



Лекция 6

Методы анализа данных. Основы математической статистики.



Составитель: доц. Космачева И.М.

ОСНОВНЫЕ НАПРАВЛЕНИЯ ПРИМЕНЕНИЯ МАТЕМАТИКО-СТАТИСТИЧЕСКИХ МЕТОДОВ В МЕДИЦИНЕ И ЗДРАВООХРАНЕНИИ:

- 1) Наиболее эффективный сбор данных и обобщение полученных результатов.
- 2) Сравнение и определение достоверности различия двух и более групп результатов.
- 3) Изучение взаимосвязи между факторами и явлениями;
- 4) Анализ динамики процессов.
- 5) Анализ прогностических факторов.
- 6) Анализ зависимостей между факторами.
- 7) Анализ выживаемости (анализ времени жизни в одной и более группах, сравнение групп по времени жизни, оценка влияния факторов на время жизни пациентов).
- 8) Вычисление необходимого объема выборки, анализ мощности критериев.
- 9) Прогнозирование исхода лечения.

СТАТИСТИЧЕСКИЕ ВЕЛИЧИНЫ

Ст
мат
по
Аб
явл
От
хар
кач

	А	В	С	Д	Е	Ф	Г	Н	И	Ж	К
1	заболеваемость на 2005 год										
2		абсолютное количество	относительное число (в%)	заболеваемость (на 1000 населения)							
3	инфекционные болезни	847	=B3/B\$21								
4	новообразование	1472									
5	болезни эндокринной системы	1419									
6	сахарный диабет	733									
7	ожирение	97									
8	болезни крови и кроветворных органов	315									
9	психические расстройства	2230									
10	болезни нервной системы	3100									
11	болезни органов глаза	2625									
12	болезни органов уха	2354									
13	болезни органов кровообращения	8965									
14	болезни органов дыхания	10283									
15	болезни органов пищеварения	4123									
16	болезни кожи и подкожной клетчатки	2562									
17	болезни костно-мышечной системы	3109									
18	болезни мочеполовой системы	2703									
19	беременность, роды	668									
20	травмы и отравления	1766									
21	Всего	49371									
22	Численность населения	43200									

СТАТИСТИЧЕСКИЕ ПАКЕТЫ

Статистический пакет - программный продукт, предназначенный для статистической обработки данных. Существуют специализированные статистические пакеты и другие пригодные для проведения статистических расчетов приложения.

- Зарубежные: *STATGRAPHICS, SPSS, SYSTAT, BMDP, SAS, CSS, STATISTICA, S-plus* и др.,
- Отечественные: *STADIA, ЭВРИСТА, МЕЗОЗАВР, ОЛИМП: Стат-Эксперт, Статистик-Консультант, САНИ, КЛАСС-МАСТЕР, Deductor Academic (basegroup.ru)* и др.
- *Mathcad, EXCEL*



DATA MINING – КЛАССЫ РЕШАЕМЫХ ЗАДАЧ

- Классификация
- Регрессия
- Кластеризация
- Ассоциация
- Последовательность



ЭТАПЫ ОПЕРАЦИИ С МЕДИКО-БИОЛОГИЧЕСКИМИ ДАННЫМИ

1. *Сбор и первичная обработка данных* – это накопление результатов исследований в том объеме, который задан условиями поставленной задачи или необходимостью принять адекватное решение. Исследователь располагает ограниченным числом наблюдений.
2. *Оценка эффективности измерения данных* – это определение степени точности и величины погрешности зарегистрированных сигналов и полученных данных.
3. *Сохранение данных* – это регистрация данных в виде твердых копий или на магнитных носителях.
4. *Формализация и стандартизация данных* – это сведение всех полученных данных к единой форме, которая должна соответствовать требованиям компьютерной обработки и обеспечивать сопоставимость всех данных между собою, а также доступность их для всех заинтересованных пользователей.

ЭТАПЫ ОПЕРАЦИИ С МЕДИКО-БИОЛОГИЧЕСКИМИ ДАНЫМИ

5. *Фильтрация и очищение данных* – это отсеивание лишних сигналов, обусловленных неточностью работы регистрирующих приборов, некорректно собранной информацией о состоянии изучаемого явления.
6. *Кодировка данных* – это унификация формы представления данных на бумажных или магнитных носителях.
7. *Сортировка данных* – это упорядочение данных по заданному признаку или совокупности их характеристик .
8. *Преобразование данных* – это изменение формы данных по заданному алгоритму или между различными типами носителей.
9. *Сжатие и архивация данных* – это уплотнение данных на носителях и организация их хранения, нередко связана с изменением их формы.
10. *Защита данных* – это приведение данных по специальному алгоритму к форме, которая недоступна для несанкционированного их использования (шифрование, или криптографическая обработка данных).
11. *Транспортировка данных* – это передача данных на расстояния с помощью механических или телекоммуникационных каналов связи.

ЭТАПЫ ОПЕРАЦИИ С МЕДИКО-БИОЛОГИЧЕСКИМИ ДАННЫМИ

- **Высокая точность измерения достигается при минимальных рандомизированных и систематических погрешностях.**
- **Систематические погрешности** - погрешности, соответствующие отклонению измеренного значения от истинного значения величины всегда в одну сторону (повышения или занижения). **При повторных измерениях погрешность остается прежней.**
- **Систематические ошибки** отражают неправильное выполнение исследований, ошибки в настройке аппаратуры, неправильную калибровку лабораторного оборудования (электронных приборов, весов, тонометров и др.), неадекватный подбор больных, нарушение технологии приготовления растворов в биохимических лабораториях, ошибки, допущенных в расчетах.
- **Возможность их избежать** — тщательно контролировать исправность медицинской аппаратуры, проводить регулярную поверку в специальных лабораториях, следить за правильностью выполнения диагностических и расчетных процедур, корректно выполнять эти расчеты.

ЭТАПЫ ОПЕРАЦИИ С МЕДИКО-БИОЛОГИЧЕСКИМИ ДАННЫМИ

- ❑ **Случайные погрешности** - это погрешности, которые непредсказуемым образом меняют свое численное значение. Такие погрешности вызываются большим числом неконтролируемых причин, влияющих на процесс измерения (неровности на поверхности объекта, дуновение ветра, скачки напряжения и т.д.).

Рандомизированные или случайные, ошибки могут быть связаны:

- ❑ с неправильными (или недостаточно точными) заключениями врача-специалиста, проводящего исследования.
- ❑ с ограниченной точностью инструментов,
- ❑ вариабельностью измеряемого объекта (колебания биологических параметров человека во время исследования (суточные биоритмы), наводка по электросетям во время снятия биопотенциалов (ЭКГ, электроэнцефалография)).

Влияние случайных погрешностей может быть уменьшено при многократном повторении опыта.

ОЦЕНКА КАЧЕСТВА ИЗМЕРЕНИЙ

- ▣ ***Сходимость измерений*** - характеризует величину случайных ошибок. Критерий показывает, насколько близки друг к другу измерения, выполненные в одинаковых условиях, т.е. в одной и той же лаборатории и на одном и том же приборе.
- ▣ ***Воспроизводимость измерений*** - показывает, как близки между собою будут результаты измерений, выполненные в различных условиях, т.е. в различных лабораториях, на различных аппаратах и различными людьми.

Величину расхождения между измерениями, выполненными в одних и тех же условиях, обозначают как среднее квадратическое, или *стандартное, отклонение* (σ).

ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ ВЕРОЯТНОСТИ

- ▣ **Вероятность** – количественная мера объективной возможности появления события при реализации определенного комплекса условий.
- ▣ Вероятность события A обозначается как $p(A)$ и выражается в долях единицы или в процентах. **Мера вероятности** – диапазон ее числовых значений: от 0 до 1.
- ▣ **Случайное событие** – событие, которое при реализации определенного комплекса условий может произойти или не произойти.
- ▣ **Достоверное событие** - событие, которое при реализации определенного комплекса условий произойдет непременно.
Невозможное событие - событие, которое при реализации определенного комплекса условий не произойдет никогда. Его вероятность будет равна 0.

ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ ВЕРОЯТНОСТИ

- ▣ **Частота появления события (статистическая вероятность)** – это отношение числа случаев, в которых реализовался определенный комплекс условий (m), к общему числу случаев (n):

$$p(A) = m/n.$$

- ▣ Вероятность отсутствия события: $q = 1 - p$.
- ▣ **Случайная величина (СВ)** – величина, которая при реализации определенного комплекса условий может принимать различные значения.
- ▣ По своей природе СВ бывают **непрерывными** и **дискретными**.
- ▣ Если СВ принимает конечное число значений, **распределение задается функцией плотности вероятности** $P(X = x)$, ставящей каждому возможному значению x случайной величины X вероятность того, что $X = x$.
- ▣ Распределение может быть задано с помощью т.н. **функции распределения** $F(x) = P(X < x)$, определяющей для всех действительных x вероятность того, что случайная величина X принимает значения, меньшие x .

ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ ВЕРОЯТНОСТИ И МАТСТАТИСТИКИ

Если СВ принимает бесконечно много значений, распределение СВ в генеральной совокупности носит теоретический, идеальный характер, а ее выборочный аналог является эмпирическим распределением.

- **Генеральная совокупность** (Г.С. в англ. — *population*) — совокупность всех объектов (единиц), относительно которых наблюдатель намерен делать выводы при изучении конкретной проблемы.
- **Выборка** или **выборочная совокупность** — множество случаев (испытуемых, объектов, событий, образцов), выбранных с помощью определённой процедуры из генеральной совокупности для участия в исследовании.

ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ ВЕРОЯТНОСТИ И МАТСТАТИСТИКИ

Различают **бесповторные** (чаще используются) и **повторные выборки** (отобранный объект возвращается в Г.С. Перед выборкой следующего объекта).

Объём выборки — число случаев, включённых в выборочную совокупность.

Независимость выборок предполагает, что значения переменных в одной выборке не связаны со значениями в другой, с которой производится сравнение.

Пример независимых выборок - Показатели артериального давления в группе мужчин по сравнению с группой женщин.

Пример зависимых выборок — Измеренные показатели артериального давления у пациентов в 9 утра и у них же в 9:15.

ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ ВЕРОЯТНОСТИ И МАТСТАТИСТИКИ

- **Репрезентативность** - это способность выборочной совокупности как количественно, так и качественно отражать свойства генеральной совокупности.
- **Количественная** репрезентативность достигается достаточностью числа наблюдений.
- **Качественная** - соответствием признаков единиц наблюдения в выборочной и генеральной совокупностях.
- **Репрезентативная выборка** повторяет пропорции генеральной совокупности, представляет собой меньшую по размеру модель генеральной совокупности за счет случайного отбора объектов.

ПРИЕМЫ ПОЛУЧЕНИЯ РЕПРЕЗЕНТАТИВНОЙ ВЫБОРКИ

- **Случайный (рандомизированный) отбор.** Предполагает обеспечение таких условий, чтобы каждый член генеральной совокупности имел равные с другими шансы попасть в выборку.
- **Отбор по свойствам генеральной совокупности.**
 1. Предварительное определение качеств, которые могут влиять на изменчивость изучаемого свойства (пол, наследственность и т.д.).
 2. Определение процентного соотношения численности различающихся по этим качествам групп в генеральной совокупности и обеспечение идентичного процентного соотношения соответствующих групп в выборке.
 3. В каждую подгруппу выборки испытуемые подбираются по принципу простого случайного отбора.

ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ ВЕРОЯТНОСТИ И МАТСТАТИСТИКИ

- Любое значение параметра, вычисленное на основе ограниченного числа наблюдений, непременно содержит элемент случайности. Результат эксперимента - случайная величина.
- Такое приближенное, случайное значение называется **оценкой параметра**.
- **Оценкой параметра** называют функцию результатов наблюдений над случайной величиной (статистику), с помощью которой судят о значении параметра .
- $\hat{a}(N)$ – статистическая оценка параметра a по данным N опытов (прогонов).
- Генеральная совокупность характеризуется одним или несколькими параметрами: μ , σ^2 , σ и т.д.

ЗАКОН БОЛЬШИХ ЧИСЕЛ

- При достаточно **большом числе наблюдений** случайные отклонения взаимно погашаются и проявляется **устойчивость некоторых параметров**, которая выражается в основной тенденции (закономерности).
- При **большом числе случайных величин их средний результат перестает быть случайным** и может быть предсказан с большой степенью определенности.

ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ ВЕРОЯТНОСТИ И МАТСТАТИСТИКИ

Общие рекомендации по численности выборки:

- Наибольший объем выборки необходим при разработке диагностической методики - от 200 до 1000-2500 человек.
- Если необходимо сравнивать 2 выборки, их общая численность должна быть не менее 50 человек; численность сравниваемых выборок должна быть приблизительно одинаковой.
- Если изучается взаимосвязь между какими-либо свойствами, то объем выборки должен быть не меньше 30-35 человек.
- Чем больше изменчивость изучаемого свойства, тем больше должен быть объем выборки.
- Поэтому изменчивость можно уменьшить, увеличивая однородность выборки, например, по полу, возрасту и т. д. При этом, естественно, уменьшаются возможности генерализации выводов.

КЛИНИЧЕСКИЕ МЕТОДЫ ИССЛЕДОВАНИЯ

- **Методы клинического исследования** - опрос пациента, лабораторное исследование, функциональная диагностика с использованием сложных систем типа МРТ, ПЭТ.
- Любое исследование начинается с постановки цели, статистической гипотезы, выбора подходящего дизайна исследования.
- К наиболее распространенным дизайнам клинических исследований относят:



ДИЗАЙН КЛИНИЧЕСКИХ ИССЛЕДОВАНИЙ

- ▣ **Исследование случай-контроль.** Относится к ретроспективным исследованиям, т.е. таким, в которых ставится задача понять какие события в прошлом повлияли на состояние пациента в данный момент. При таком клиническом исследовании сравнивают людей, имеющих какое-то заболевание, с людьми, не страдающими им. Это делается для выявления зависимости между тем или иным клиническим исходом и воздействовавших ранее разнообразных факторов риска. Например, как курение повлияло на развитие раковой опухоли.
- ▣ **Когортное исследование.** Относится к классу проспективных исследований, т.е. где изучается воздействие факторов риска до начала заболевания, а не после, как в ретроспективном. В этом исследовании формируются 2 группы. Одну группу подвергают воздействию вредных факторов, а вторую нет. А потом у этих групп сравнивают клинические исходы. Такой тип клинического исследования может быть очень длительным.
- ▣ **Исследование перекрестного дизайна.** При перекрестном исследовании набирается группа пациентов. Пациенты рандомизируются так, что одна группа получает лечебный препарат, а другая плацебо. Через какое-то время эти группы меняются местами так, что та, которая получала лечебный препарат стала получать плацебо, а та, что получала плацебо, стала получать лечебный препарат. Фиксируются различия.
- ▣ **Рандомизированное контролируемое исследование.** Используется для изучения взаимосвязей между лечением и исходом заболевания. При таком дизайне клинического исследования пациенты случайным образом разбиваются на группы. Ни исследователи, ни пациенты не знают какое лечение проводится в каждой из групп. После исследования или на какой-то фазе смотрят различия между группами.

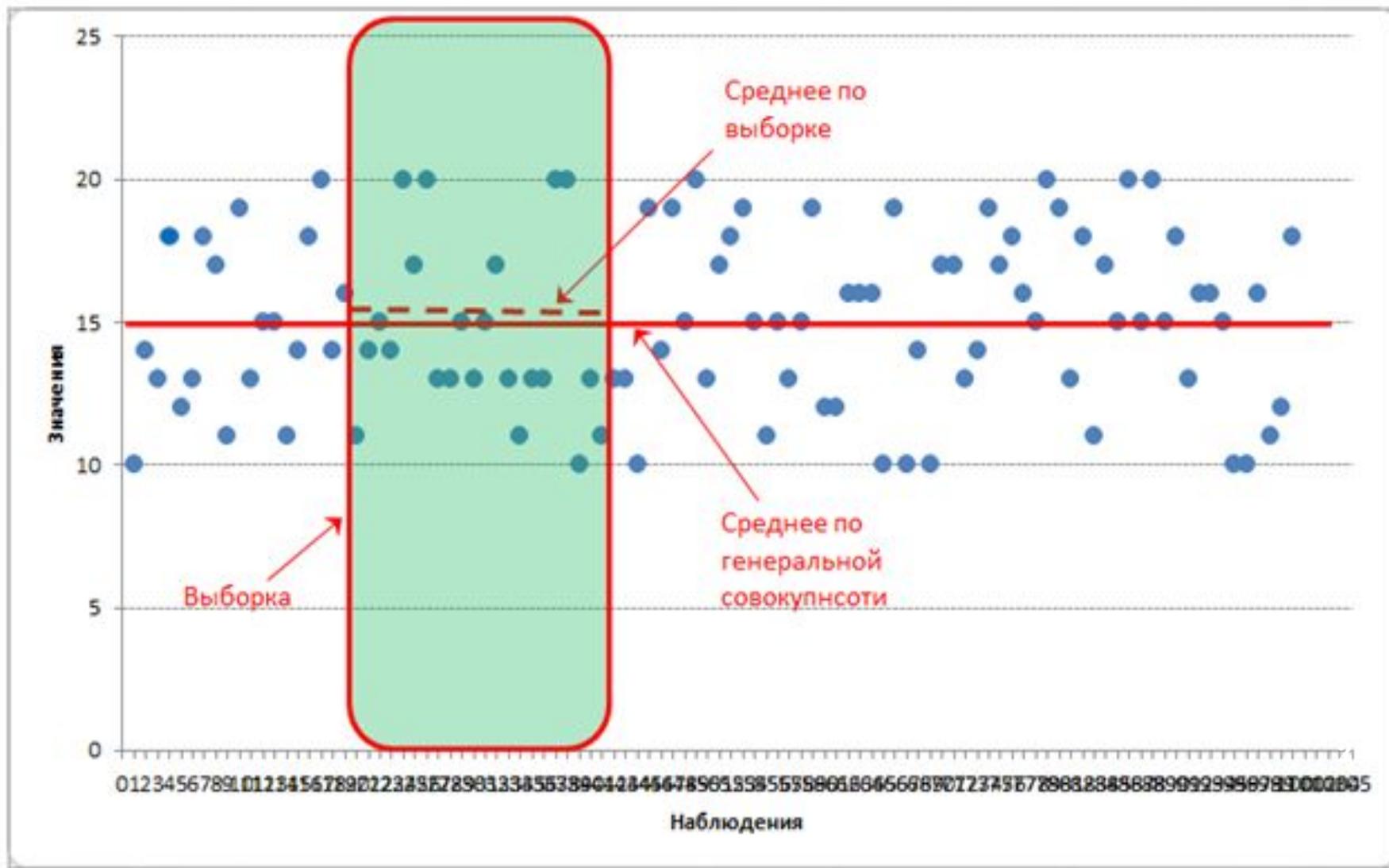


ОСНОВНЫЕ СТАТИСТИКИ

- ▣ *Выборочное среднее \bar{x} – оценка математического ожидания, среднее арифметическое элементов выборки.*
- ▣ *Выборочная дисперсия S^2 – среднее квадратов отклонения элементов выборки от выборочного среднего, является оценкой дисперсии, характеризует разброс выборочных значений.*
- ▣ *Стандартное отклонение S – корень из дисперсии.*
- ▣ *Коэффициент вариации – отношение выборочного среднего квадратического отклонения к выборочной средней, характеризует рассеяние вне зависимости от размерности вариант .*
- ▣ *Размах варьирования- разность между наибольшей и наименьшей вариантами.*
- ▣ *Медиана Me .*
- ▣ *Мода Mo .*
- ▣ *Коэффициент эксцесса E .*
- ▣ *Коэффициент асимметрии A .*
- ▣ *Процентиль.*

ОСНОВНЫЕ СТАТИСТИКИ

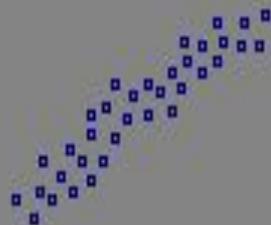
Отклонение выборочной средней от генеральной показано на рисунке.



ОСНОВНЫЕ СТАТИСТИКИ

$$\begin{aligned}\tilde{K}_{xy} &= \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) = \\ &= \frac{1}{N-1} \left[\sum_{i=1}^N x_i y_i - \frac{1}{N} \sum_{i=1}^N x_i \sum_{i=1}^N y_i \right]\end{aligned}$$

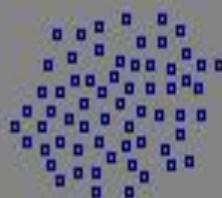
Y



X

Сильная положительная корреляция (KR стремится к 1)

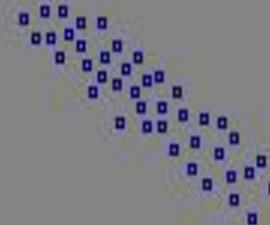
Y



X

Корреляция отсутствует (KR близко к 0)

Y



X

Сильная отрицательная корреляция (KR стремится к -1)

ОСНОВНЫЕ СТАТИСТИКИ

- Корреляция** (от лат. **correlatio**), **корреляционная зависимость** - взаимозависимость двух или нескольких случайных величин (взаимосвязь между ростом и весом детей, между стажем работы и производительностью труда).

Значение	Интерпретация
от 0 до 0,3	очень слабая
от 0,3 до 0,5	слабая
от 0,5 до 0,7	средняя
от 0,7 до 0,9	высокая
от 0,9 до 1	очень высокая
- Статистическая зависимость** – это зависимость, при которой изменение одной СВ влечет изменение распределения другой СВ.
- Корреляционная зависимость** - при изменении значения одной переменной происходит закономерное изменение (уменьшение или увеличение) среднего значения другой(-их) переменной(-ых).
- Корреляционная зависимость - **вероятностная зависимость между показателями, которая проявляется только в массе наблюдений.**
- Корреляционная зависимость отражает только **взаимосвязь между переменными** и не говорит о причинно-следственных связях (если величины независимы, то коэффициент корреляции равен нулю, обратное не всегда верно).
- Коэффициент корреляции может варьировать в пределах от **-1** (отрицательная корреляция) до **+1** (положительная корреляция).



ОСНОВНЫЕ СТАТИСТИКИ

Формулы для случая, если значения признака выборки объема N различны имеют соответствующие частоты , n_i

причем их сумма равна N

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i \cdot n_i = \bar{x}_g$$

$$\sum_{i=1, k} n_i = N$$

$$\bar{D} = \frac{1}{N-1} \sum_{i=1}^N n_i \cdot (x_i - \bar{x}_g)^2 = \frac{1}{N-1} \left(\sum_{i=1}^N n_i \cdot x_i^2 - \frac{1}{N} \left(\sum_{i=1}^N n_i \cdot x_i \right)^2 \right) =$$

$$= \frac{N}{N-1} D_g = s^2$$



ОСНОВНЫЕ ХАРАКТЕРИСТИКИ

ВАРИАЦИОННОГО РЯДА

- ▣ **Медиана** - это такое значение признака, которое делит упорядоченное (ранжированное) множество данных пополам так, что одна половина всех значений оказывается меньше медианы, а другая - больше.

Если данные содержат нечетное число значений (8, 9, **10**, 13, 15), то медиана есть центральное значение;

Если данные содержат четное число значений (5, **8**, **9**, 11), то медиана есть точка, лежащая посередине между двумя центральными значениями.

- ▣ **Мода** - это такое значение из множества измерений, которое встречается наиболее часто. Когда два соседних значения встречаются одинаково часто и чаще, чем любое другое значение, мода есть среднее этих двух значений.

ОСНОВНЫЕ СТАТИСТИКИ

$$K_{\text{выб}} = \frac{\sum n_{xy}xy - n\bar{x}\bar{y}}{n\bar{\sigma}_x\bar{\sigma}_y}$$

Коэффициент корреляции

$$\bar{a} = \frac{\bar{\mu}_3}{\bar{\sigma}^3}$$

Коэффициент асимметрии

$$\bar{\mu}_3 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x}_g)^3$$

Выборочный центральный момент третьего порядка

$$\bar{\mu}_4 = \frac{\sum_{i=1}^N (x_i - \bar{x})^4 \cdot n_i}{N\bar{s}^4} - 3$$

Выборочный эксцесс

$$D_{\sigma} = \frac{1}{N} \sum_{i=1}^N n_i \cdot (x_i - \bar{x}_g)^2, \quad \bar{s} = \sqrt{\frac{1}{N} \sum_{i=1}^N n_i \cdot (x_i - \bar{x}_g)^2}$$

Дисперсия и среднеквадратическое отклонение



ОСНОВНЫЕ СТАТИСТИКИ

- Если значения переменной распределены несимметрично относительно центра, то группы лучше описывать с помощью **медианы и квантилей (процентилей, квартилей, децилей)**.
- **Квантилью x_p** (p -квантилью, квантилью уровня p) случайной величины, имеющей функцию распределения $F_x(x)$, называют решение x_p уравнения $F_x(x) = p$. Для некоторых p уравнение $F_x(x) = p$ может иметь **несколько решений**, для некоторых - **ни одного**.
- Квантили, наиболее часто встречающиеся в практических задачах, имеют свои названия:
 - медиана* - квантиль уровня 0.5;
 - нижняя квартиль* - квантиль уровня 0.25;
 - верхняя квартиль* - квантиль уровня 0.75;
 - децили* - квантили уровней 0.1, 0.2, ..., 0.9;
 - процентили* - квантили уровней 0.01, 0.02, ..., 0.99.

ОСНОВНЫЕ СТАТИСТИКИ

- ▣ **Процентиль на уровне P** - это такое значение, ниже которого расположено P процентов наблюдений данной переменной. Например, значение 50-й процентиля указывает, что 50% значений располагается ниже этого уровня.
- ▣ **Коэффициент эксцесса E** - характеризует «островерхость» гистограммы или полигона по сравнению с кривой Гаусса нормального распределения.
- ▣ **Коэффициент асимметрии A** - характеризует степень симметричности гистограммы или полигона по сравнению с кривой Гаусса. Если коэффициенты асимметрии и эксцесса близки к нулю, то форму распределения можно считать близкой нормальному.

ИНТЕРВАЛЬНЫЕ ОЦЕНКИ ПАРАМЕТРОВ

- *Интервальной оценкой* параметра θ называется числовой интервал (a, b) который с заданной вероятностью p (*надежностью*) покрывает неизвестное значение параметра θ .
- *Величина доверительного интервала зависит от объема выборки (уменьшается с ростом n) и надежности p (увеличивается с ростом p).*
- Такой интервал (a, b) называется *доверительным*, а вероятность p *доверительной вероятностью*. Вместо нее часто задают величину $\alpha = 1 - p$, называемую *уровнем значимости*.

p : 0,95; 0,99; 0,999

α : 0,05; 0,01; 0,001

СТАТИСТИЧЕСКИЕ ГИПОТЕЗЫ

Статистическая гипотеза – некоторое предположение о свойствах генеральной совокупности, из которой взята выборка (о виде или параметрах неизвестного закона распределения).

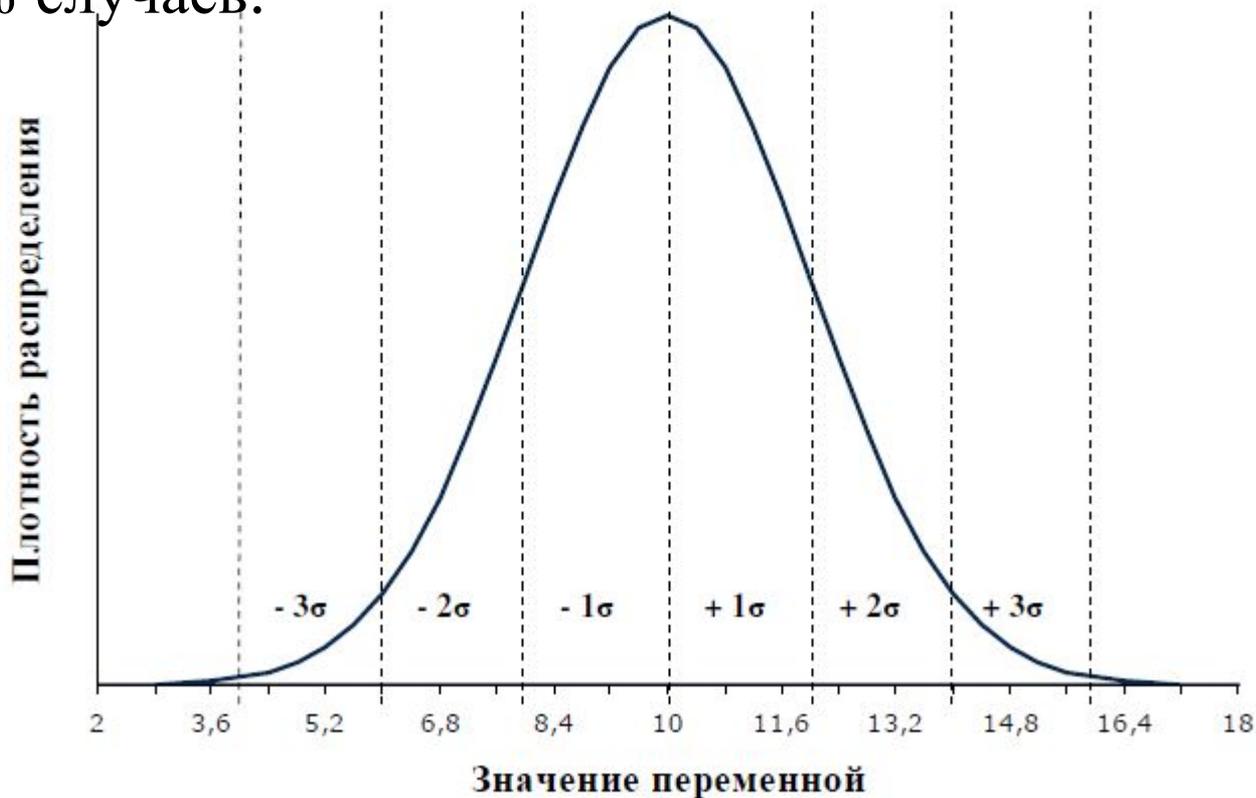
- ▣ **нулевая гипотеза H_0 и альтернативная гипотезы H_1 .**
- ▣ **параметрическая (мат. ожидания, дисперсия, т.д.) и непараметрическая (подчинение СВ нормальному закону, закону Пуассона и т.д.)**
- ▣ **простая и сложная.**

ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ ВЕРОЯТНОСТИ

Основными типами распределений признаков являются «нормальным», но в практике медицинского эксперимента оно встречается лишь в 30% случаев.

- дискретные (биномиальное, распределение Пуассона, распределение Бернулли)
- непрерывные (нормальное, логнормальное, экспоненциальное, гамма-распределение, бета-распределение, распределение Вейбулла)

В состав статистических параметров распределений для асимметричных распределений



усса),
нципа
й.
гально
чаях
а их

Составитель: доц. Космачева И.М.

ВЕРОЯТНОСТЬ ОШИБКИ В ВЫВОДАХ

H_0	верна	неверна
отвергаем	1-го рода α	правильное решение
принимаем	правильное решение	2-го рода β

СТАТИСТИКА В МЕДИЦИНЕ

При проведении медицинского исследования, как и при выполнении любого диагностического теста, могут встретиться следующие варианты врачебных заключений.

1. Интерпретация “положительный” у больных с наличием заболевания. Это *истинно положительные случаи (TP, true positive)*.

2. Интерпретация “отрицательный” у пациентов без заболевания. Это – *истинно отрицательные случаи (TN, true negative)*.

3. Интерпретация “положительный” у пациентов с отсутствием заболевания (у здоровых лиц). Это *ложноположительные случаи (FP, false positive)* (“ложные тревоги”. В практической медицине риск совершить такого рода ошибки каждой врач устанавливает по-своему. Однако общепринятой критической величиной ложной тревоги является величина β риска, равная 0,05, или 5%. Этот уровень обозначается в медицинской статистике как уровень статистически значимого показателя.

Он обозначается как $p \leq 0,05$. Практически это означает, что врач ошибается в 1 случае из 20.

4. Интерпретация “отрицательный” у больных с заболеваниями. Это *ложноотрицательные случаи*.

Как ложноположительные, так и ложноотрицательные ошибки представляют собою совокупность ошибок диагностики и выдвигают основное требование перед врачом-диагностом – их минимизации.

СТАТИСТИКА В МЕДИЦИНЕ

Для характеристики информативности диагностического метода исследования служат объективные параметры, именуемые *операционными характеристиками теста*. К операционным характеристикам метода диагностики относятся:

- *чувствительность* (Se , *sensitivity*),
- *специфичность* (Sp , *specificity*).

Чувствительность (Se) – это пропорция правильных положительных результатов теста среди всех больных. Определяется по формуле:

$$Se = TP \cdot 100\% / D_+$$

где Se – чувствительность; TP – верно положительные случаи; D_+ – число больных искомым заболеванием.

СТАТИСТИКА В МЕДИЦИНЕ

- **Чувствительность** априори показывает, какова будет доля больных, у которых данное исследование даст положительный результат.
- Чем выше чувствительность теста, тем чаще с его помощью будет выявляться заболевание, тем, следовательно, он более эффективен.
- Если высокочувствительный тест оказывается отрицательным, то наличие заболевания маловероятно. Поэтому их следует применять для **исключения заболеваний**. Тесты с высокой чувствительностью рекомендуется применять на ранних этапах диагностического процесса, когда требуется сузить круг предполагаемых заболеваний.
- Но высокочувствительный тест дает много “ложных тревог”, что требует дополнительных финансовых затрат на дальнейшее обследование.

СТАТИСТИКА В МЕДИЦИНЕ

- ❑ **Специфичность (Sp)** – это пропорция правильных отрицательных результатов теста среди здоровых пациентов. Данный показатель определяется по формуле

$$Sp = TN \cdot 100\% / D_-$$

где Sp – специфичность; TN – истинно отрицательные случаи; D_- – здоровые пациенты.

- ❑ Определив специфичность, можно априори предполагать, какова доля здоровых лиц, у которых это исследование даст отрицательный результат.
- ❑ Чем выше специфичность метода, тем надежнее с его помощью **подтверждается заболевание**, тем, следовательно, он более эффективен.
- ❑ Тестирование эффективно на втором этапе диагностики, когда круг предполагаемых заболеваний сужен и необходимо с большой уверенностью доказать наличие болезни.
- ❑ Отрицательным фактором высокоспецифичного теста является тот факт, что его использование сопровождается весьма значительным числом пропусков заболевания.

ВЕРОЯТНОСТЬ ОШИБКИ В ВЫВОДАХ

- Вероятность α допустить ошибку 1-го рода (отвергается верная), называется **уровнем значимости критерия**.

$\alpha: 0,01; 0,05; 0,1$

- Вероятность не допустить ошибку 2-го рода (принимается ложная) **$(1-\beta)$** называется **мощностью критерия** . Т.е. вероятность отвергнуть гипотезу, когда она неверна.
- Принятые в биомедицинской статистике жесткие рамки, определяющие долю ложноположительных результатов **не более 5%** - необходимость, так как на основании результатов медицинских исследований внедряется либо отклоняется новое лечение, а это вопрос жизни многих тысяч людей, **мощностью критерия** не менее **80%**.

ПРОВЕРКА СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

Критериальная (тестовая) проверка - когда выборочная совокупность данных параметрически (своей функцией распределения) сравнивается с генеральной совокупностью или другой выборкой.

Проверка гипотезы основывается на вычислении некоторой случайной величины — критерия, точное или приближенное распределение которого известно.

$F_{\text{набл}}$ - выборочная характеристика (статистика)

$F_{\text{кр}}$ - критическое значение точного или приближенного распределения СВ

ПРОВЕРКА СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

$F_{\text{наб}} \leq F_{\text{кр}}$, гипотеза H_0 принимается.

Принятие гипотезы H_0 следует расценивать как правдоподобное, не противоречащее опыту утверждение.

$F_{\text{наб}} > F_{\text{кр}}$, гипотеза H_0 отвергается.

ПРОВЕРКА СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

Задачи тестирования гипотез можно разделить на две группы:

- Первая группа отвечает на вопрос, имеются ли различия между группами по уровню некоторого показателя

Например, различия в уровне печеночных трансаминаз у пациентов с гепатитом и здоровых людей.

- Вторая группа позволяет доказать наличие связи между двумя или более показателями, например, функции печени и иммунной системы.

В практическом плане задачи из первой группы можно разделить на два подтипа:

- сравнение показателя только в двух группах(здоровые и больные, мужчины и женщины);
- сравнение трех и более групп (изучение разных доз препарата).

ПРИМЕРЫ

1. Сравнение среднего результата одной серии экспериментов со средним результатом другой серии.

$$H_0 : \bar{X} = \bar{Y}$$

2. Сравнение точности приборов, инструментов, методов измерений и т.д.

$$H_0 : D_x = D_y$$

3. Сравнение долей признака в двух совокупностях

$$H_0 : P_0 = P_1$$

СРАВНЕНИЕ 2-Х СРЕДНИХ

$$H_0 : M(X) = M(Y)$$

X и Y - нормальные генеральные совокупности

\bar{x} и \bar{y} выборочные средние

Если $|F_{\text{наб}}| < F_{\text{кр}}$, H_0 принимается

$$F_{\text{набл}} = \frac{\bar{x} - \bar{y}}{\sqrt{D(X)/n + D(Y)/m}}$$

$$\Phi_{F_{\text{кр}}} = (1 - \alpha) / 2,$$

$$\Phi(t) = \frac{1}{\sqrt{2 \cdot \pi}} \int_0^t e^{-\frac{x^2}{2}} dx$$

СПАСИБО ЗА ВНИМАНИЕ