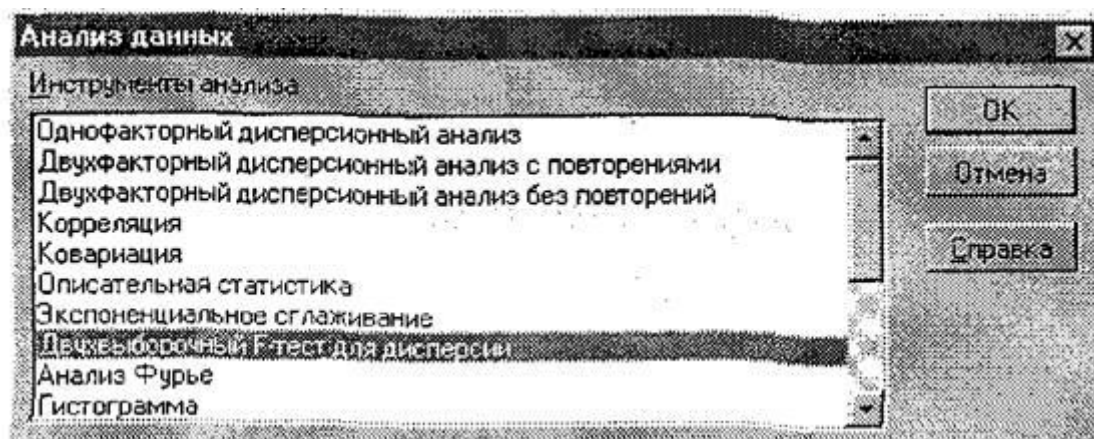


Рисунок 5.2 Окно выбора метода обработки

В этом окне выбираете «*Двухвыборочный F-тест для дисперсий*». В результате появится окно вида, показанного на рис. 5.3. Здесь задаются интервалы (номера ячеек) первой и второй переменной, уровень значимости (альфа) и место, где будет находиться результат. Задавайте все необходимые параметры и нажимайте ОК. Результат работы приведен на рис. 5.4

Следует отметить, что функция проверяет односторонний критерий и делает это правильно. Для случая, когда критериальное значение больше 1, вычисляется верхнее критическое значение.

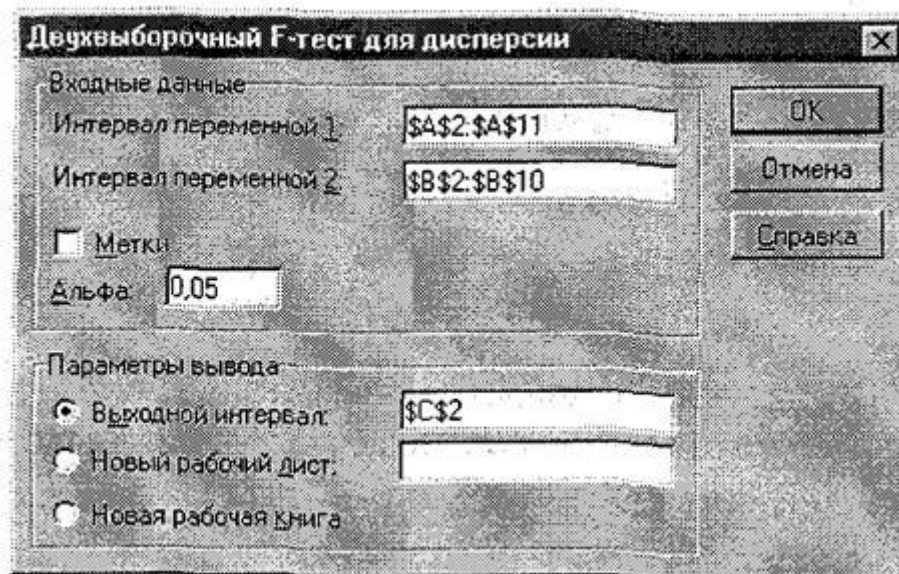


Рисунок 5.3 Окно задания параметров

Когда критериальное значение меньше 1, то вычисляется нижнее критическое.

Напоминаем, что гипотеза о равенстве дисперсий отвергается, если критериальное значение больше верхнего критического или меньше нижнего.

	A	B	C	D	E	F	G	H	I	J
1	Группа 1	Группа 3								
2	1,85	2,27	Двухвыборочный F-тест для дисперсии							
3	1,87	2,09								
4	1,87	2,09		Переменная 1	Переменная 2					
5	2,3	2,41	Среднее	2,001	2,162222222					
6	2,52	2,31	Дисперсия	0,083254444	0,019469444					
7	1,89	2,17	Наблюдения	10	9					
8	2,37	2	df	9	8					
9	1,7	2,1	F	4,276159224						
10	1,7	2,02	P(F<=f) одностороннее	0,026382941						
11	1,94		F критическое одностороннее	3,388123559						
12										
13										
14										

Рисунок 5.4 Проверка равенства дисперсий

### Формулы, используемые при регрессионном анализе

Формула для вычислений	Функция или инструмент Анализа данных в Excel	Результат вычислений /Примечания
<p><b>Оценка параметров модели парной и множественной линейной регрессии</b></p> $A = (X^T X)^{-1} X^T Y$	<p>Для вычисления параметров уравнения регрессии следует воспользоваться инструментом <b>Регрессия</b> из пакета <b>Анализ данных</b></p>	<p>Возвращает подробную информацию о параметрах модели, качестве модели, расчетных значениях и остатках в виде четырех таблиц: <i>Регрессионная статистика</i>, <i>Дисперсионный анализ</i>, <i>Коэффициенты</i>, <i>Вывод остатка</i>.</p> <p>Также могут быть получены <i>график подбора</i> и <i>график остатков</i></p>
Оценка качества модели регрессии		
<p><b>F-критерий Фишера для проверки значимости модели регрессии</b></p> $F = \frac{R^2 / k}{(1 - R^2) / (n - k - 1)}$	<p><b>=FРАСПОБР (вероятность; степени_свободы1; степени_свободы2)</b></p> <p><b>вероятность</b> – это вероятность, связанная с F-распределением</p> <p><b>степени_свободы 1</b> – это числитель степеней свободы (<math>v_1 = k</math>)</p> <p><b>степени_свободы 2</b> – это знаменатель степеней свободы (<math>v_2 = (n - k - 1)</math>),</p>	<p>Возвращает обратное значение для F-распределения вероятностей.</p> <p>FРАСПОБР( ) можно использовать, чтобы определить критические значения F-</p>

	где $k$ – количество факторов, включенных в модель)	распределения.  Чтобы определить критическое значение $F$ , нужно использовать уровень значимости $\alpha$ как аргумент <b>вероятность</b> для ФРАСПОБР( ).
--	---	---

<p><b>Коэффициент детерминации</b></p> $R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$	<p><b>Коэффициент детерминации</b> показывает долю вариации результирующего признака, находящегося под воздействием изучаемых факторов, то есть определяет, какая доля вариации признака <math>Y</math> учтена в модели и обусловлена влиянием на него факторов.</p> <p>Чем ближе <math>R^2</math> к 1, тем выше качество модели</p>
--	--

<p><b>Коэффициент множественной корреляции</b> (индекс корреляции) <math>R</math></p>	<p>Данный коэффициент является универсальным, так как он отражает тесноту связи и точность модели, а также может использоваться при любой форме связи переменных.</p> <p>Чем ближе <math>R</math> к 1, тем выше качество модели</p>
---	---

<p><b><math>t</math>-критерий Стьюдента</b> для оценки значимости параметров модели линейной регрессии:</p> $t_{aj} = \hat{a}_j / S_{aj}$	<p>Вычисленное значение <math>t_{aj}</math> сравнивается с критическим значением <math>t</math>-критерия, которое берется из таблицы значений <math>t</math>-распределения Стьюдента с учетом заданного уровня значимости и числа степеней свободы <math>(n - k - 1)</math>. В Excel критическое значение <math>t</math>-критерия можно получить с помощью функции</p> <p><b>СТЮДРАСПОБР</b> (вероятность; степени _свободы)</p> <p><b>вероятность</b> – вероятность, соответствующая двустороннему</p>
---	---



	<p>распределению Стьюдента</p> <p><b>степени _свободы</b> – число степеней свободы, характеризующее распределение</p>
<p><b>Средняя относительная ошибка аппроксимации</b></p> $E_{\text{отн}} = \frac{1}{n} \sum_{i=1}^n \frac{ e_i }{y_i} \cdot 100\%$	<p>Средняя относительная ошибка аппроксимации – оценка точности модели</p>
<p><b>Оценка влияния отдельных факторов на зависимую переменную на основе модели</b></p>	
<p><b>Коэффициенты эластичности</b></p> $\varepsilon_j = a_j \cdot \frac{\bar{x}_j}{\bar{y}}$	<p>Коэффициент эластичности показывает, на сколько процентов изменится значение исследуемой величины при изменении соответствующего фактора на 1%</p>
<p><b>Бета-коэффициенты</b></p> $b_j = \hat{a}_j \cdot \frac{s_{x_j}}{s_y}$	<p>Бета-коэффициент показывает, на какую часть своего СКО изменится значение исследуемой переменной при изменении соответствующего фактора на 1 СКО</p>
<p><b>Дельта-коэффициенты</b></p> $\Delta_j = r_{y,x_j} \cdot \beta_j / R^2$	<p>Дельта-коэффициент показывает среднюю долю влияния соответствующего фактора в совокупном влиянии всех факторов, включенных в модель</p>

## 6. Цель деятельности аспирантов на занятии:

### Аспирант должен знать:

25. Регрессия. Виды регрессии.
26. Понятия предиктора, отклика.
27. Уравнение регрессии.
28. Коэффициент множественной детерминации. Коэффициент множественной корреляции.
29. F- критерий Фишера.
30. Регрессионный анализ.

### Аспирант должен уметь:

3. Применять регрессионный коэффициент по t-критерию Стьюдента для определения вклада предиктора в оценку отклика.
4. Применять коэффициент множественной детерминации для измерения точности предсказания значений отклика и вероятных ошибок.
5. Проводить регрессионный анализ с помощью пакетов анализа.

### Содержание обучения:

#### Теоретическая часть:

8. Общий вид модели линейной множественной регрессии.
9. Стандартный вид уравнений регрессии.
10. Основные показатели качества регрессии. Их значения и соотношения.
11. Оценка качества регрессии с помощью F-критерия Фишера.
12. Регрессионный анализ при помощи пакета анализа EXCEL, с помощью метода ввода в пакете SPSS.

**Практическая часть:**

*Условие.* В шести кабинетах производственного обучения межшкольного учебно-производственного комбината, в мастерской сош и пту изучалось влияние шума на организм учащихся.

*Найти* регрессионную связь уровней шума на рабочих местах с октавными частотами (Var1-63-Var8-8000Гц) и уровнем звука в ДБА(Var9). (использовать метод ввода).

**19. Перечень вопросов для проверки исходного уровня знаний:**

11. Регрессия. Виды регрессии.
12. Понятия предиктора, отклика.
13. Уравнение регрессии.
14. Коэффициент множественной детерминации.  
Коэффициент множественной корреляции.

**20. Перечень вопросов для проверки конечного уровня знаний:**

12. F- критерий Фишера для оценки уровня статистической значимости коэффициента множественной корреляции.
13. Проведение регрессионного анализа с помощью регрессионных коэффициентов по t-критерию Стьюдента и с помощью метода ввода.

**21. Хронокарта учебного занятия:**

1. Организационный момент – 10 мин.
2. Разбор темы – 40 мин.
3. Текущий контроль (тестирование, практическая работа) - 90 мин.
4. Подведение итогов занятия – 10 мин.

**22. Самостоятельная работа аспиранта.**

Применение регрессионного анализа для прогнозирования результатов (пакет Statistica).

**23. Перечень учебной литературы к занятию:**

2. Есауленко И.Э., Семенов С.Н. Основы практической информатики в медицине; Воронеж, 2005.
3. Жижин К. С. Медицинская статистика; Ростов н/Д, 2007.

## ТЕМА 9: «Дисперсионный анализ. ОДА. ДДА.»

### 1. Научно-методическое обоснование темы:

Дисперсионный анализ - это метод математической статистики, предназначенный для моделирования количественного выходного параметра-отклика на воздействующие входные факторы, уровни которых оцениваются качественно, по номинальной шкале. Например, фактор А - тяжесть заболевания на трех уровнях: (легкая, средняя, тяжелая).

В зависимости от количества контролируемых факторов различают одно- двух- и многофакторный дисперсионный анализ (ДА). Каждый контролируемый фактор может фиксироваться на двух, трех и более уровнях. Моделируемый параметр-отклик на воздействующие факторы оценивается количественно по интервальной или порядковой шкале для каждого сочетания уровней факторов.

Сущность ДА заключается в разложении дисперсии параметра  $Y$  на составляющие:

- дисперсию вследствие влияния контролируемых факторов;
- дисперсию, вызываемую действием неконтролируемых, случайных факторов и ошибками измерения.

По доле дисперсии, обусловленной контролируемыми факторами, определяется степень и значимость влияния входных факторов на параметр  $Y$ .

По средним значениям параметра  $Y$  на различных уровнях факторов изучается характер изменения параметра при изменении уровней воздействующих факторов, дается прогноз ожидаемых значений параметра при заданных уровнях факторов.

По результатам моделирования множества выходных параметров дается оценка их информативности на воздействующие факторы, что имеет большое значение при оценке весомости параметров и выработке комплексной интегральной оценки состояния объекта исследования.

### 2. Краткая теория:

Часто применяемым методом проверки выборок на однородность и поиска причинно-следственных связей является дисперсионный анализ, разработанный Р. Фишером. Существует несколько вариантов этого вида статистической обработки экспериментальных данных. Наиболее актуальны из них следующие:

1. Однофакторный, или одномерный, дисперсионный анализ (ДА по одному признаку), который называется ANOVA либо однофакторный дисперсионный анализ.
2. Многофакторный, или многовариантный, дисперсионный анализ по нескольким признакам (MANOVA).

Сущность первого из них в отыскании причинно-следственных связей при воздействии одного (вид ANOVA) или группы факторов (вид MANOVA). В роли факторных нагрузок могут выступать различные условия проведения измерений: временные, ситуационные, психологические и др. *Для выявления влияния надо*

**располагать результатами измерений, соответствующими не менее чем трем уровням фактора.** К примеру, ANOVA применяется для анализа не менее трех выборок и основан на сравнении их дисперсий.  
ОДА.

Это классический ОДА, он параметрический и предполагает, что при расчете так называемого F-критерия Фишера выборки взяты из генеральных совокупностей, распределенных по нормальному закону. В медицине, биологии это условие нарушается, что послужило толчком для разработки непараметрических аналогов ОДА: для несвязанных выборок – критерии Краскела - Уоллиса и Джонкира, для связанных – критерии Фридмана и Пейджа

Не стоит только думать, что применение классического ОДА с использованием ПК избавляет исследователя от четкого продумывания сути эксперимента или тщательного подбора анализируемого материала. Данный вид статистической обработки данных всего лишь подтверждает или отвергает концепцию, рожденную исследователем за письменным столом. Однако он существенно отличается от корреляционного анализа уже тем, что здесь мы можем дать оценку, выраженную в цифрах, причинно-следственным связям между анализируемыми признаками.

Обязательное условие при использовании ANOVA/MANOVA:

- Перед проведением аналитической работы проверить – соблюдается ли условие нормальности и данные представляют собой случайные выборки из нормально распределенных генеральных совокупностей;

- Также тщательно проверить, соблюдается ли условие однородности (гомогенности) дисперсий: имеют ли выборки равные дисперсии;

- Убедиться в том, что выборки независимы, т.е. нельзя априори предсказать значение какого-либо наблюдения по значению другого.

ОДА дает корректные результаты при нарушении однородности дисперсий в том случае, если уравнены объемы выборок или отличие их будет очень незначительным. Если сформировать выборки большого объема, то и первое и второе допущение можно перекрыть.

Нулевая гипотеза ОДА свидетельствует о равенстве средних величин у рассматриваемых совокупностей; соответственно альтернативная гипотеза отвергает значимые отличия в средних, обусловленные воздействием рассматриваемого фактора.

F-критерий Фишера рассчитывается по следующей формуле:

$$F = \frac{\sigma_{bg}^2}{\sigma_{wg}^2}$$

Эта формула выражает отношение двух дисперсий: межгрупповой (она в числителе дроби) и внутригрупповой (в знаменателе дроби). Как правило, внутригрупповая дисперсия обусловлена случайными причинами, а воздействие фактора проявляется в наличии межгрупповой дисперсии. Особую роль при применении ОДА играет сумма квадратов отклонений SS, так как с нее начинается расчет дисперсий, входящих в приведенную формулу. Каждая из этих дисперсий вычисляется как отношение соответствующей суммы квадратов отклонений к количеству степеней свободы:

$$S_{bg}^2 = \frac{SS_{bg}}{df_{bg}},$$

$$S_{wg}^2 = \frac{SS_{wg}}{df_{wg}},$$

где SS – сумма квадратов отклонений, соответствующая внутри – и межгрупповой дисперсии;

$df_{bg}$  ( или  $k - 1$  ) - число степеней свободы межгрупповой дисперсии;  
 $df_{wg}$  ( или  $N - k$  ) - число степеней свободы внутригрупповой дисперсии;  
 $k$  - количество градаций (уровней) фактора, соответствующее числу выборок;  
 $N$  - общее число наблюдений в выборках.

Дисперсионный анализ: однофакторный.

Однофакторный дисперсионный анализ (*ANOVA – analysis of variance*) используется для сравнения средних значений для трех и более выборок. *Фактором* называется *независимая* переменная, влияние которой изучается на *зависимую* переменную.

Например, фактором может быть уровень образования, вид деятельности, возрастная группа респондентов, степень лояльности к торговой марке и т.д.

Анализ основан на расчете ***F – статистики*** (статистика Фишера), которая представляет собой отношение двух *дисперсий*: межгрупповой и внутригрупповой. *F – тест* в однофакторном дисперсионном анализе устанавливает, значимо ли отличаются средние нескольких независимых выборок. Он заменяет *t – тест* для независимых выборок при наличии более двух выборок и дает тот же результат в случае двух выборок.

Процедура выполнения однофакторного дисперсионного анализа:

1. Определение независимых и зависимых переменных
2. Разложение полной дисперсии ( $SS$ )
3. Измерение эффекта ( $\eta^2$ )
4. Проверка значимости ( $F$ )
5. Представление результата

**Необходимым условием для проведения дисперсионного анализа является то, чтобы независимая переменная была категориальной, а зависимая – метрической.**

Набор данных в *ANOVA* состоит из ***k*** – независимых одномерных выборок, элементы которых измерены в одинаковых единицах (долл., кг., баллы). Допустимы различные объемы (размеры) выборок.

1 этап. Подготовка данных для анализа выглядит следующим образом:

	Независимая переменная – фактор (напр., вид деятельности) (количество выборок $k = 4$ )			
	Выборка 1 – (экономисты)	Выборка 2 – (медики)	Выборка 3 – (биологи)	Выборка k – (химики)
Зависимая:	$X_{1,1}$	$X_{2,1}$	$X_{3,1}$	$X_{k,1}$
Зависимая:	$X_{1,2}$	$X_{2,2}$	$X_{3,2}$	$X_{k,2}$
Зависимая:	$X_{1,3}$	$X_{2,3}$	$X_{3,3}$	$X_{k,3}$
Зависимая:	$X_{1,4}$	$X_{2,4}$	$X_{3,4}$	$X_{k,4}$
Зависимая:	$X_{1,5}$	$X_{2,5}$		$X_{k,5}$
Зависимая:		$X_{2,6}$		$X_{k,6}$
Зависимая:		$X_{2,7}$		
Объем $n=n_1+n_2+n_3+...+n_k$	$n_1 = 5$	$n_2 = 7$	$n_3 = 4$	$n_k = 6$
Среднее	$X_1$	$X_2$	$X_3$	$X_k$
Ст. отклонение	$\sigma_1$	$\sigma_2$	$\sigma_3$	$\sigma_k$

Нулевая гипотеза в однофакторном дисперсионном анализе утверждает, что все средние значения из различных генеральных совокупностей (которые представлены выборочными средними) равны между собой.

$H_0 : \mu_1 = \mu_k$  (все равны). (или  $X_1 = X_2 = \dots = X_k$ )

Альтернативная гипотеза утверждает, что хотя бы два любых средних не равны между собой.

$H_1 : \mu_1 \neq \mu_k$  (хотя бы две не равны). (или  $X_1 \neq X_k$ )

F – тест состоит в расчете F – статистики и сравнении ее с табличным значением (аналогично с t – тестом).

Поскольку нулевая гипотеза утверждает, что средние всех генеральных совокупностей равны, необходимо оценить это среднее значение по всем выборкам, т.е. рассчитать *общее среднее*. **Общее среднее представляет собой среднее всех значений из всех выборок.**

Если размеры выборок не равны, то среднее рассчитывается как средневзвешенное с учетом размера выборок:

2 этап. Для изучения различий между зависимыми переменными проводится разложение полной дисперсии:

$$SS = SS_{between} + SS_{within},$$

где  $SS_{between}$  – межгрупповая вариация и  $SS_{within}$  – внутригрупповая вариация.

**Межгрупповая вариация ( $SS_{between}$ )** показывает, насколько выборочные средние отличаются между собой. Она равна нулю, если средние равны и тем больше, чем сильнее различаются средние. Расчет межгрупповой дисперсии (вариации):

$$SS_{between} = \sum_{i=1}^k n_i * (X_i - X)^2 \quad \text{и средний квадрат:}$$

$$MS_b = \frac{SS_{between}}{k - 1}$$

**Внутригрупповая вариация ( $SS_{within}$ )** показывает, насколько отличаются между собой значения по каждой выборке.

$$SS_{within} = \sum_{i=1}^k (n_i - 1) * \sigma_i^2 \quad \text{и средний квадрат:}$$

$$MS_w = \frac{SS_{within}}{n - k}$$

3 этап. Эффект влияния независимой переменной на зависимую переменную рассчитывается через корреляционное отношение  $\eta^2$  (эта-квадрат), которое рассчитывается по формуле:

$$\eta^2 = \frac{SS_{between}}{SS}$$

Значение корреляционного отношения находится в пределах от 0 до 1. Оно равно 0, когда все выборочные средние равны, т.е. независимая переменная не влияет на зависимую, и, наоборот, влияние увеличивается с ростом этого значения. Другими словами, величина  $\eta^2$  представляет собой меру вариации зависимой переменной, вызванную влиянием на нее независимой переменной.

4 этап фактически сводится к процедуре статистической проверки гипотезы о равенстве средних (наличии различий) путем расчета  $F$  – статистики:

$$F = \frac{MS_{between}}{MS_{within}}$$

5 этап. Для того, чтобы сделать окончательный вывод, необходимо обратиться к  $F$  – таблице, содержащей критические значения  $F$  - статистики при истинной нулевой гипотезе. Чтобы найти критическое значение, необходимо учесть количество степеней свободы ( $df$  – degree freedom) и соответствующий уровень проверки (по умолчанию 5%).

Степень свободы для межгрупповой вариации составляет « $k - 1$ », а для внутригрупповой вариации « $n - k$ ».

$F$  – тест заключается в сравнении  $F$  – статистики, рассчитанной по имеющимся данным с критическим значением  $F$  – таблицы. Результат является значимым, если  $F_{\text{стат}} > F_{\text{критич}}$ , поскольку это говорит о наличии существенных различий между средними значениями по группам.

**Результат вычислений с помощью Excel: однофакторная ANOVA – таблица.**

**Пример 1.** Поставки лекарственных препаратов для вашей аптеки осуществляются тремя поставщиками («Мега+», «Коста» и «Трамп») в разное время: дневные часы, ночные смены и даже в пересменки. Естественно, с вашей стороны, контроль за качеством продукции в дневное время выше, чем в другое время. Вами собраны данные с оценками качества (в баллах), и вы стремитесь узнать, есть ли отличие в качестве продукции, которая поставляется в разное время?

	Дневная смена	Ночная смена	Пересменка
«Мега+»	77,06	93,12	77,05
«Коста»	81,14	88,13	78,11
«Трамп»	82,02	81,18	79,91

Однофакторный дисперсионный анализ						
ИТОГИ						

Группы	Счет	Сумма	Среднее	Дисперсия		
Столбец 1	3	240,2	<b>80,0733333</b>	7,00373333		
Столбец 2	3	262,4	<b>87,4766666</b>	35,9610333		
Столбец 3	3	235,0	<b>78,3566666</b>	2,09053333		
Дисперсионный анализ						
Источник вариации	SS	df	MS	F	P-Значение	F-критич.
Между группами	140,930688	2	70,4653444	<b>4,69192377</b>	<b>0,05932788</b>	<b>5,143252</b>
Внутри групп	90,1106	6	15,0184333			
Итого	231,041288	8				

Результаты расчета показывают, что  $F_{\text{стат}} < F_{\text{критич}}$  ( $4,691 < 5,14$ ), следовательно, отличие в качестве поставляемых лекарственных средств в разное время отсутствует. Кроме того, *P-значение* (вероятность истинности нулевой гипотезы о равенстве средних) превышает 0,05, т.е. она не может быть отклонена.

Можно считать доказанным тот факт, что качество поставляемых лекарств не зависит от времени поставки и является одинаковым в разное время.

**Команды на выполнение в Excel:**

«Сервис» - «Анализ данных» - «Однофакторный дисперсионный анализ».

**Пример 2.** Требуется оценить влияние уровня рекламы внутри аптеки на объемы продаж. Имеются следующие данные по 30 торговым точкам:

	Уровень рекламы		
	высокий	средний	низкий
	Продажи, тыс. руб.		
1	10	8	5
2	9	8	7
3	10	7	6
4	8	9	4
5	9	6	5
6	8	4	2
7	9	5	3
8	7	5	2
9	7	6	1
10	6	4	2



Однофакторный дисперсионный анализ						
ИТОГИ						
<i>Группы</i>	<i>Счет</i>	<i>Сумма</i>	<i>Среднее</i>	<i>Дисперсия</i>		
Столбец 1	10	83	8,3	1,788888889		
Столбец 2	10	62	6,2	3,066666667		
Столбец 3	10	37	3,7	4,011111111		
Дисперсионный анализ						
<i>Источник вариации</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-Значение</i>	<i>F критическое</i>
Между группами	106,0666667	2	53,03333	17,94360902	1,10362E-05	3,354130829
Внутри групп	79,8	27	2,955556			
Итого	185,8666667	29				

Результаты анализа показывают, что разница в объемах продаж в аптеках с разным уровнем рекламы, является значимой (существенной). Об этом свидетельствует значение  $F$ :  $17,943 > 3,354$ , а также малая вероятность принятия нулевой гипотезы ( $p$  – значение =  $1,10 \text{ E}-0,05$ ).

Следовательно, нулевая гипотеза отклоняется и принимается альтернативная о том, что уровень рекламы влияет на объемы продаж, причем наблюдается прямая зависимость, т.е. более высокому уровню рекламы соответствуют более высокие объемы продаж.

### Двухфакторный дисперсионный анализ. (ДДА)

ДДА выявляет влияние на зависимую переменную уже двух рассматриваемых факторов не только по отдельности, но в их совместном воздействии.

Он в целом не меняет общую логику дисперсионного анализа, но несколько усложняет саму процедуру проведения, так как появляется необходимость в оценке еще и межфакторного метода взаимодействия.

Подобная ситуация возникает в тех случаях, когда совместное влияние двух факторов в отдельности проявляется слабо. Именно в исследовании межфакторного взаимодействия и заключаются особенность и основное достоинство ДДА.

Статистические гипотезы для ДДА формулируются отдельно как для каждого фактора, так и для их совместного влияния (взаимодействия). Для проверки статистических гипотез в ДДА, как и в ОДА, используется то же соотношение дисперсий, тот же самый критерий  $F$  Фишера. Дисперсионный анализ может осуществляться в условиях бесповторного и опыта с повторениями.

### 7. Цель деятельности аспирантов на занятии:

**Аспирант должен знать:**

1. Нормальный закон распределения.
2. Дисперсия. Виды дисперсий.
3. Гипотезы Виды статистических гипотез.
4. Формулу расчета критерия Фишера.

**Аспирант должен уметь:**

1. Выдвигать нулевую и альтернативную гипотезы.
2. Использовать F-критерий Фишера для анализ данных в excel и соответственно в пакете SPSS или Statistica.

**Содержание обучения:**

**Теоретическая часть:**

13. Понятие ДА.
14. Виды ДА.
15. F-критерий Фишера.
16. Применение ОДА и ДДА.

**Практическая часть:**

*Задача 1.* В течение нескольких дней подопытные животные подвергались радиоактивному облучению. Можно ли говорить об изменении радиоактивности крови в связи с длительностью облучения в разных группах животных? (применение ОДА для выявления фактора) (см. упр. 24. Жижин К.С. Медицинская статистика.)

*Задача 2.* Врачом-гигиенистом исследовался процесс окраски детских игрушек из дерева четырьмя видами краски при четырех способах нанесения этой краски на изделие. Необходимо ответить на вопрос: какая из комбинаций: краска + способ окрашивания дают наиболее устойчивое окрашивание. (ДДА с повторениями для выявления факторов). (см. упр. 26. Жижин К.С. Медицинская статистика.)

**24. Перечень вопросов для проверки исходного уровня знаний:**

1. Сущность нормального закона распределения.
2. Понятие дисперсии. Формулы вычисления дисперсии.
3. Формула Фишера. Сущность критерия Фишера.

**25. Перечень вопросов для проверки конечного уровня знаний:**

1. Сущность ОДА.
2. Непараметрические аналоги ОДА.
3. Обязательное условие при использовании ДА.
4. Алгоритм проведения ДА при помощи excel.

**26. Хронокарта учебного занятия:**

1. Организационный момент – 10 мин.
2. Разбор темы – 40 мин.
3. Текущий контроль (тестирование, практическая работа) - 90 мин.
4. Подведение итогов занятия – 10 мин.

**27. Самостоятельная работа аспиранта.**

Непараметрический аналог ОДА для несвязанных выборок – критерий Краскела - Уоллиса и для связанных выборок – критерий Фридмана.

**28. Перечень учебной литературы к занятию:**

5. Есауленко И.Э., Семенов С.Н. Основы практической информатики в медицине; Воронеж, 2005.
6. Жижин К. С. Медицинская статистика; Ростов н/Д, 2007.

## **ТЕМА 10 Многомерные статистические методы.**

### **1. Научно-методическое обоснование темы:**

В последние годы в статистике получают все большее распространение непараметрические методы оценки различий двух групп наблюдений, оценки связи (корреляции) между двумя рядами наблюдений и отнесения наблюдений к одному из двух классов, многомерные статистические методы, так как на практике исследуемая совокупность имеет значительную размерность с большим количеством признаков; выявляет структуру, которую обычными методами анализа выявить просто невозможно.

Сейчас уже ясно, что совершенно недостаточно владеть одним из методов статистической оценки различий двух групп наблюдений. В каждом случае необходимо выбирать подходящий критерий. Это позволяет не только повысить эффективность статистической обработки, но и, как будет ясно из дальнейшего, снизить ее трудоемкость. В большинстве медицинских исследований наиболее подходящим оказывается один из непараметрических критериев различий, которые в настоящее время в медицине применяются относительно редко.

Еще реже применяются в исследовательских работах и клинической практике статистические методы диагностики и прогнозирования.

Следует отметить, что применение непараметрических критериев статистики в медицине и биологии, несомненно, заслуживает значительно более фундаментального изучения. Представляет интерес более подробное изучение общих принципов непараметрической статистики, рассмотрение ряда непараметрических критериев различий (критерия Колмогорова — Смирнова, критерия Ван дер Вардена и др.), более подробное рассмотрение последовательной статистической процедуры, анализ методов и подходов к составлению машинных алгоритмов и программ, использующих принципы непараметрической статистики, применения кластерного анализа в медико-биологических исследованиях, которая вполне доступна для понимания исследователей и практиков.

На практике при разведочном анализе, когда исследователь испытывает дефицит достоверной информации, предпочитают агломеративную стратегию, чтобы оптимизировать количество кластеров, что позволит исследователю определить количество кластеров, которое позволит ориентироваться в ходе дальнейшего конфирматорного (утрачивающего) анализа выборочной совокупности.

Кластерный анализ позволяет разделить эмпирическую выборку на несколько классов(кластеров), но не дает ни правил, ни четких критериев оценки качества классификации, хотя эти правила и критерии важны прежде всего в вопросах диагностики редких, патологических процессов, симптоматика которых весьма размыта. И особенно в процессе оказания экстренной медицинской помощи, когда у врача на перебор вариантов лечебно-диагностической тактики считанные минуты. Для решения таких задач и существует дискриминантный анализ

Кластеризация, многомерное шкалирование, эмпирическое классифицирование основывается на экспертных оценках на основании профессионального опыта врача-диагноста.

### **2. Краткая теория:**

Многомерные статистические методы целесообразно применять в двух основных случаях:

- 1) когда анализируемая совокупность имеет значительную размерность, с большим количеством признаков;
  - 2) когда эксплораторный анализ не обеспечивает информацией о структуре данных.
- Данные методы анализа:

- позволяют уменьшить размерность и получить такой же результат, а возможно, даже открыть иные закономерности;
- выявляют в рассматриваемой совокупности данных так называемую «(латентную)» структуру, которую обычными методами анализа выявить просто невозможно.

Оба этих направления к великому сожалению, еще не используются достаточно широко в медико-биологических исследованиях, и врачи, и биологи о них (за исключением отдельных энтузиастов) очень слабо информированы.

В прикладной статистике этими методами долгое время не могли пользоваться из-за отсутствия вычислительной техники для обработки больших массивов данных. Активно эти методы стали развиваться со второй половины XX в. при появлении быстродействующих компьютеров, выполняющих за доли секунды необходимые вычисления, на которые до этого уходили дни, недели, месяцы.

Мы рассмотрим два основных многомерных метода, которые в медицинской статистике представлены не так широко, как хотелось бы, разберем их достоинства и недостатки, это:

- кластерный анализ;
- дискриминантный анализ;
- факторный анализ.

## **1. КЛАСТЕРНЫЙ АНАЛИЗ.**

***Общая схема применения кластерного анализа в медико-биологических исследованиях.***

Кластерный анализ - это математический метод решения задач классификации, разделения эмпирической выборки на ряд непересекающихся групп, таксонов.

Термин «кластер» (от англ. Cluster) - «гроздь, пучок, скопление» с общим свойством»; а термин «таксон» (от англ. Taxon) обозначает систематизированную группу любой категории. И еще, элементы, объединенные в один кластер, более схожи по сравнению с остальными.

Кластерный анализ не использует никаких дополнительных априорных предположений: например, о характере распределения вероятностей в генеральной совокупности и опирается, как правило, только на данные о самой эмпирической выборке. Как правило, результаты считаются окончательными и не пересматриваются для данной эмпирической выборки, хотя при получении дополнительных данных или при выборе другого метода классификация, вполне понятно, может быть иной.

Иногда можно встретить в литературе информацию, когда кластерный анализ относят к категории статистических методов, предназначенных для так называемой классификации без обучения (в отличие от дискриминантного анализа, который называют классификацией с обучением). Из других названий кластерного анализа можно упомянуть: кластер-анализ, автоматическая классификация, таксономия, распознавание образов без обучения.

Теоретические основы метода были заложены в середине XX в. и продолжают интенсивно развиваться и совершенствоваться в настоящее время. Жаль, что кластерный анализ даже после появления персональных компьютеров, т. е. начиная примерно с 80-х годов прошлого века, в медико-биологических (уж в медицинских - точно!) кругах не стал серьезным и массовым (подчеркнем это слово!) подспорьем в работе ни научных работников, ни практических врачей.

В то же время количество научных публикаций, содержащих результаты, полученные с помощью кластерного анализа, постоянно растет, причем количество работ, посвященных собственно кластерному анализу, до сих пор остается сравнительно небольшим.

Процедура кластерного анализа вполне доступна для понимания исследователей и практиков, не имеющих специальной математической подготовки и не только на интуитивном уровне. Однако обширный арсенал методов кластерного анализа и конкретных задач кластеризации велики, и этот факт - одна из причин того, что в

отечественной литературе работы, посвященные применению кластерного анализа в медицине и биологии, встречаются редко.

Мы хотим показать аспирантам реализацию данного вида обработки экспериментальных данных с использованием пакетов SPSS, Statistica. Они, с нашей точки зрения, должны снять завесу «чрезвычайной сложности» с данного способа анализа, помочь уверенно ориентироваться при использовании в реальных исследованиях кластерного анализа.

#### **Этапы применения кластерного анализа**

1. Получение с помощью конкретных измерительных шкал выборки эмпирических данных, представление ее в виде матрицы «объект - признак».
2. Определение направления кластеризации, классификации: пациенты, респонденты, наблюдения, измеренные признаки, или и то и другое одновременно.
3. Распределение эмпирических данных в виде точек многомерного метрического пространства с определенными координатами; определение меры сходства или различия между его точками.
4. Выбор основного принципа разделения выборки на кластеры.
5. Выбор конкретного алгоритма кластеризации с характерным приемом. определения мер сходства или различия между кластерами, т. е. способа определения межкластерных расстояний, и, естественно, способа оценки качества кластеризации.
6. Выполнение кластеризации или разбиения исходной выборки на кластеры.
7. Интерпретация результатов кластеризации.

Основные приемы кластерного анализа: по измерительным шкалам, направлению кластеризации и используемой метрике.

Все три этапа процедуры кластеризации целесообразно рассмотреть совместно, так как и в теории, и тем более на практике они тесно взаимосвязаны между собой.

Выборка данных - результат измерения ряда признаков, характеристик процессов, состояний, свойств: «X», некоторой совокупности объектов, пациентов: «A». Получение такой выборки предполагает наличие определенных измерительных методик.

Согласно им, результаты измерения могут быть представлены в номинальной, порядковой, интервальной шкалах или' шкале отношений. Математическим основанием здесь являются интервальные шкалы и шкалы отношений.

Но в кластерном анализе в отличие, например, от факторного, дискриминантного или дисперсионного анализа требования к типу шкалы не являются столь жесткими: они могут повлиять на выбор конкретного метода кластеризации, но не на допустимость кластеризации.

В случаях смешения типа шкал вопрос о выборе подходящего метода кластеризации должен решаться особенно тщательно: теоретические основы анализа при использовании смешанных шкал исследованы недостаточно, и велика опасность ошибки кластеризации, а, следовательно, и окончательного вывода в исследовании.

Наиболее надежными выходами являются следующие:

- применить метод, предназначенный для номинальной шкалы;
- выбрать меру расстояния, предназначенную для шкал смешанного типа;
- самый простой путь - стремиться избегать использования шкал разного типа.

Как показывает практика, тип шкалы, к сожалению, никак не определяет характера процедуры кластеризации в целом. Дело в том, что использованные при измерениях шкалы влияют на выбор подходящего метода кластеризации не прямо, а косвенно - через выбор необходимого метрического пространства.

Эмпирические данные формируются в виде матрицы «объект - признак»). Это прямоугольная таблица чисел, строки которой соответствуют измеренным объектам (пациенты, подопытные животные, препараты, процедуры), а столбцы - измеряемым признакам (процессов, состояний или свойств).

Однако технологически исследователь сначала заносит данные в таблицу «объект-признак»:

Объект	Признак			
	X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>	X <sub>4</sub>
A <sub>1</sub>	A <sub>11</sub>	A <sub>12</sub>	A <sub>13</sub>	A <sub>14</sub>
A <sub>2</sub>	A <sub>21</sub>	A <sub>22</sub>	A <sub>23</sub>	A <sub>24</sub>
A <sub>3</sub>	A <sub>31</sub>	A <sub>32</sub>	A <sub>33</sub>	A <sub>34</sub>
A <sub>4</sub>	A <sub>41</sub>	A <sub>42</sub>	A <sub>43</sub>	A <sub>44</sub>

И только потом появляется матрица. Она будет отличаться от таблицы «объект-признак» только тем, что в матрице явно не присутствуют заголовки строк и столбцов. При построении матрицы «объект-признак» нередко возникает проблема из-за разнотипности шкал измерения признаков, подобное требует нормирования показателей, т. е. введения условной единицы измерения, допускающей формальные сопоставления объектов, но нельзя упускать из виду, что способы нормирования применимы лишь к результатам измерений в шкалах интервалов и отношений. Приложение их к номинальным или порядковым данным является некорректным. Ситуация, однако, не фатальна и в таких ситуациях существуют адекватные меры различия или сходства.

Отметим, что способы нормирования обычно выполняются «по столбцу», однако при необходимости аналогичное нормирование можно выполнить и «по строке».

После этого данные представляются в виде точек многомерного пространства, но до этого необходимо принять решение о направлении кластеризации, т. е. о том, что и как именно будет подвергаться разделению на кластеры.

При кластеризации в ее классическом понимании осуществляется и кластеризация объектов, и кластеризация признаков. Одновременная кластеризация используется редко, и интересующихся мы отсылаем к специальной литературе (Hartigan G.A. Clustering algorithmus. - New York, 1975). В зависимости от выбранного направления кластеризации (объекты или признаки) исследователь может представить выборку эмпирических данных в качестве набора точек многомерного пространства двумя различными способами:

- Набор точек - как объекты.
- Набор точек - как признаки.

Сам исследователь решает, исходя из поставленной цели, что и как он будет анализировать. В зависимости от этого образуемое для представления данных метрическое пространство будет иметь размерность: равную либо  $n$  - числу объектов, либо  $m$  - числу измеренных признаков каждого объекта.

Естественно, принципиальных различий для кластеризации объектов или признаков нет: это для кластерного анализа, в определенном смысле «все равны».

Безусловно, для осуществления кластеризации полученное многомерное пространство данных надо превратить в метрическое, указав способ определения расстояния (метрики) между его точками.

*Метрическое пространство* - это пространство, включающее серию объектов, называемых его элементами, между которыми задана функция расстояния «а», называемая метрикой, определенная на всех упорядоченных парах точек множества и удовлетворяющая следующим условиям:

- Неотрицательность.
- Рефлексивность.
- Симметричность.
- Транзитивность.

Нередко требования к расстоянию ослабляют, отказываясь от некоторых из них: чаще всего - от транзитивности или симметричности. В этом случае мы имеем дело уже с «ослабленной. величиной расстояния, так как для нее выполняются не все фигурирующие в определении требования.

Во многих методах кластерного анализа использование псевдометрик является корректным в силу того, что недостающие метрические свойства не используются. По-

этому разговор о мерах различия: метрики и псевдометрики - особой роли, кроме как с теоретической точки зрения, не играет.

Мера различия ведет от матрицы «объект-признак» к матрице попарных расстояний между эмпирическими точками построенного метрического пространства (в рассматриваемом случае - между объектами.)

Существенный признак матрицы в том, что она, во-первых, симметрична, во-вторых, по диагонали идут нули. Получив матрицу расстояний, можно перейти к последующим этапам процедуры кластеризации.

При одной и той же стратегии кластеризации могут использоваться различные меры различия или сходства. Каждая из них имеет свои особенности.

Наиболее часто используемое понятие «евклидово расстояние» наиболее популярно. Хотя имеет ограничения на применение только к данным, измеренным в шкалах интервалов или отношений, но на практике часто применяется и для данных, полученных в других шкалах (хотя и не всегда корректно). Наибольший эффект получается, если использовать евклидово расстояние для переменных, измеренных в одних и тех же единицах (или для нормированных данных); в противном случае следует использовать ее нормированный вариант.

Расстояние «Манхэттен» применяется для номинальных и дихотомических признаков как сумма покоординатных различий между точками. Во многом аналогично евклидову, однако при его применении сглаживается эффект больших различий по отдельным координатам.

Расстояние Минковского является обобщением случаев евклидова расстояния «Манхэттен» и ряда других. В силу этого парадигму Минковского удобно использовать при экспериментах с подбором расстояния.

Есть еще один коэффициент сходства, разработанный Гауэром. Он позволяет одновременно использовать признаки, измеренные в трех различных шкалах: интервальных, порядковых и дихотомических. В этом его явное преимущество, тем более, что мер сходства для работы со смешанными шкалами разработано мало. К сожалению, коэффициент Гауэра практически не реализован в рассматриваемых нами статистических пакетах.

В кластерном анализе применяется множество иных мер сходства или различия:

- Для интервальных данных - «квадрат евклидова расстояния», Чебышева, Махаланобиса, коэффициент корреляции Пирсона.
- Для порядковых данных - Хи-квадрат, Фи-квадрат, коэффициенты ранговой корреляции Спирмена, Кендалла, Чупрова.
- Для номинальных и дихотомических данных - рассеяние, дисперсия, четырехпольный коэффициент корреляции Фи и др.
- Для данных, измеренных в смешанных шкалах, применяются меры близости отечественных исследователей - Журавлева, Воронина, Миркина.

Понятно, что сколь бы ни была сильна субъективная точка зрения на выбор меры различия или сходства, все же она в первую очередь определяется объективными свойствами исследуемого явления, напрямую связанными с характером измерительных шкал.

## **КЛАССИФИКАЦИЯ МЕТОДОВ КЛАСТЕРНОГО АНАЛИЗА ПО СТРАТЕГИЯМ КЛАСТЕРИЗАЦИИ**

После того, как построено метрическое пространство, дальнейшая часть процедуры кластерного анализа достаточно, автономна: здесь уже неважно, как именно задавалась метрика и что именно (объект или признаки) представлялось в виде точек пространства. Главное, что к этому этапу построена матрица попарных расстояний (или попарных мер сходства), которая используется на последующих ступенях кластерного анализа.

Какова же стратегия кластеризации, т. е. основного принципа ее осуществления? Классификация методов кластерного анализа не является самоцелью уже потому, что



весьма непроста по сути, чтобы четко и односложно сориентировать читателя в необъятном море разработанных методов и алгоритмов кластеризации. Для практического применения, и тем более на начальных этапах освоения метода, вполне достаточно иметь представление о следующих приемах кластеризации:

- иерархические;
- итеративные;
- алгоритмы разрезания графа.

Для начального ознакомления и практического использования сосредоточим основное внимание на иерархических и итеративных методах кластеризации.

В иерархических методах выстраивается «граф, или дерево», кластеров, где в окончательных кластерах можно увидеть динамику отдельных точек метрического пространства данных.

В итеративных методах разбиение на кластеры ведет к последовательным перерасчетам приближений, итераций. И тот и другой методы подразделяют на дивизивные (разделительные) и агломеративные (объединительные). Это деление отражает желаемый результат применения кластерного анализа, а не его технологию (итеративное, или «прямое», построение кластеров).

В дивизивных иерархических методах множество исходных данных формирует один большой кластер, который дробится на заранее заданное количество мелких кластеров. Процесс завершается, когда получено заданное число кластеров при определенном удовлетворяющем исследователя качестве разделения. В дивизивных методах иерархические приемы обработки доминируют над итеративными.

Иногда заранее выделяют некоторое количество так называемых «эталонных» кластеров, к которым постепенно присоединяются все оставшиеся эмпирические точки пространства данных. Процесс кластеризации заканчивается, когда получено удовлетворительное качество разбиения.

Популярным приемом является метод k-средних.

В любом случае вопрос о выборе критериев качества разбиения на кластеры является достаточно сложным.

Агломеративные методы, напротив, насыщены не итеративными, а иерархическими приемами обработки данных. В них каждый элемент эмпирической выборки представляется отдельным кластером. Затем идет объединение; при этом на каждом шаге группируются наиболее близкие друг к другу кластеры. Это кластеры более высокого уровня в иерархии кластеров, отсюда подобные приемы называют методами иерархической кластеризации. Кластеризация имеет конечное число шагов, в итоге формируется единственный, «глобальный общий» кластер, идентичный исходной эмпирической выборке.

То есть если в агломеративных методах кластеризация множества одноэлементных кластеров формирует парадигму одного кластера. В дивизивных методах все наоборот: один общий глобальный кластер дробится на большое число мелких кластеров. Максимальное количество отдельных кластеров не может превосходить количества элементов в этой выборке.

Это в теории, а на практике исследователь сам задает количество кластеров, на которые надо разделить выборку, исходя из условий, диктуемых особенностями постановки эксперимента. Классификация иерархических агломеративных методов кластерного анализа по способам определения межкластерных расстояний.

Выполняя иерархическую агломеративную кластеризацию, надо решить вопрос о выборе конкретного способа определения межкластерных расстояний. И дело в том, что в кластерном анализе расстояние рассматривается в двух смыслах:

- 1) как расстояние между объектами внутри кластера;
- 2) как межкластерное расстояние.

Иначе при решении задач кластерного анализа возникнет проблема выбора наиболее подходящего способа определения межкластерных расстояний.

Эта проблема общая для дивизивных и агломеративных; для иерархических и итеративных методов кластеризации. Однако в каждом функционально полном статистическом пакете программ для этого имеются соответствующие возможности, хотя сами наборы способов определения межкластерных расстояний, могут существенно отличаться.

Вот наиболее существенная их подборка:

- Простая связь, одиночная связь, метод «ближнего соседа» - здесь расстояние между кластерами рассматривается попарно между двумя самыми близкими. Обладает сильной компрессией, формирует минимальный граф объединения.
- Полная связь, или метод «дальнего соседа», - здесь исходное пространство растягивается.
- Невзвешенная попарногрупповая средняя - в этом случае расстояние между двумя кластерами трактуется как среднее по всем парным расстояниям, метод не меняет размерность исходного внутрикластерного пространства.
- Метод Уорда - этот метод сильно изменяет метрическое признаковое пространство и формирует резко выраженные кластеры. Хорош для выявления трудноуловимых различий, однако в этом варианте анализа легко выдать желаемое за действительное, т. е. усмотреть в случайности стойкую закономерность.

Агломеративная кластеризация фигурирует в литературных источниках в следующих модификациях:

- Взвешенная попарно-групповая средняя.
- Невзвешенная попарно-групповая центроидная.
- Взвешенная попарно-групповая центроидная медианная).
- Межгрупповое связывание.
- Внутригрупповое связывание.
- Центроидная кластеризация.
- Медианная кластеризация.

«Разброс» стратегий, как видно из перечня, широк, и если мы хотим получить от кластеризации наибольший эффект, лучше ее осуществлять несколькими методами, выбирая наиболее предпочтительную. Это, между прочим, характерно для всех многомерных методик: не столько важна методика статистической обработки, сколько ее интерпретация.

## **ПРИЕМЫ КЛАСТЕРНОГО АНАЛИЗА В МЕДИКО-БИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ**

Агломеративные и дивизивные методы кластеризации в решении задач, возникающих в медико-биологических исследованиях. В статистических пакетах SPSS и Statistica.

Поскольку для большей части исследователей-врачей, или биологов данный раздел статистической обработки будет совершенно необычным и новым, в самом начале кластеризации стоит объяснить стратегическую направленность кластеризации, что является сутью исследования: агломерация (объединение) и дивизияция (разделение).

На практике при разведочном (эксплораторном) анализе, когда исследователь испытывает дефицит достоверной информации, предпочитают агломеративную стратегию, чтобы оптимизировать количество кластеров. Такой подход позволяет исследователю определить количество кластеров, которое позволит ориентироваться в ходе дальнейшего конфирматорного (уточняющего) анализа выборочной совокупности.

Важно подчеркнуть, результат в полной зависимости от того, насколько эта выборка репрезентативна, чтобы, опираясь на ее результаты, характеризовать генеральную совокупность. Этот момент должен быть исследован отдельно: с помощью

дискриминантного анализа, методов получения репрезентативной выборки, ее необходимого объема, валидности методик и т. д.

Итак, как мы сказали, данный вид анализа носит эвристический характер и соответственно не имеет под собой достаточных статистических оснований. В любой момент может возникнуть потребность повторного проведения кластерного анализа с использованием иных методов кластеризации.

**Примечание.** *Неопытными исследователями результаты кластеризации выдаются за окончательные и единственно возможные. Это глубокое заблуждение, поскольку кластеризация - начало статистического разделительного анализа.*

Даже в научных статьях подчас никакого обсуждения устойчивости, сравнительного анализа применения различных стратегий кластеризации, как правило, не приводится. Тем не менее вполне реальна возможность радикального изменения выводов экспериментального исследования при отступлении от используемых кластеризационных процедур. Пренебрежение этими установками может приводить к полярным результатам кластеризации одних и тех же эмпирических данных.

Алгоритм применения кластерного анализа в любом исследовании при использовании статистических пакетов программ должен учитывать: а) Типы измерительных шкал, примененных для получения выборки: интервальные, порядковые, номинальные, дихотомические шкалы, их однотипность.

б) Подходит или нет статистический пакет кластерного анализа.

в) Направление кластеризации, меру сходства или различия для построения метрического пространства данных, глобальную стратегию кластеризации.

г) Содержательную интерпретацию кластеризации, дополнительную проверку на других приемах кластеризации, других статистических пакетов.

Приложение предложенного алгоритма к реальной ситуации на практике может высветлить явное отличие от приведенной канонической схемы. Сущность этого несоответствия обусловлена наличием тех самых разнотипных измерительных шкал, о которых сказано выше, и в силу этого для определения сходства между объектами обязательно применение коэффициент Гауэра.

## **2. ДИСКРИМИНАНТНЫЙ АНАЛИЗ**

### ***Основы дискриминантного анализа***

Кластерный анализ позволяет разделить эмпирическую выборку на несколько классов (кластеров), однако не дает, ни правил, ни четких критериев оценки качества классификации. В то же время и правила, и критерии важны прежде всего в вопросах диагностики редких, нетипичных патологических процессов, симптоматика которых весьма размыта. И особенно в процессе оказания ургентной (экстренной) медицинской помощи, когда у врача на перебор вариантов лечебно-диагностической тактики считанные минуты.

Для решения подобных задач и существует дискриминантный анализ. И хотя дискриминантный и кластерный анализы близки по сути (направлены на решение задач классификации), но подходами к классификации принципиально различаются.

Дискриминантный анализ, как и кластерный анализ, направлен на разделение выборки на ряд кластеров, но его конечная цель - отнесение некоторого объекта к одному из уже построенных классов, а также проверка непротиворечивости классификации.

Термин «дискриминация» (от лат. *discriminatio* - разделение) означает не только разделение объектов на классы, но и ограничение такого разделения.

Это ряд методов, с помощью которых мы можем отнести новый объект к одному из заранее построенных классов, а также проверить качество построенной классификации. Еще дискриминантный анализ называют анализом с обучающей выборкой для распознавания образов или классификацией с обучением.

Кластеризация, многомерное шкалирование, эмпирическое классифицирование основывается на экспертных оценках на основании профессионального опыта врача-диагноста.

Алгоритм дискриминации таков:

1. Проверить, создана ли выборка данных в интервальных шкалах или в шкалах отношений, имеют ли признаки нормальное распределение вероятностей.

2. Проверить, разделена ли выборка на конечное число (не менее двух) непересекающихся классов, известна ли для каждого объекта его принадлежность к определенному классу. (Можно ограничиться значениями вероятностями принадлежности объекта к каждому классу.)

3. Если все обстоит так, то можно приступать к решению основных вопросов дискриминации:

- Принадлежит ли произвольно выбранный объект из генеральной совокупности к одному из классов, на которые разделена эмпирическая выборка, и можно ли конструировать правило классификации. Можно ли систему распознавания научить определять принадлежность объекта к тому или иному классу?

- Каково качество построенной классификации: насколько она чутка к разделению объектов на классы, насколько такая дифференцировка достоверна?

- Каковы информативные признаки из числа измеряемых у исследуемых объектов, какие из них имеют наибольшее значение для правильного и качественного дифференцирования.

Существует ряд разновидностей дискриминантного анализа, но математическая суть у них едина, поэтому для практического применения рассмотрим три основных направления дискриминантного анализа, реализованных в большинстве статистических пакетов:

- линейный дискриминантный анализ Фишера;
- канонический дискриминантный анализ;
- пошаговый дискриминантный анализ.

Линейный дискриминантный анализ Фишера (линейная дискриминация Фишера, дискриминантный анализ) предложен Р. Фишером. Суть его в том, что разбиения выборочной совокупности строятся на так называемой линейной комбинации значений измеренных признаков. Ее аналитическое выражение таково:

$$h_k = b_{ko} + \sum_{j=1}^m b_{kj} X_j$$

2. Новый объект можно отнести к какому-то классу согласно классифицирующей функции, если значение конкретного признака является максимальным среди всех значений, вычисленных на этом объекте.

В основе метода Фишера лежит еще одно предположение, накладываемое на ковариации переменных: признаки должны иметь статистически идентичные ковариационные матрицы.

Ковариация двух переменных - мера их совместного изменения, равноценна коэффициенту корреляции Пирсона. Однако показатель ковариации в отличие от коэффициента Пирсона может принимать произвольные значения, а не только в пределах:  $[-1 \leq r \leq +1]$ .

Канонический дискриминантный анализ - схема обратна первому виду анализа: здесь разделение объектов ведется по минимальным значениям дискриминирующей функции. Вопрос отнесения объекта к определенному классу возможно положительно решить только тогда, когда евклидово расстояние от центра кластера до оцениваемого показателя минимально. Такой вид анализа, конечно, более сложен и трудоемок в реализации.

Тем более, на основе проведенных численных экспериментов ряд авторов отмечают, что результаты анализа Фишера и канонического дискриминантного анализа совпадают.

С вводом в обиход персональных компьютеров широкое распространение получил так называемый (пошаговый метод дискриминации). Он, как и линейный вид анализа, достаточно прост в реализации и помогает наглядно за счет последовательного включения (исключения) наиболее информативных дискриминантных переменных на каждом шаге (для каждого текущего набора дискриминантных переменных) оценивать качество полученной классификации.

**Примечание.** *Следует обратить самое серьезное внимание на обязательную нормальность распределения в генеральной совокупности, которая часто не выполняется для эмпирических данных. пренебрежение этим может привести к серьезным ошибкам классификации.*

### 3. ФАКТОРНЫЙ АНАЛИЗ

#### *Теоретические основы факторного анализа*

Факторный анализ сегодня самый популярный из всех многомерных методов анализа, но, как правило, в далеких от медицины областях, если не считать фундаментальной работы на эту тему немецкого ученого врача Карла Иберлы, вышедшей в 80-е годы XX столетия. К сожалению, знаком с этим методом, возможно, не более чем один врач из тысячи. Наша сегодняшняя задача состоит в том, чтобы донести до читателя основу этого вида анализа и на практических примерах показать, как он влияет на синтез, интеграцию и интерпретацию конечного результата обработки данных.

В отличие от всех ранее описанных в данной книге приемов обработки эмпирической информации факторный анализ не только позволяет сжать объемы информации, но на совершенно новой основе строит доказательство влияния этих факторов. Если все предыдущие методы цифрами всего лишь подтверждали влияние подразумеваемого, предполагаемого фактора или группы факторов, то в данном случае выявляется этот самый скрытый (латентный) фактор или группа и цифрами объясняется его влияние.

Факторный анализ в принципе – мультифакторный анализ, но «начало» его в однофакторном анализе ч. Спирмена и двухфакторном (бифакторном) – К. Холзингера.

Естественно, математические основы, к примеру, у одно-двухфакторного и мультифакторного анализа существенно отличаются.

Это сугубо математический метод, в котором обязательно используется корреляционная матрица – матрица попарных коэффициентов линейной корреляции Пирсона между исследуемыми признаками.

Процедура извлечения факторов с помощью корреляционной матрицы исходных данных называется факторизацией.

Концепция факторного анализа заключена в следующем:

- Истинные причины изучаемого явления не могут быть непосредственно наблюдаемы и доступны, их число также неизвестно исследователю.
- Признаки измерены в интервальных шкалах.
- Предполагается нормальность распределения исследуемых эмпирических данных в генеральной совокупности.
- Постулируется ортогональность и независимость выявляемых факторов, хотя это положение на практике трудновыполнимо.

Из всего сказанного ясно, что применимость методов факторного анализа является весьма «жесткой», ограничительной, и корень зла прежде всего, в частности, в интервальных измерительных шкалах и соответствии вероятностного распределения признаков нормальному закону.

В медико-биологических исследованиях эти постулаты часто не выполняются, и, естественно, теоретические основы факторного анализа фактически являются весьма условными

Тем не менее, как пишет К. Иберла, упомянутые ограничения на применение факторного анализа можно если не совсем обойти, то в той или иной степени ослабить использованием методов эвристического склада, т. е. позволяющих получить решение без его исчерпывающего теоретического обоснования.

Мы не будем подробно останавливаться на классификации его методов, тем более что некоторые авторы отмечают, что различные методы дают принципиально одинаковые результаты.

***Модель факторного анализа такова:***

1. Имеется  $N$  объектов (например, испытуемых), для каждого из которых измерено  $n$  признаков (например, некоторых свойств). Результаты представлены в виде матрицы «объект-признак».

2. Исходные эмпирические данные нормируются.

Идея факторного анализа состоит в том, чтобы представить нормированные значения матрицы «объект - признак» в виде линейной комбинации небольшого числа скрытых (латентных) факторов, т. е. упростить структуру признакового пространства.

С помощью данной модели вводится в рассмотрение ряд базовых понятий факторного анализа:

- Общие факторы - выделяемые при факторном анализе, как мы уже сказали, - латентные факторы, их нельзя измерить непосредственно, но можно выделить статистическими методами.
- Специфические факторы - выделяемые при факторном анализе латентные факторы, воздействующие на какой-либо один' определенный признак.
- Факторные нагрузки - не известные заранее коэффициенты общих и специфических факторов.
- Общность - вклад общих факторов в дисперсию признака.
- Характерность - вклад специфических факторов в дисперсию.
- Факторная матрица - матрица, составленная из координат общих факторов.
- Факторные веса коэффициенты факторной матрицы.
- Объясненная дисперсия - часть общей дисперсии, объясняемая с помощью выделенных факторов.
- Собственные значения - рассматриваемые в математике собственные значения матрицы «объект -признак».
- Факторная структура - набор общих факторов, которые заменяют собой исходные признаки.

Центральное звено факторного анализа составляет оценка факторных нагрузок, приемы могут быть разные: метод главных компонент, метод главных факторов и т. д. Факторы задают по сути новые оси в пространстве признаков в декартовой системе координат. В этом плане главная особенность факторного анализа - вращение факторов, позволяющее получить более простую и легче интерпретируемую факторную структуру.

В ФА применяется множество видов вращения факторов:

- (Варимакс) и (Нормализованный Варимакс);
- (Биквартимакс) и (Нормализованный Биквартимакс);
- (Квартимакс) и (Нормализованный Квартимакс);
- (Эквимакс) и (Нормализованный Эквимакс).

Общего «рецепта» вращения не существует. Исследователь сам подбирает наиболее подходящий метод вращения факторов эмпирическим путем.

Проиллюстрируем применение факторного анализа на конструировании из множества исходных признаков небольшого количества новых переменных (главных компонент), объясняющих значительную часть общей дисперсии. Метод главных компонент специфичен, но часто рассматривается как один из методов факторного анализа: именно в таком качестве он реализован в пакетах SPSS и Statistica.

В пакете Statistica реализовано множество разновидностей методов факторного анализа:

- главных компонент;
- главных факторов;
- главных осей;
- максимального правдоподобия;
- центроидный.

Представление данных при проведении факторного анализа может осуществляться, или в виде «сырых», первичных данных, или в виде готовой матрицы корреляций. Это непринципиально, но вносит определенные дополнительные коррективы в процесс обработки эмпирических данных.

Факторный анализ — это многомерный метод, который применяется при изучении существующих между значениями переменных взаимосвязей. По нему, известные переменные зависят от случайной ошибки и меньшего количества переменных, которые неизвестны. Впервые факторный анализ возник в психометрике, однако уже в наше время его широко используют в социологии, нейрофизиологии, политологии, а также статистике, экономике и других науках. С появлением новейших компьютерных программ его начали и в них применять. Факторный анализ в excel помог значительно упростить общие задачи, который стоят перед эти видом исследования.

С помощью факторного анализа можно решить несколько важнейших проблем исследования, а именно: всесторонне описать объект измерения; сделать это компактно. Этот вид анализа позволяет найти скрытые переменные факторы, которые отвечают за наличие статистических линейных связей корреляций, которые существуют между исследуемыми переменными.

В общем понимании следует выделить две главные цели факторного анализа:

- поиск взаимосвязей между переменными;

сокращение, для описания данных, необходимого числа переменных. (готовый пример см. файл Faktor\_rent.xls)

## ФАКТОРНЫЙ АНАЛИЗ В EXCEL: ПРИМЕР

Факторным называют многомерный анализ взаимосвязей между значениями переменных. С помощью данного метода можно решить важнейшие задачи:  
всесторонне описать измеряемый объект (причем емко, компактно);  
выявить скрытые переменные значения, определяющие наличие линейных статистических корреляций;  
классифицировать переменные (определить взаимосвязи между ними);  
сократить число необходимых переменных.

### ПРИМЕР проведение факторного анализа.

Допустим, нам известны продажи каких-либо товаров за последние 4 месяца. Необходимо проанализировать, какие наименования пользуются спросом, а какие нет.

	A	B	C	D	E	F
1	Наименование	1 мес.	2 мес.	3 мес.	4 мес	Итого
2	Товар 1	310	421	513	646	1890
3	Товар 2	183	235	272	317	1007
4	Товар 3	136	110	428	319	993
5	Товар 4	112	231	349	473	1165
6	Товар 5	57	175	234	400	866
7	Товар 6	68	179	305	425	977
8	Товар 7	38	10	87	133	268
9	Товар 8	47	85	132	143	407
10	Итого:	951	1446	2320	2856	7573

### Исходные данные.

Посмотрим, за счет, каких наименований произошел основной рост по итогам второго месяца. Если продажи какого-то товара выросли, положительная дельта – в столбец «Рост». Отрицательная – «Снижение». Формула в Excel для «роста»: =ЕСЛИ((C2-B2)>0;C2-B2;0), где C2-B2 – разница между 2 и 1 месяцем. Формула для «снижения»: =ЕСЛИ(J3=0;B2-C2;0), где J3 – ссылка на ячейку слева («Рост»). Во втором столбце – сумма предыдущего значения и предыдущего роста за вычетом текущего снижения.

Н	I	J	К
Факторы		Рост	Снижение
Итог 1 мес.	951	0	0
Товар 1	951	111	0
Товар 2	1062	52	0
Товар 3	1088	0	26
Товар 4	1088	119	0
Товар 5	1207	118	0
Товар 6	1325	111	0
Товар 7	1408	0	28
Товар 8	1408	38	0
Итог 2 мес.	1446		
	=I3+J3-K4		

Рост по итогам.

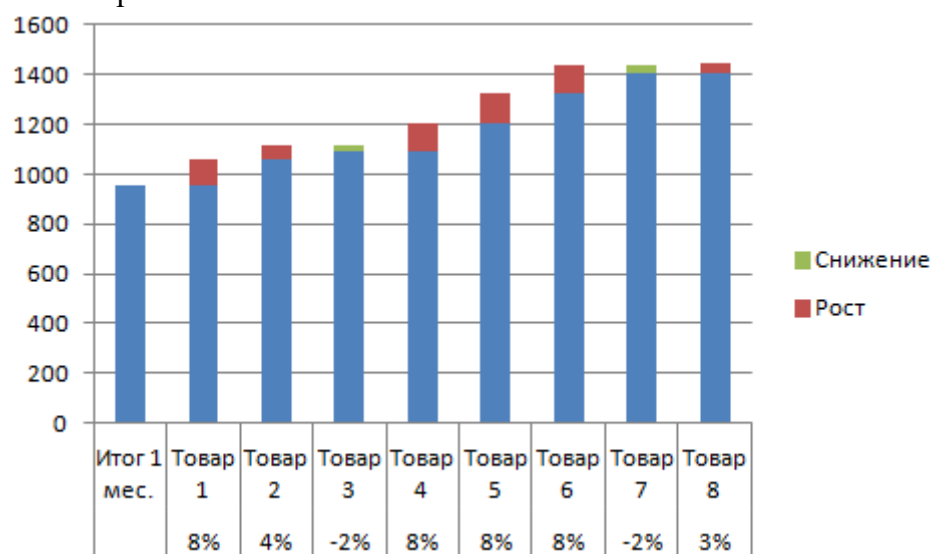


Рассчитаем процент роста по каждому наименованию товара. Формула: =ЕСЛИ(J3/\$I\$11=0;-K3/\$I\$11;J3/\$I\$11). Где J3/\$I\$11 – отношение «роста» к итогу за 2 месяц, ; -K3/\$I\$11 – отношение «снижения» к итогу за 2 месяц.

G	H	I	J	K
% роста	Факторы		Рост	Снижение
	Итог 1 мес.	951	0	0
8%	Товар 1	951	111	0
4%	Товар 2	1062	52	0
-2%	Товар 3	1088	0	26
8%	Товар 4	1088	119	0
8%	Товар 5	1207	118	0
8%	Товар 6	1325	111	0
-2%	Товар 7	1408	0	28
3%	Товар 8	1408	38	0
	Итог 2 мес.	1446		

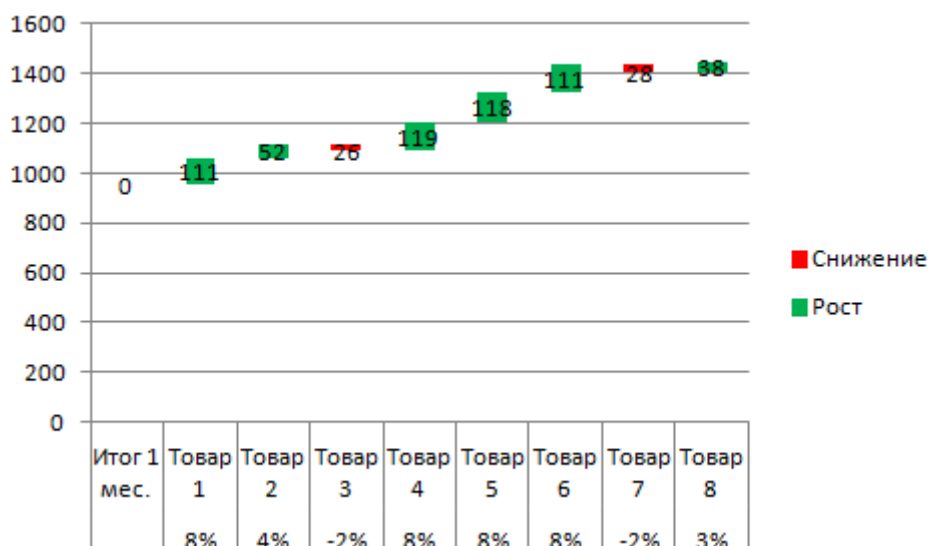
Детализация роста.

Выделяем область данных для построения диаграммы. Переходим на вкладку «Вставка» - «Гистограмма».



Гистограмма.

Поработаем с подписями и цветами. Уберем накопительный итог через «Формат ряда данных» - «Заливка» («Нет заливки»). С помощью данного инструментария меняем цвет для «снижения» и «роста».



Теперь наглядно видно, продажи какого товара дают основной рост.

## ДВУХФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ В EXCEL

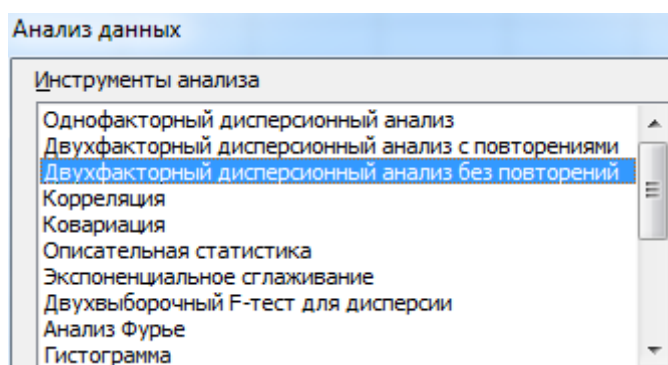
Показывает, как влияет два фактора на изменение значения случайной величины. Рассмотрим двухфакторный дисперсионный анализ в Excel на примере.

Задача. Группе мужчин и женщин предъявляли звук разной громкости: 1 – 10 дБ, 2 – 30 дБ, 3 – 50 дБ. Время ответа фиксировали в миллисекундах. Необходимо определить, влияет ли пол на реакцию; влияет ли громкость на реакцию.

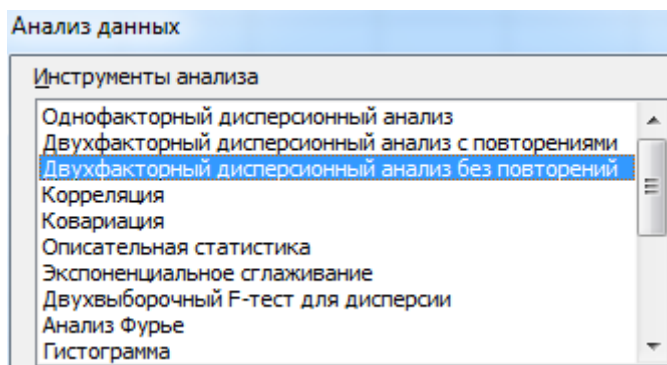
	A	B	C	D
1		Громкость		
2	Пол	1	2	3
3	муж.	304	272	223
4	жен.	262	269	183

Исходная таблица.

Переходим на вкладку «Данные» - «Анализ данных» Выбираем из списка «Двухфакторный дисперсионный анализ без повторений».



Заполняем поля. В диапазон должны войти только числовые значения.



Результат анализа выводится на новый лист (как было задано).

	A	B	C	D	E	F	G
1	Двухфакторный дисперсионный анализ без повторений						
3	ИТОГИ	Счет	Сумма	Среднее	Дисперсия		
4	Строка 1	3	6	2	1		
5	Строка 2	3	799	266,3333	1664,33333		
6	Строка 3	3	714	238	2281		
7							
8	Столбец 1	3	567	189	26949		
9	Столбец 2	3	543	181	24033		
10	Столбец 3	3	409	136,3333	13733,3333		
11							
12							
13	Дисперсионный анализ						
14	Источник вариации	SS	df	MS	F	P-Значение	F критическое
15	Строки	126370,9	2	63185,44	82,6013509	0,000558863	6,94427191
16	Столбцы	4832,889	2	2416,444	3,15898032	0,150290749	6,94427191
17	Погрешность	3059,778	4	764,9444			
18							
19	Итого	134263,6	8				

Так как *F*-статистики (столбец «F») для фактора «Пол» больше критического уровня *F*-распределения (столбец «F-критическое»), данный фактор имеет влияние на анализируемый параметр (время реакции на звук).

Для фактора «Громкость»:  $3,16 < 6,94$ . Следовательно, данный фактор не влияет на время ответа.

## 8. Цель деятельности аспирантов на занятии:

**Аспирант должен знать:**

- Многомерные статистические методы. Виды.
- Случаи использования многомерных статистических методов.
- Схема применения кластерного анализа в медико-биологических исследованиях.
- Классификация кластерного анализа. Приемы кластеризации.
- Алгоритм дискриминации.
- Концепцию и базовые понятия факторного анализа.

**Аспирант должен уметь:**

- Применять кластерный анализ на практике, используя приемы кластеризации (пакеты SPSS или Statistica).

4. Применять пошагово дискриминантный анализ в клинической практике (пакеты SPSS или Statistica).
5. Применять разные методы факторного анализа при конструировании из множества исходных признаков небольшого количества новых переменных, извлекать факторы с помощью корреляционной матрицы исходных данных.

### Содержание обучения:

#### Теоретическая часть:

17. Многомерные статистические методы и их виды.
18. Схема применения кластерного анализа в медико-биологических исследованиях.
19. Классификация кластерного анализа. Приемы кластеризации.
20. Приемы кластерного анализа в медико-биологических исследованиях.
21. Основы дискриминантного анализа.
22. Теоретические основы факторного анализа.

#### 29. Практическая часть:

**Задача №1.** Агломеративная кластеризация для эксплораторного анализа данных (пакеты SPSS и Statistica).

*Условие:* профессиональный отбор врачей –лаборантов сопровождается анализом их профессионально значимых функций(ПЗФ), уровень развития которых оценивается по психофизиологическим реакциям, в частности: распределение внимания по таблицам Шульте-Платонова (ШП, сек), срывам дифференцированной реакции на сложный световой раздражитель(СД, абс. число срывов), тактильной чувствительности (ТЧ, мм).

*Вопрос:* можно ли разделить 32 претендента на группы, сколько таких групп может получиться исходя из результатов профотбора, поскольку руководитель организации стоит перед сложной материальной проблемой оснащения не более четырех лабораторий?

№	ШП	СД	ТЧ	№	ШП	СД	ТЧ
1	66	6	4	17	50	5	4
2	40	4	2	18	52	5	3
3	50	4	2	19	48	5	3
4	70	6	2	20	47	5	4
5	54	5	3	21	48	5	3
6	70	6	3	22	70	7	4
7	50	5	4	23	50	5	4
8	49	4	3	24	54	5	5
9	48	5	3	25	60	4	5
10	70	6	4	26	70	7	4
11	45	5	3	27	50	4	5
12	70	6	3	28	48	5	4
13	47	5	7	29	51	5	3
14	54	5	5	30	52	4	5
15	49	5	7	31	47	5	7
16	48	5	7	32	51	5	4

(решение см. стр.102, 108 Жижин «Медицинская статистика»)

**Задача №2.** Дискриминантный анализ эмпирических данных-случай подтверждения допустимости классификации.

**Условие.** Насколько точна диагностика острого аппендицита по степени выраженности симптомов: гангренозного – 1, флегмонозного – 2, катарального – 3, другой абдоминальной патологии – 4. В разработку включены данные 100 историй болезни с тремя видами аппендицитов и из них 24 случая неподтвержденных. (решение см. стр.122, Жижин «Медицинская статистика»)

**4. Перечень вопросов для проверки исходного уровня знаний:**

- a. Понятие многомерных статистических методов.
- b. Кластерный анализ и его классификация.
- c. Дискриминация. Алгоритм дискриминации. Основные вопросы дискриминации.
- d. Факторизация. Концепция факторного анализа.

**5. Перечень вопросов для проверки конечного уровня знаний:**

1. Перечислите случаи использования многомерных статистических методов и преимущества их использования на практике.
2. Этапы применения кластерного анализа.
3. Перечислите приемы использования кластерного анализа в медико-биологических исследованиях.
4. Дискриминантный анализ. В чем разница между кластерным и дискриминантным анализом?
5. Описать модель факторного анализа.
6. Перечислить и описать коротко базовые понятия факторного анализа.

**6. Хронокарта учебного занятия:**

Организационный момент – 10 мин.

Разбор темы – 40 мин.

Текущий контроль (тестирование, практическая работа) - 90 мин.

Подведение итогов занятия – 10 мин.

**7. Самостоятельная работа аспиранта.**

Использование нейронных сетей в медицинской статистике.

**8. Перечень учебной литературы к занятию:**

2. Есауленко И.Э., Семенов С.Н. Основы практической информатики в медицине; Воронеж, 2005.
3. Жижин К. С. Медицинская статистика; Ростов н/Д, 2007.

