Введение в биостатистику

Сафонова В.Р.
Ханты-Мансийск, 2013

В медицине и здравоохранении часто используются, сознательно или неосознанно, различные статистические концепции при принятии решений по таким вопросам как:

- оценка состояния здоровья и его прогноз;
- выбор стратегии и тактики профилактики и лечения;
- оценка отдаленных результатов и выживаемости.

Статистика!!!

$$SS_{\text{total}} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} [(x_{ij} - \bar{x}_{.j}) + (\bar{x}_{.j} - \bar{x}_{..})]^2$$

$$= \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_{.j})^2 + 2 \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_{.j})(\bar{x}_{.j} - \bar{x}_{..})$$

$$+ \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\bar{x}_{.j} - \bar{x}_{..})^2$$

…..НУ И ЧТО?

СТАТИСТИКА

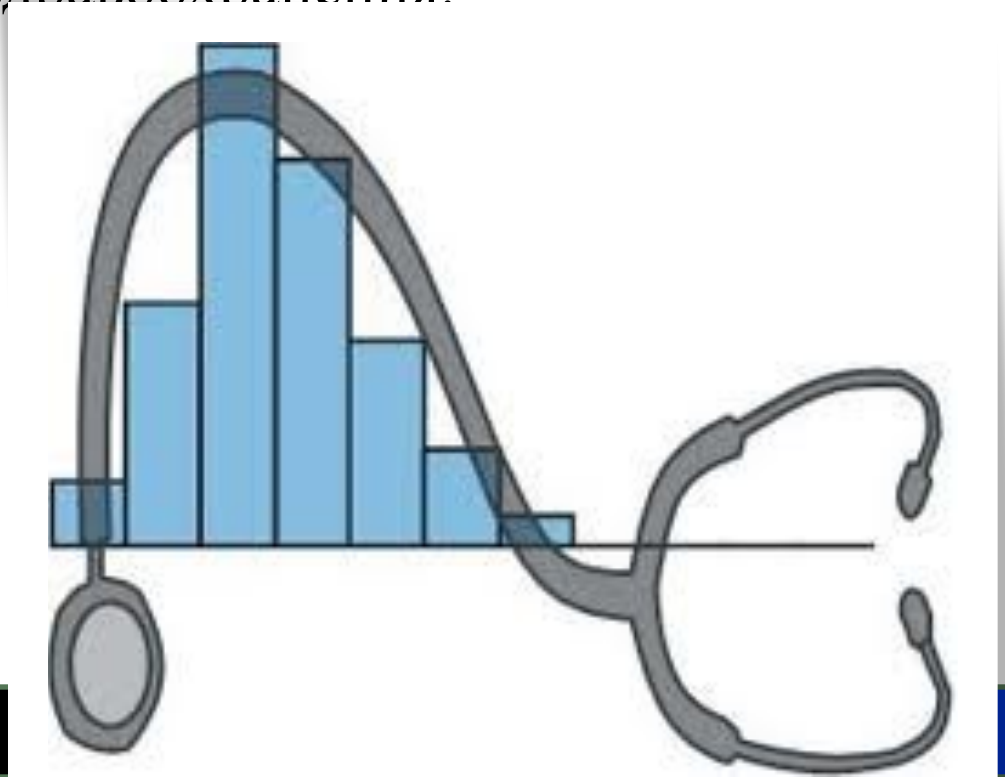
- это инструмент для анализа экспериментальных данных и результатов популяционных исследований;
- это язык с помощью которого исследователь сообщает полученные им результаты и благодаря которому он понимает медико-статистическую информацию;
- это элемент доказательной медицины;
- это база обоснования принятия управленческих решений.

СТАТИСТИКА

1. **Наука,** изучающая количественные закономерности материальных явлений в неразрывной связи с их качественной стороной.
2. **Точная** наука, изучающая методы сбора, обработки, систематизации, анализа и интерпретации данных, которые описывают массовые действия, явления и процессы.
3. **(от лат. status — состояние дел) наука,** сочетающая учет и анализ, фиксирующая, систематизирующая и изучающая показатели наиболее типичных, массовых экономических процессов и их изменение во времени.

БИОСТАТИСТИКА

приложение общей теории статистики для решения научно-практических проблем в области биологии, медицины и здравоохранения.



СТАТИСТИКА (Statistics)- наука о сборе, представлении и анализе данных.

БИОСТАТИСТИКА - статистическая наука (statistics) в приложении к живому миру. Включает в себя демографию, эпидемиологию и организацию клинических испытаний. Синоним - биометрия.

Oxford Dictionary of Statistics, 2002

ВЕРОЯТНОСТЬ

количественная мера объективной возможности появления события при реализации определенного комплекса условий.

Вероятность события А обозначается как p(A) и выражается в долях единицы или в процентах.

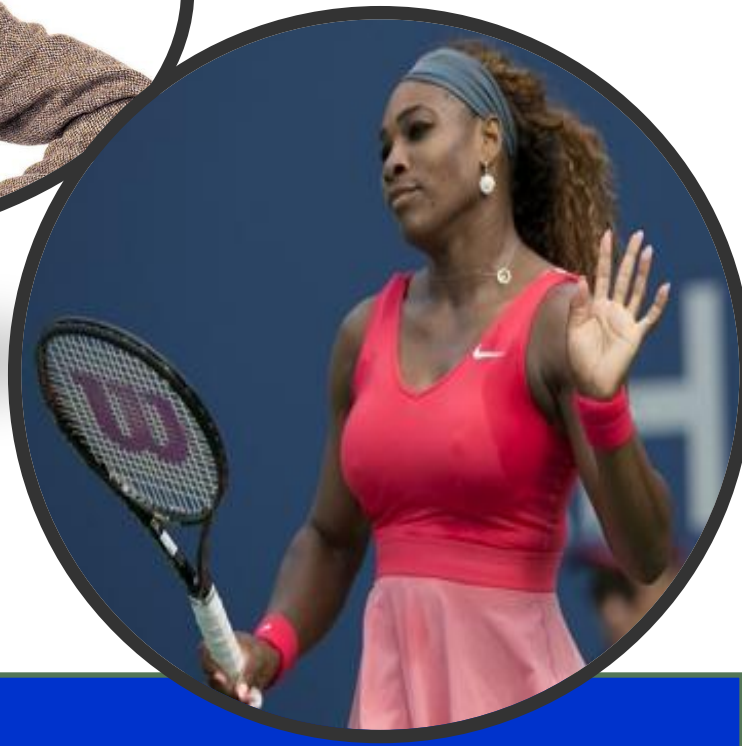
Мера вероятности – диапазон ее числовых значений: от 0 до 1 или от 0 до 100%.

ДИЛЕММА НЕРЕШИТЕЛЬНОГО ВЛЮБЛЕННОГО

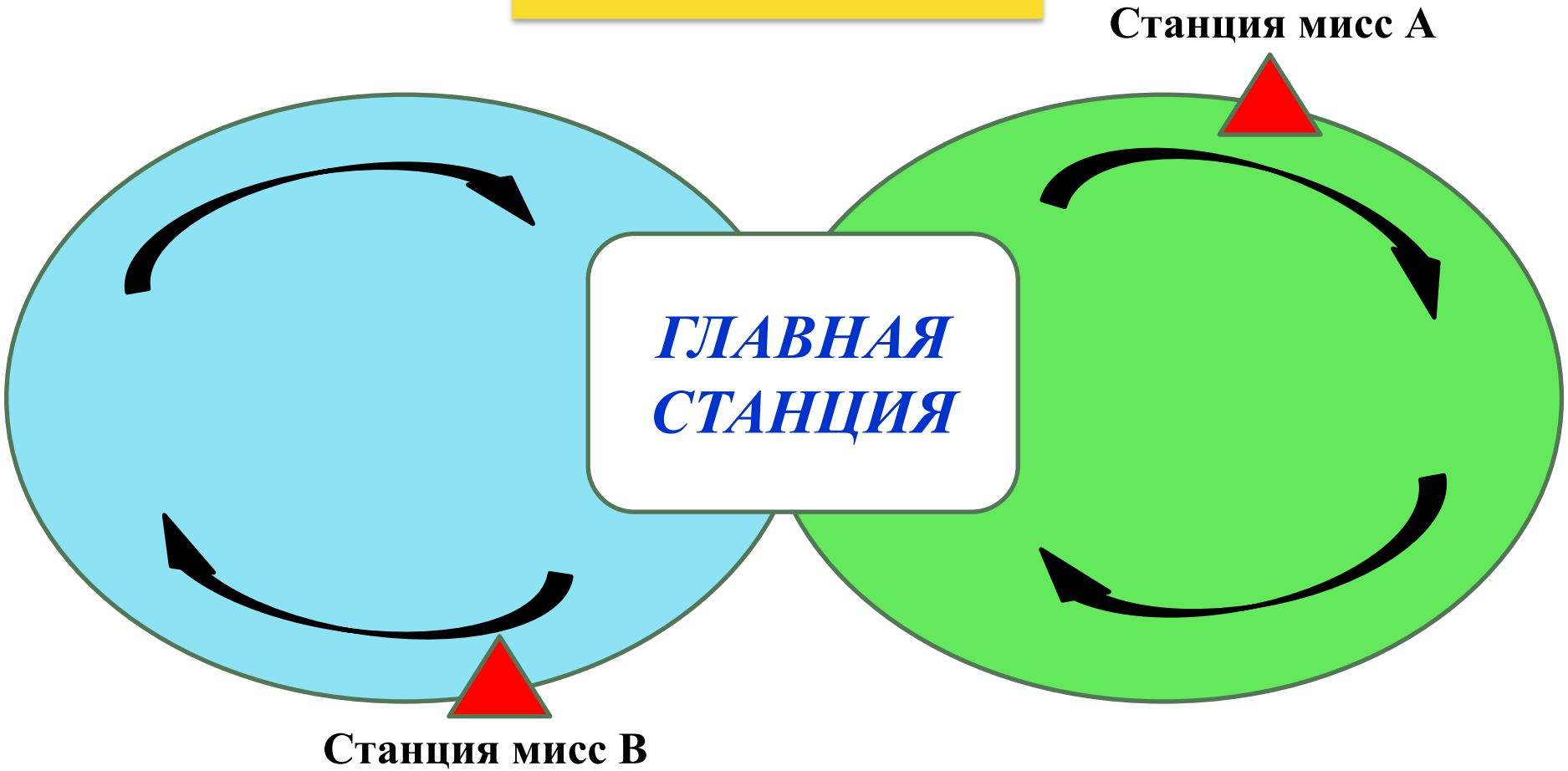
МИСТЕР Z

МИСС А

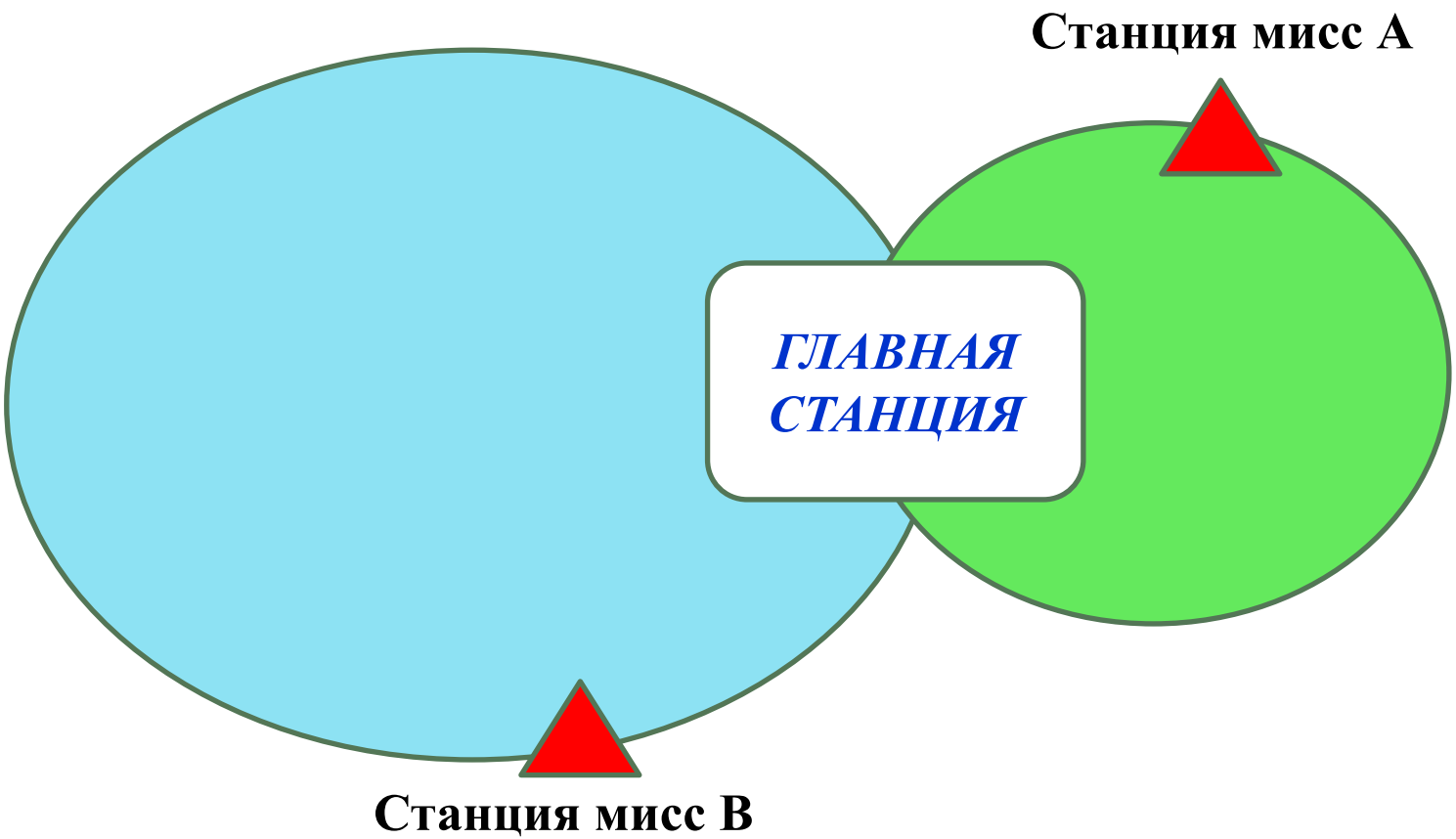
МИСС В



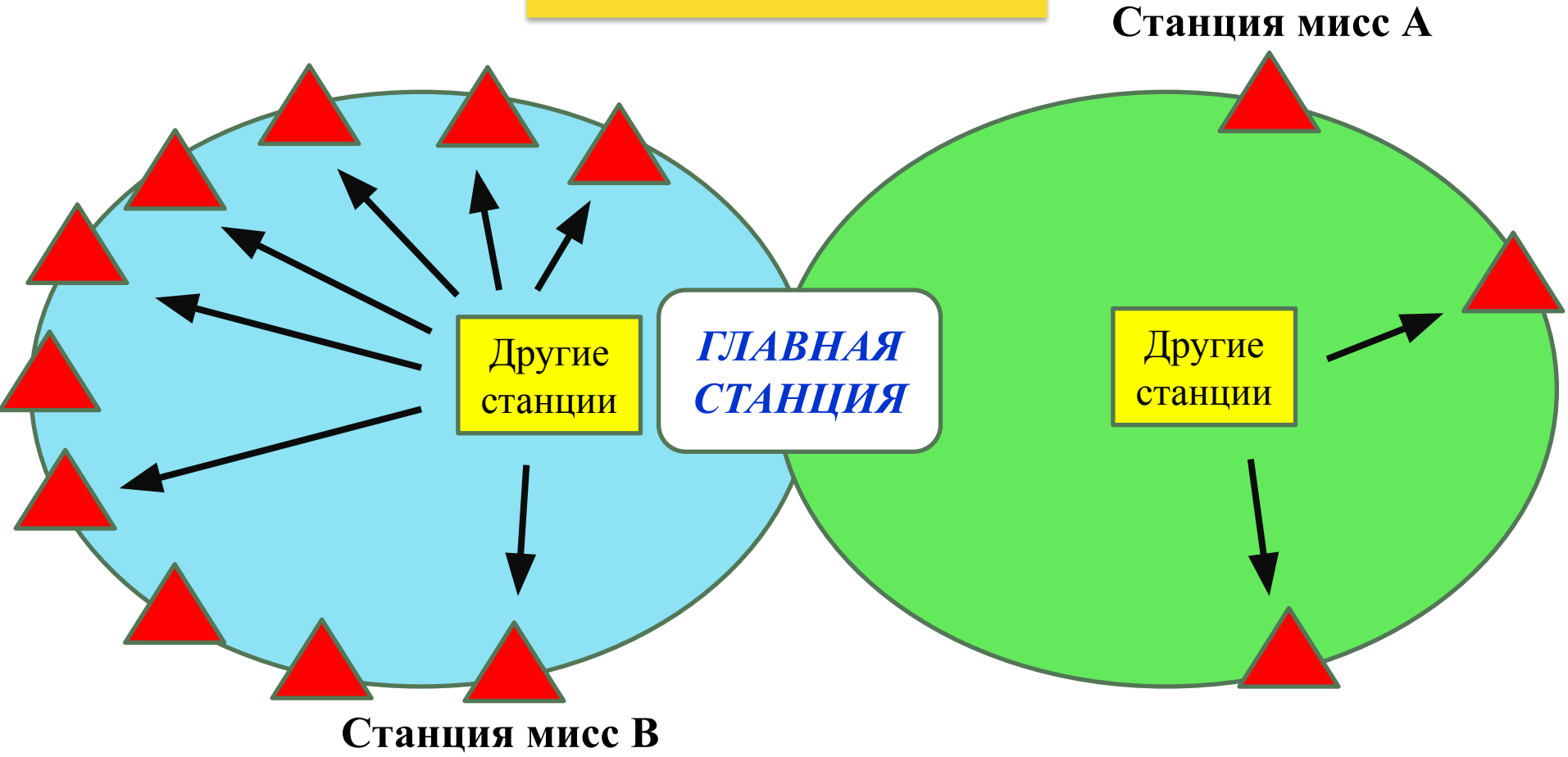
ОФИС
МИСТЕРА Z



ОФИС МИСТЕРА Z



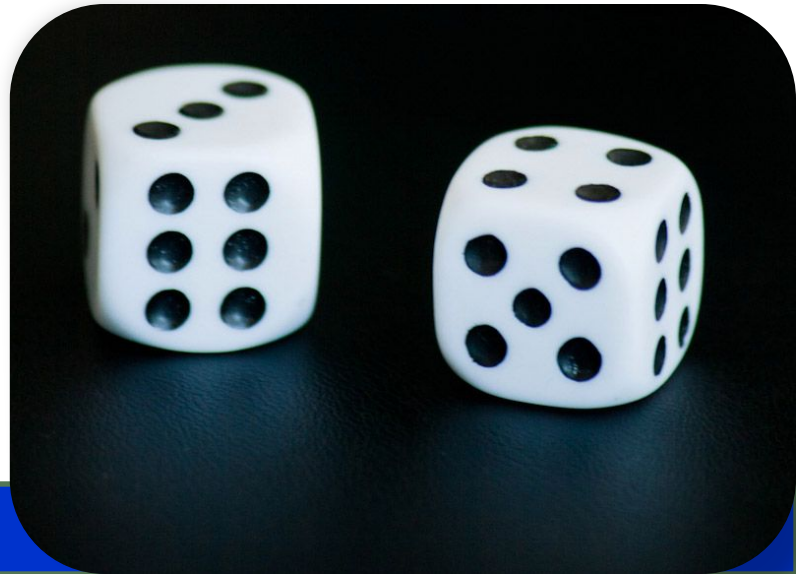
ОФИС МИСТЕРА Z



СЛУЧАЙНОЕ СОБЫТИЕ

событие, которое при реализации определенного комплекса условий может произойти или не произойти.

Его вероятность будет находиться в пределах $0<p(A)<1$ или $0<p(A)<100\%$.



ДОСТОВЕРНОЕ СОБЫТИЕ

событие, которое при реализации определенного комплекса условий произойдет непременно.
Его вероятность будет равна 1 или 100%.

НЕВОЗМОЖНОЕ СОБЫТИЕ

событие, которое при реализации определенного комплекса условий не произойдет никогда.
Его вероятность будет равна 0.

ЧАСТОТА ПОЯВЛЕНИЯ СОБЫТИЯ

(СТАТИСТИЧЕСКАЯ ВЕРОЯТНОСТЬ)

это отношение числа случаев, в которых реализовался определенный комплекс условий (m), к общему числу случаев (n):

$$\mathbf{p(A)=m/n}$$

Вероятность события: $q=1-p$.

ШАНС

это отношение вероятности того, что событие произойдет к вероятности того, что событие не произойдет.

ОТНОШЕНИЕ ШАНСОВ

(ODDS RATIO)

это отношение шансов для первой группы объектов к отношению шансов для второй группы объектов.

ПРАВИЛО СЛОЖЕНИЯ ВЕРОЯТНОСТЕЙ

Если два события, A и B, взаимоисключающие, несовместимые, то вероятность события A или B равна сумме их вероятностей:

P(A или B)= p(A)+p(B)

ПРАВИЛО УМНОЖЕНИЯ ВЕРОЯТНОСТЕЙ:

Если два события, A и B, независимы (т.е. возникновение одного события не влияет на возможность появления другого), то вероятность того, что оба события произойдут, равна произведению вероятности каждого:

$$P(A \text{ и } B)=p(A)*p(B)$$

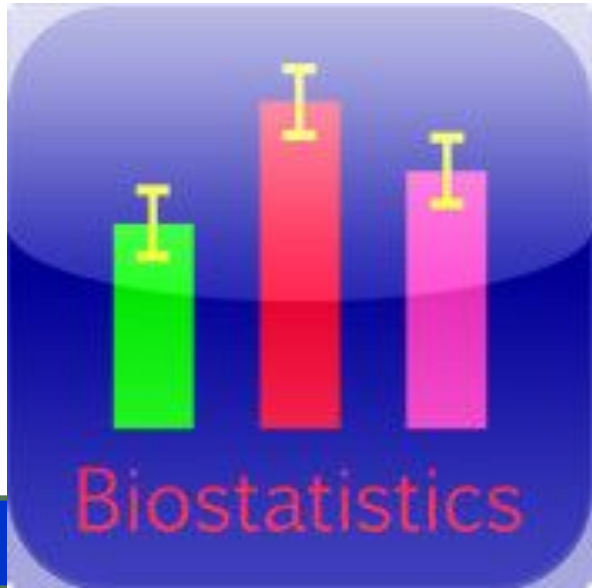
СЛУЧАЙНАЯ ВЕЛИЧИНА

величина, которая при реализации определенного комплекса условий может принимать различные значения.

Закон больших чисел:
при достаточно большом числе наблюдений случайные отклонения взаимно погашаются и проявляется основная тенденция (закономерность).

Приступая к изучению основ статистического анализа необходимо выделить два основных этапа:

- описание полученного в ходе исследования массива данных
- анализ данных и проверка различных статистических гипотез



Основные направления применения математико-статистических методов в медицине и здравоохранении:

- Наиболее эффективный сбор данных и обобщение полученных результатов;
- Сравнение и определение статистически значимых различий (достоверных) между двумя и более группами результатов;
- Изучение взаимосвязи между факторами и явлениями;
- Анализ динамики процессов;
- Анализ прогностических факторов.

Прежде чем приступить к анализу данных и проверке различных гипотез:

- Сформулируйте вопрос, на который Вы хотите ответить с помощью статистического анализа.
- Выберите наиболее адекватный для ответа на данный вопрос статистический критерий или метод.
- Правильно интерпретируйте его результаты.

Анализ организации конкретного исследования и его результатов:

- оценить адекватность дизайна научного исследования решению той или иной проблемы эпидемиологии и общественного здоровья.
- Анализ технологии приведенного исследования.
- Оценка полученных результатов.
- практическое применение полученных результатов.

ЗНАНИЕ ВОЗМОЖНОСТЕЙ СТАТИСТИЧЕСКИХ МЕТОДОВ НЕОБХОДИМО КАЖДОМУ РАБОТАЮЩЕМУ В МЕДИЦИНЕ И ЗДРАВООХРАНЕНИИ.

Изучение статистики может пригодиться:

При прочтении научных публикаций

Важно понимать статистические исследования, проводимые в интересуемой области.

Для этого необходимо знать и владеть

- *статистической терминологией,*
- *статистической символикой,*
- *знать концепцию статистических процедур, используемых в исследовании.*

В собственной научной работе и клинической практике

Для проведения исследований необходимо уметь:

- *планировать эксперимент*
- *собирать данные*
- *анализировать данные*
- *делать статистические выводы и прогнозы*

Для понимании основ доказательной медицины

ПАКЕТЫ ПРИКЛАДНЫХ ПРОГРАММ:

SPSS (Statistical Package for Social Science)
SAS
STATA
STATISTICA
BIOSTATISTICA
Epilnfo
программа «R»

ПРИМЕРЫ КЛИНИЧЕСКИХ ИССЛЕДОВАНИЙ

- Изучение эффективности нового лекарства
- Оценка нового диагностического теста
- Сравнительный анализ схем ведения больного
- Изучение причин и факторов риска болезни
- Прогноз развития заболевания

ЭТАПЫ НАУЧНО-ПРАКТИЧЕСКОГО ИССЛЕДОВАНИЯ:

1. Формулирование цели и задач исследования.
2. Организация исследования.
3. Сбор информации.
4. Обработка информации.
5. Анализ результатов исследования.
6. Внедрение результатов исследования в практику и оценка эффективности.

I ЭТАП: ЦЕЛИ И ЗАДАЧИ ИССЛЕДОВАНИЯ

Этот этап включает в себя обоснование актуальности **проблемы и цели** исследования.

Цель – это конечный результат или желаемое состояние. Цель должна быть сформулирована четко и недвусмысленно.

I ЭТАП: ЦЕЛИ И ЗАДАЧИ ИССЛЕДОВАНИЯ

Название **темы** должно соответствовать цели исследования.

Для раскрытия поставленной цели необходимо определить **задачи исследования**, т.е. те конкретные действия, которые последовательно ведут к достижению цели исследования. Для небольших исследований намечают 4-6 задач.

I ЭТАП: ЦЕЛИ И ЗАДАЧИ ИССЛЕДОВАНИЯ

Большую помощь при формировании цели и задач исследования оказывает **рабочая гипотеза,** т.е. тот основной специфический вопрос исследования, на который необходимо ответить в ходе эксперимента, основная идея исследования, предвидение ожидаемых результатов.

I ЭТАП: ЦЕЛИ И ЗАДАЧИ ИССЛЕДОВАНИЯ

Анализ литературы помогает:

- Оценить степень разработки темы;
- Определить дизайн исследования и методы исследования;
- Оценить полученные ранее результаты;
- Изучить исторические аспекты проблемы, ее возникновение и подходы к решению.

II ЭТАП: ОРГАНИЗАЦИЯ ИССЛЕДОВАНИЯ

(DESIGN STUDY)

II ЭТАП: ОРГАНИЗАЦИЯ ИССЛЕДОВАНИЯ

Выбор объекта наблюдения:

- Под объектом наблюдения понимают статистическую совокупность, состоящую из отдельных предметов или явлений – единиц наблюдений, взятых в определенных границах времени и пространства.

- Формирование критериев включения и исключения.

II ЭТАП: ОРГАНИЗАЦИЯ ИССЛЕДОВАНИЯ

Единица наблюдения – первичный элемент статистической совокупности, являющийся носителем признаков (variables), подлежащих регистрации, изучению в ходе исследования.

Признаки или переменные (variables), могут принимать различные конкретные значения (values).

II ЭТАП: ОРГАНИЗАЦИЯ ИССЛЕДОВАНИЯ

Типы признаков *(виды шкал):*

- **Переменные**
- Категориальные
- (качественные)
  - Номинальные

  - Порядковые
  - (ординальные)

- Числовые
- (количественные)
  - Дискретные

  - Непрерывные

II ЭТАП: ОРГАНИЗАЦИЯ ИССЛЕДОВАНИЯ

Перечень признаков, подлежащих изучению в ходе исследования, оформляется в виде ***регистрационного документа*** (анкета, бланк, карта и т.п.), включающего вопросы, которые исследователь хочет изучить в ходе эксперимента и в дальнейшем заполняется на каждую единицу наблюдения.

II ЭТАП: ОРГАНИЗАЦИЯ ИССЛЕДОВАНИЯ

В зависимости *от степени охвата объекта* исследования принято различать:

- **сплошное исследование** (генеральная совокупность - population);
- **выборочное исследование** (выборочная совокупность - sample).

ГЕНЕРАЛЬНАЯ СОВОКУПНОСТЬ

Это совокупность всех мыслимо возможных объектов данного вида, над которыми проводятся наблюдения с целью получения конкретных значений определенной случайной величины.

РЕПРЕЗЕНТАТИВНОСТЬ

✔ **Репрезентативность** означает, что все пропорции генеральной совокупности должны быть представлены в выборке.

✔ **Репрезентативность** выборки обеспечивается **случайностью** отбора. Это означает, что любой объект выборки отобран случайно, при этом все объекты имеют одинаковую вероятность попасть в выборку.

II ЭТАП: ОРГАНИЗАЦИЯ ИССЛЕДОВАНИЯ

- *репрезентативность* – это представительность выборочной совокупности по отношению ко всей (генеральной) совокупности;

- *репрезентативность* должна быть ***количественной и качественной.***

II ЭТАП: ОРГАНИЗАЦИЯ ИССЛЕДОВАНИЯ

- Репрезентативность выборки зависит от …
- Главное требование, предъявляемое к отбору - …
- Случайность отбора достигается путем …

РАНДОМИЗАЦИЯ

Процесс создания репрезентативной выборки достигается путем **рандомизации** (random - случайный (англ.)), т.е. процессом случайного отбора элементов генеральной совокупности в выборку.

В процессе отбора следует избегать участия человека.

Следует использовать объективные (механические или электронные) средства рандомизации.

Существуют различные методы отбора объектов генеральной совокупности в выборку.

Чаще всего, элементы генеральной совокупности нумеруют, затем прибегают к одному из нижеперечисленных способов.

МЕТОДЫ СЛУЧАЙНОГО ОТБОРА ОБЪЕКТОВ

Механический отбор с повтором и без повтора. Отбор с помощью **таблиц** или **генератора** случайных чисел.

Многоступенчатая выборка.

Например, опрос студентов: сначала случайным образом выбираем вуз, затем случайно выбираем факультет, затем студента. В этом случае результат менее точный, чем при случайном выборе студентов сразу, без разделения по вузам и факультетам.

МЕТОДЫ СЛУЧАЙНОГО ОТБОРА ОБЪЕКТОВ

Кластерная выборка – похожа на многоступенчатую, отличие состоит в том, что исследуются все объекты последней ступени (в нашем случае, все студены данного факультета. Факультет и есть кластер).

Стратифицированная выборка – случайная выборка применяется отдельно для каждой группы (страты).

Систематическая выборка – например из списка объектов выбирается каждый 10-тый. Такая выборка наименее случайна.

II ЭТАП: ОРГАНИЗАЦИЯ ИССЛЕДОВАНИЯ

Важное место при решении организационных вопросов исследования принадлежит так называемому пробному, предварительному **(пилотному)** исследованию.

Пилотное исследование позволяет решить следующие основные задачи:

- отработать программу исследования;
- проверить варианты сбора данных;
- оценить вариабельность (разнообразие признаков);
- оценить затраты (время, деньги, штаты), необходимые для проведения исследования.

III ЭТАП: СБОР ИНФОРМАЦИИ

- На этом этапе основное внимание должно быть уделено соблюдению правил регистрации, охвату всех включенных в исследование единиц наблюдения, достоверности собранных данных.

- Выбор способа сбора данных определяется целью и задачами исследования и зависит от программы наблюдения, численности обследуемых единиц, уровня подготовки как организатора исследования, так и изучаемых лиц.

III ЭТАП: СБОР ИНФОРМАЦИИ

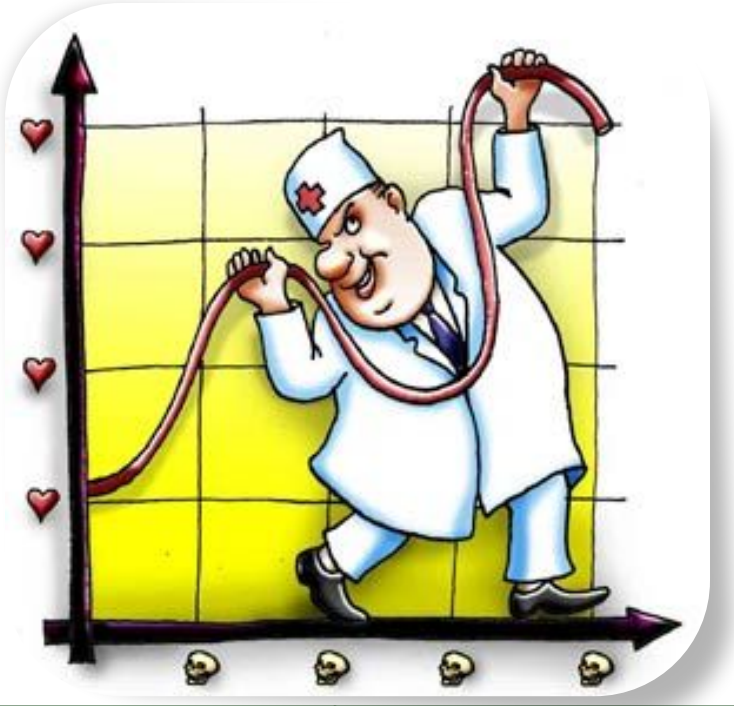
Способы сбора данных:

- ✔ отчетный (с помощью системы учетно-отчетной документации);
- ✔ экспедиционный (при обследовании деятельности отдельных учреждений, служб здравоохранения и т.п.);
- ✔ саморегистрация (предполагает самостоятельное заполнение обследуемым регистрационного документа);
- ✔ анкетный (сведения получают при помощи специальных вопросников, анкет рассылаемых или публикуемых в печати);
- ✔ корреспондентский (динамическое наблюдение за определенной группой лиц).

IV ЭТАП: ОБРАБОТКА ДАННЫХ

- СОЗДАНИЕ И ПОДГОТОВКА БАЗЫ ДАННЫХ

V ЭТАП: АНАЛИЗ РЕЗУЛЬТАТОВ ИССЛЕДОВАНИЯ



ТИПЫ ПРИЗНАКОВ (ВИДЫ ШКАЛ):

- **Переменные**
- Категориальные
- (качественные)
  - Номинальные

  - Порядковые
  - (ординальные)

- Числовые
- (количественные)
  - Дискретные

  - Непрерывные

Типы признаков (виды шкал):

Качественные категориальные (qualititative, categorical)

✔*Номинальные (Nominal);*

✔*Дихотомические (Binary - dichotomous);*

✔*Порядковые, ординальные, ранжируемые (Ordinal).*

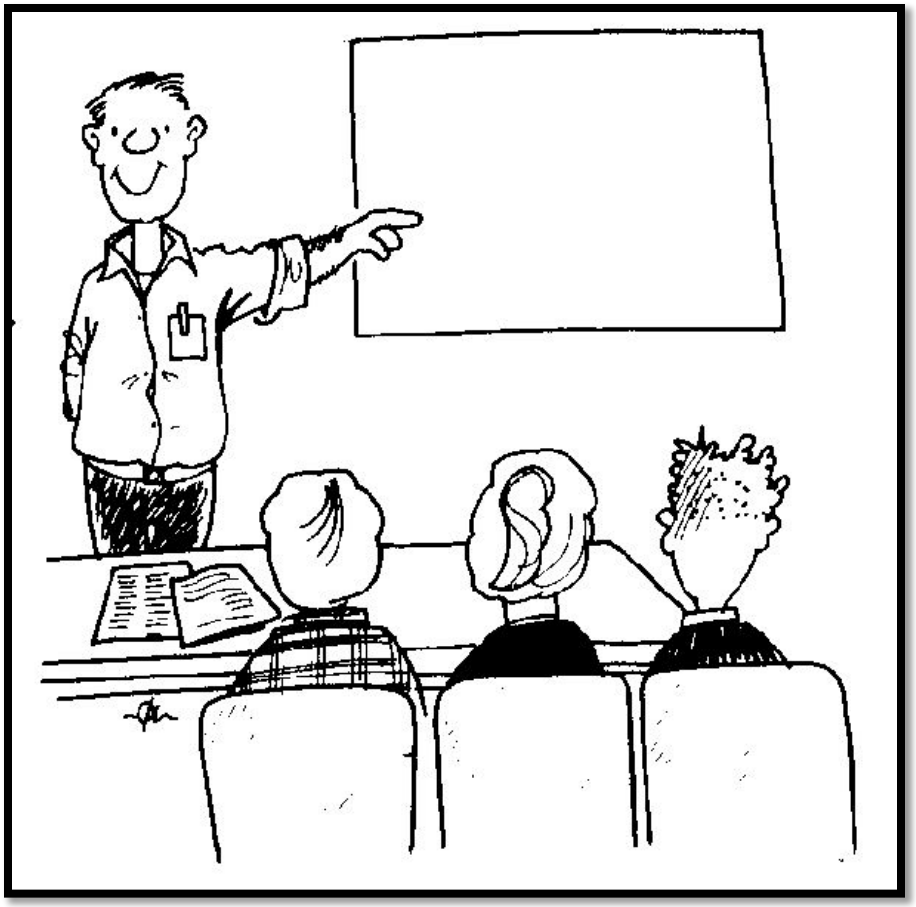
Количественные, интервальные (quantitative, numerical, interval)

✔*Дискретные (Discrete)*

✔*Непрерывные (Continuous)*

РАЗЛИЧИЕ МЕЖДУ ТИПАМИ ДАННЫХ

В зависимости от того, оказываются ли данные категориальными или числовыми, используют различные статистические методы.



ПРОИЗВОДНЫЕ (ВТОРИЧНЫЕ) ДАННЫЕ

Проценты. Могут возникать при рассмотрении вопроса относительно улучшения состояния больного во время лечения.

Пропорции или отношения. Иногда встречается два варианта пропорций или отношений. Например, индекс массы тела (индекс Кетле).

Интенсивность. Относительная частота заболеваний, где количество заболеваний делят на общее число лет, в течение которых вели наблюдения за пациентами в этом исследовании, общепринята при эпидемиологическом исследовании.

Метки и оценки. Произвольные значения, или метки, используют в том случае, когда невозможно изменить количество.

ЦЕНЗУРИРОВАННЫЕ ДАННЫЕ

Мы можем рассмотреть цензурированные данные на следующих примерах.

- Если мы проводим лабораторные измерения, используя прибор, который может обнаружить значения только выше некоторого предельного уровня, тогда любая величина ниже этого уровня не будет обнаружена. Например, вирус, уровень обнаружения которого ниже предела, часто рассматривается как «необнаруженный», при этом на самом деле он может находиться в образце.
- Мы можем столкнуться с цензурированными данными, например, когда некоторые больные из группы исследуемых отстраняются от испытания до окончания исследований.

ФОРМАТЫ ВВОДА ДАННЫХ

Существует несколько способов ввода данных и сохранения их в компьютере. Большинство статистических пакетов позволяют сразу же вводить данные. Однако существуют ограничения, а именно: вы не сможете переносить данные из одного пакета в другой. Простейшая альтернатива – сохранять данные либо в электронной таблице, либо в пакете баз данных. К сожалению, их статистические процедуры часто ограничены, и обычно возникает необходимость вводить данные в статистический пакет, чтобы провести исследования.

КАТЕГОРИАЛЬНЫЕ ДАННЫЕ

С нечисловыми данными могут возникнуть проблемы при занесении их в некоторые статистические пакеты, поэтому вам необходимо назначить числовые коды категориальным данным, прежде чем вводить данные в компьютер.

ЧИСЛОВЫЕ ДАННЫЕ

Должны быть введены с той же самой точностью, с которой были проведены измерения, и единица измерения должна быть едина для всех наблюдений данной переменной. Например, масса должна быть записана в килограммах или в граммах, но не попеременно то в кг, то в г.

МНОЖЕСТВЕННЫЕ ФОРМЫ НА ОДНОГО БОЛЬНОГО

Иногда информацию собирают на одного и того же больного более чем в одном случае (наблюдении). Важно отметить, что должен существовать уникальный идентификатор (например, порядковый номер), принадлежащий только одному человеку в данном наблюдении, который предоставит вам возможность соединить все данные, собранные на одного человека при исследовании.

КОДИРОВАНИЕ ОТСУТСТВУЮЩИХ (ПРОПУЩЕННЫХ) ДАННЫХ

Вам следует определить, что вы будете делать с отсутствующими данными, прежде чем вводить информацию. В большинстве случаев вы будете вынуждены использовать какой-нибудь символ для недостающих данных. Статистические пакеты предлагают для этого различные способы. Некоторые пакеты используют специальные символы.

ПРОВЕРКА ОШИБОК И ВЫБРОСОВ

При любом исследовании всегда есть опасность допустить ошибки при наборе данных либо вначале, при измерениях, либо при сборе, переписывании и вводе данных в компьютер. Довольно трудно избежать этих ошибок. Однако можно сократить количество опечаток и описок путем тщательной проверки данных, как только они будут введены. Даже бегло просмотрев таблицу, можно обнаружить очевидные ошибки.

ВЫБРОСЫ (АНОМАЛЬНЫЕ ЗНАЧЕНИЯ)

Наблюдения, которые отличаются от главной группы данных и несовместимы с остальными. Эти данные могут быть подлинными наблюдениями с очень экстремальными величинами переменной. Однако они могут появиться также в результате опечаток и в этом случае любые данные, вызывающие подозрение, должны быть проверены. Важно выяснить, имеются ли выбросы в наборе данных, так как они могут в значительной степени повлиять на результаты некоторых исследований.

ГРАФИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ ДАННЫХ

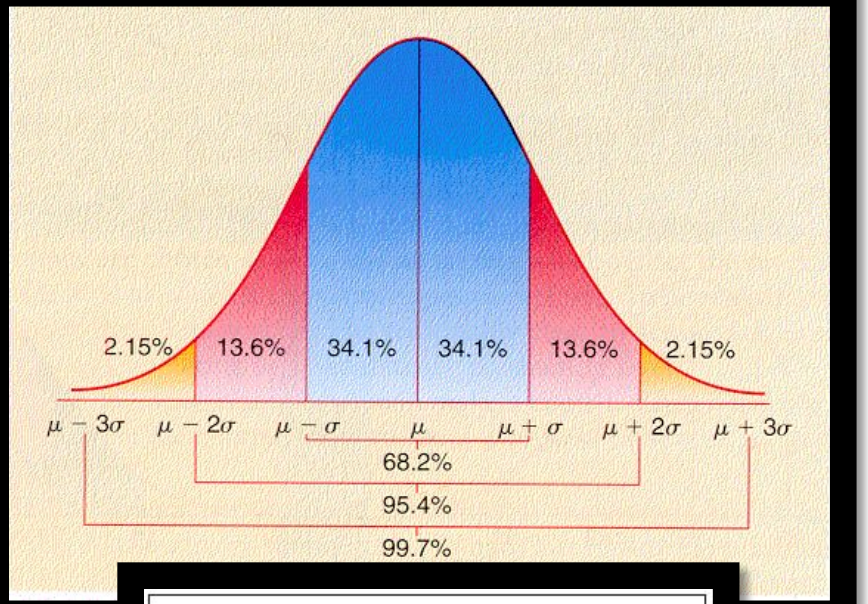
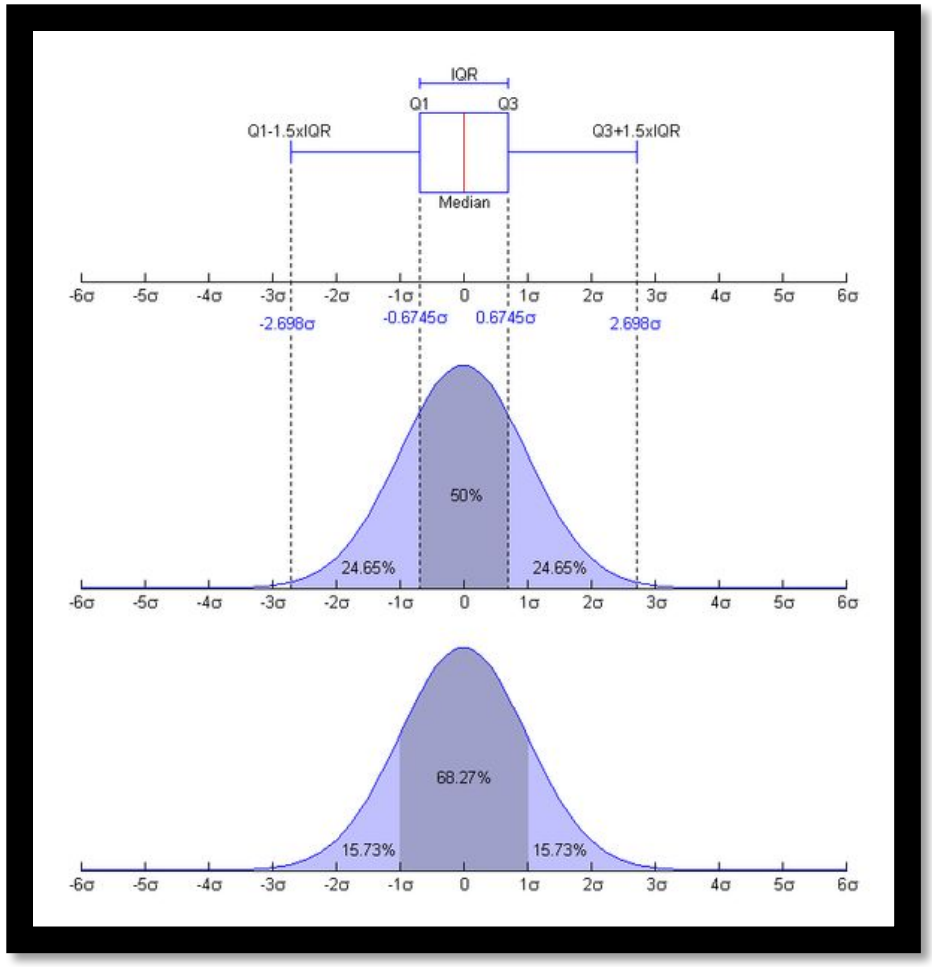
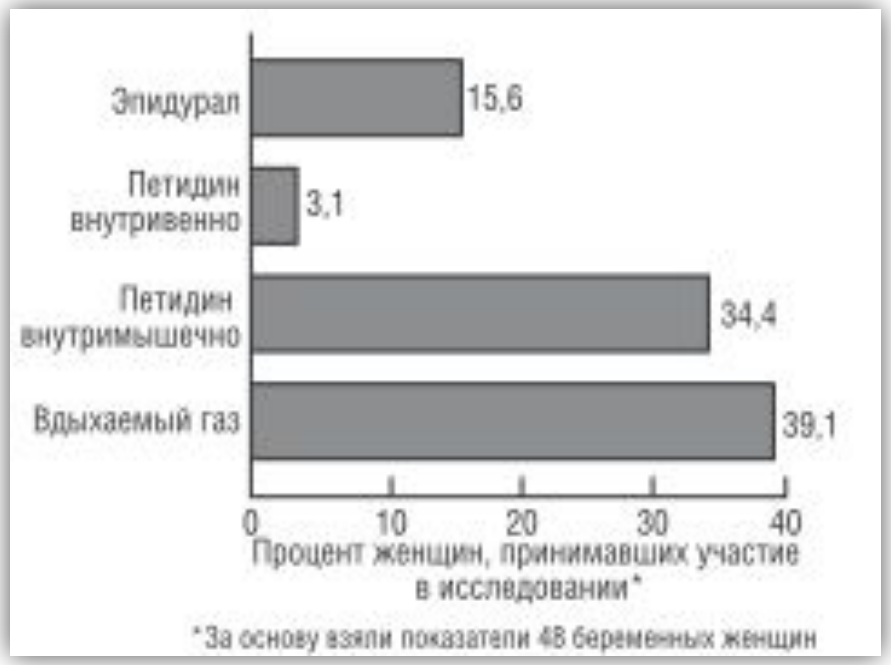
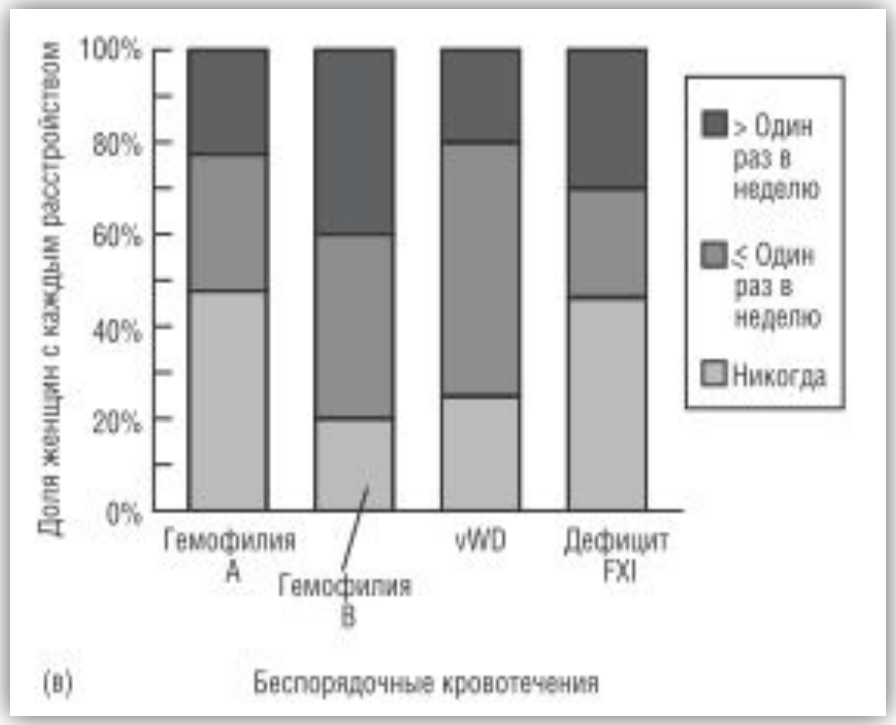


Рис. 3. Диаграмма адреноселективности различных –блокаторов

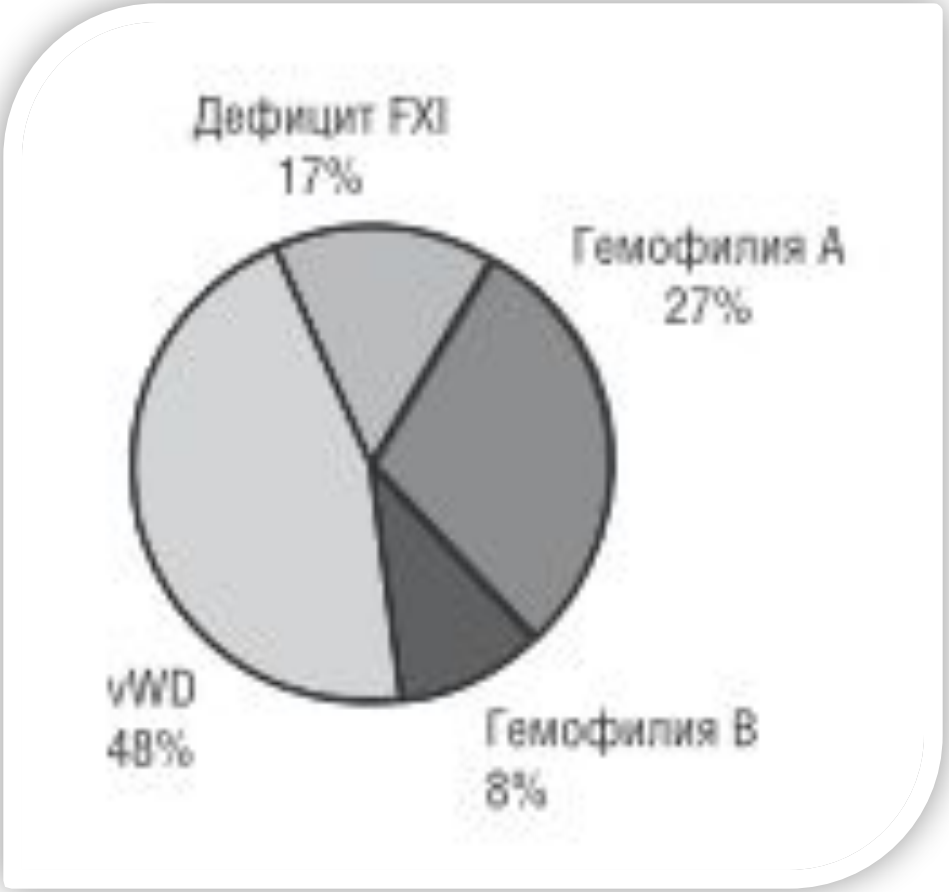


СТОЛБЧАТАЯ И КОЛОНЧАТАЯ ДИАГРАММА

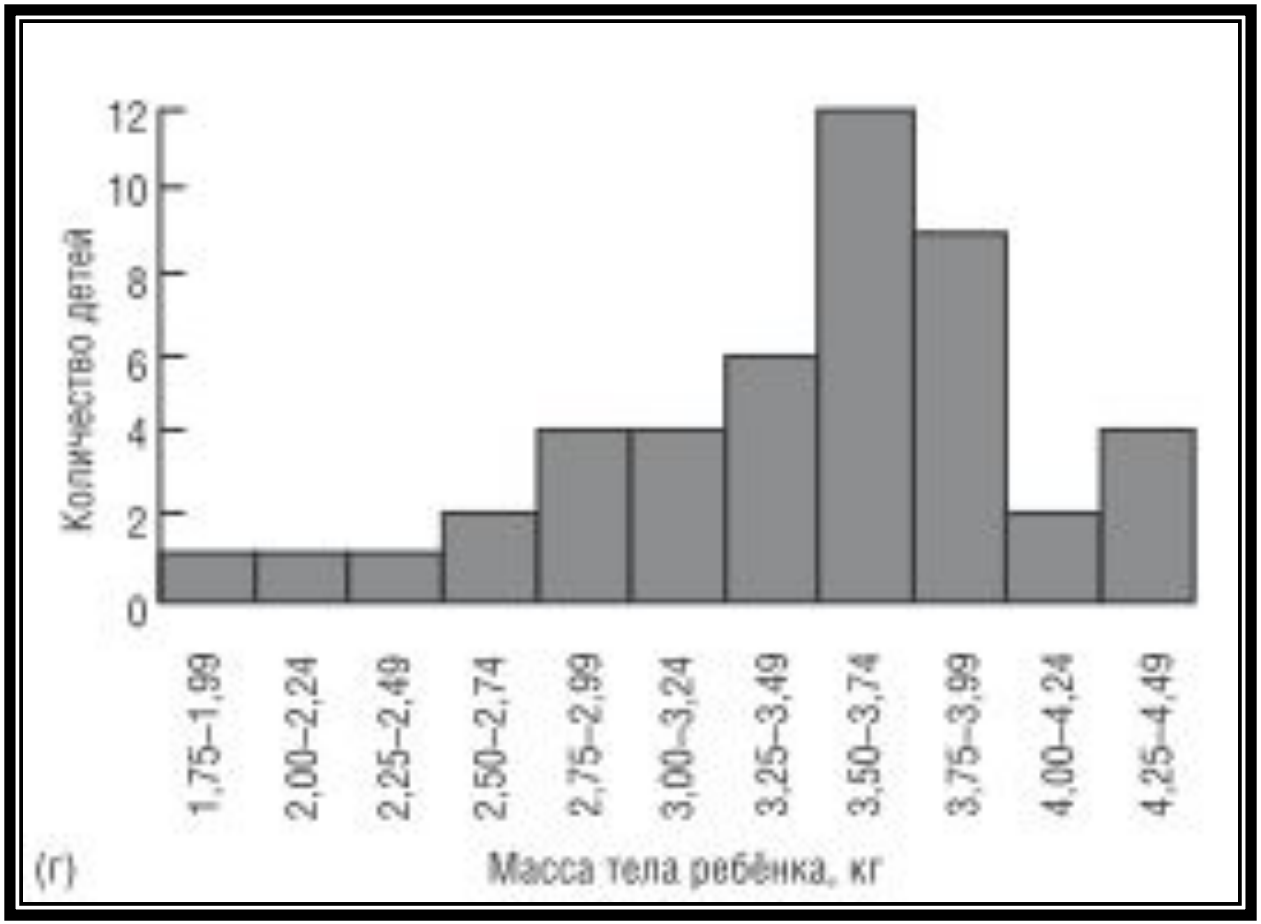




КРУГОВАЯ ДИАГРАММА



ГИСТОГРАММА



ТОЧЕЧНЫЙ ГРАФИК

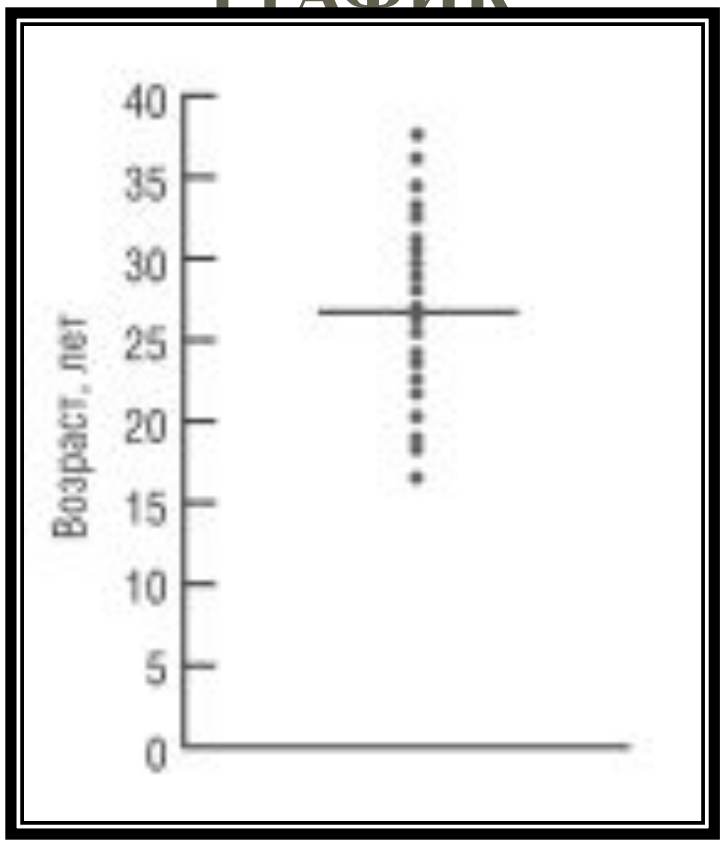


ГРАФИК «СТЕБЕЛЬ И ЛИСТЬЯ»

3	1,0	04
665	1,1	39
53	1,2	99
9751	1,3	1135677999
955410	1,4	0148
987655	1,5	00338899
9531100	1,6	0001355
731	1,7	00114569
99843110	1,8	6
654400	1,9	01
6	2,0	
7	2,1	19
10	2,2	
Беклометазон		Плацебо

ГРАФИК *BOX-PLOT*

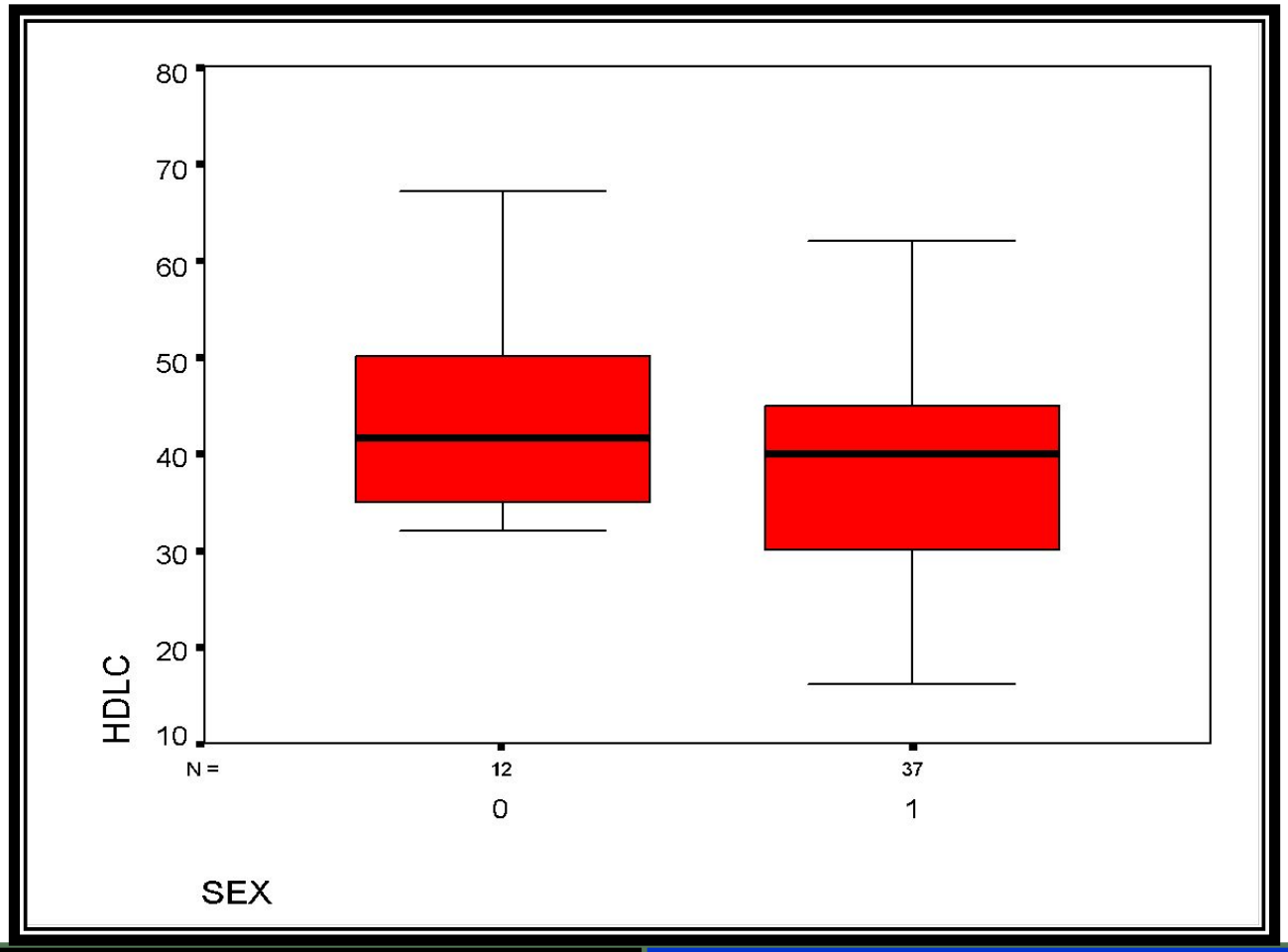


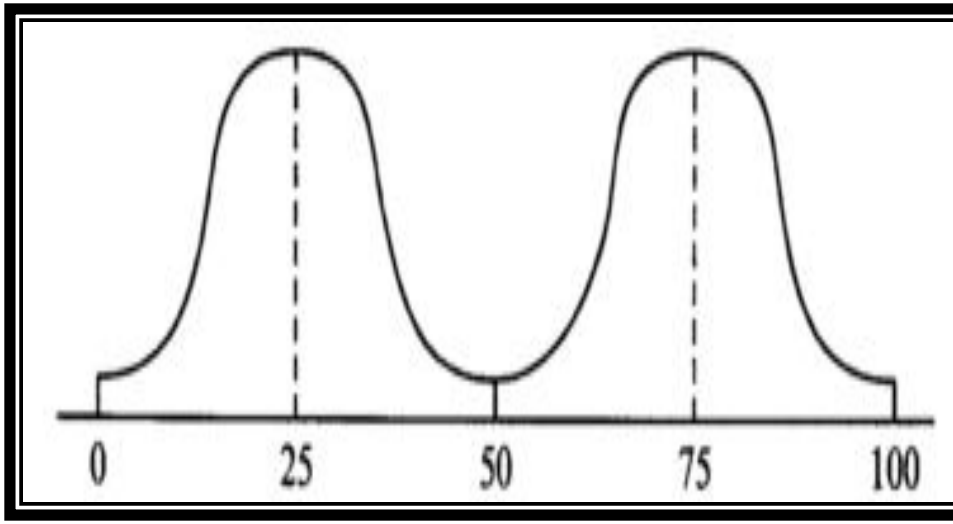
ГРАФИК *BOX-PLOT*

ФОРМЫ ЧАСТОТНОГО РАСПРЕДЕЛЕНИЯ

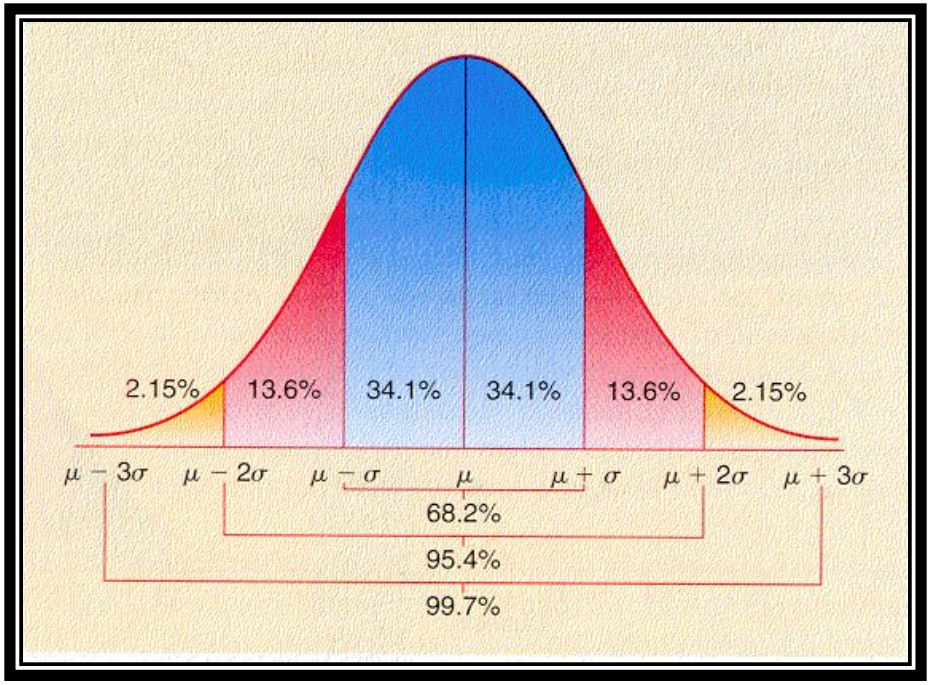
Выбор наиболее подходящего статистического метода часто зависит от формы распределения. Распределение данных чаще всего *унимодальное,* т.е. имеющее одну «вершину».

Иногда распределение *бимодальное* (две «вершины») или равномерное (каждая величина одинаково вероятна и нет «вершин»).

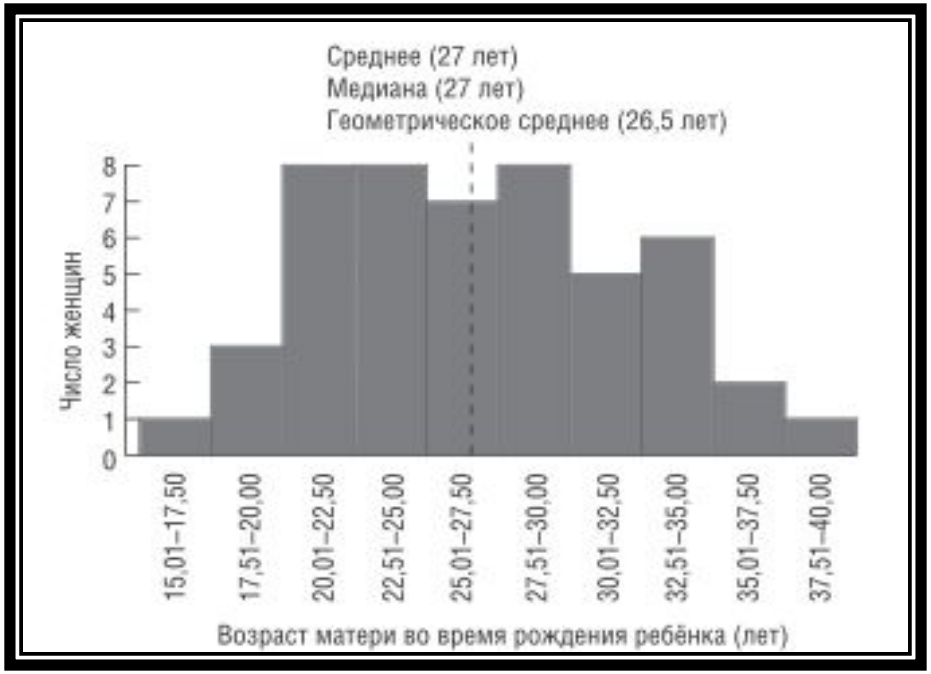
БИМОДАЛЬНОЕ



УНИМОДАЛЬНОЕ



ФОРМЫ ЧАСТОТНОГО РАСПРЕДЕЛЕНИЯ



ПОКАЗАТЕЛИ НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ

✔ **ПОКАЗАТЕЛИ ЦЕНТРАЛЬНОЙ ТЕНДЕНЦИИ**

Среднее (average, mean)

Мода (mode)

Медиана (median)

✔ **ПОКАЗАТЕЛИ РАЗБРОСА ДАННЫХ**

Дисперсия (variance)

Стандартное отклонение (standard deviation)

Интерквартильное расстояние

ОПИСАНИЕ ДАННЫХ: «МЕРЫ ПОЛОЖЕНИЯ»

СРЕДНЕЕ АРИФМЕТИЧЕСКОЕ

Одна из ***мер центральной тенденции.*** Вычисляется путем суммирования всех величин в группе и последующего деления полученной суммы на число слагаемых.

ОПИСАНИЕ ДАННЫХ: «МЕРЫ ПОЛОЖЕНИЯ»

СРЕДНЕЕ АРИФМЕТИЧЕСКОЕ (M - м)

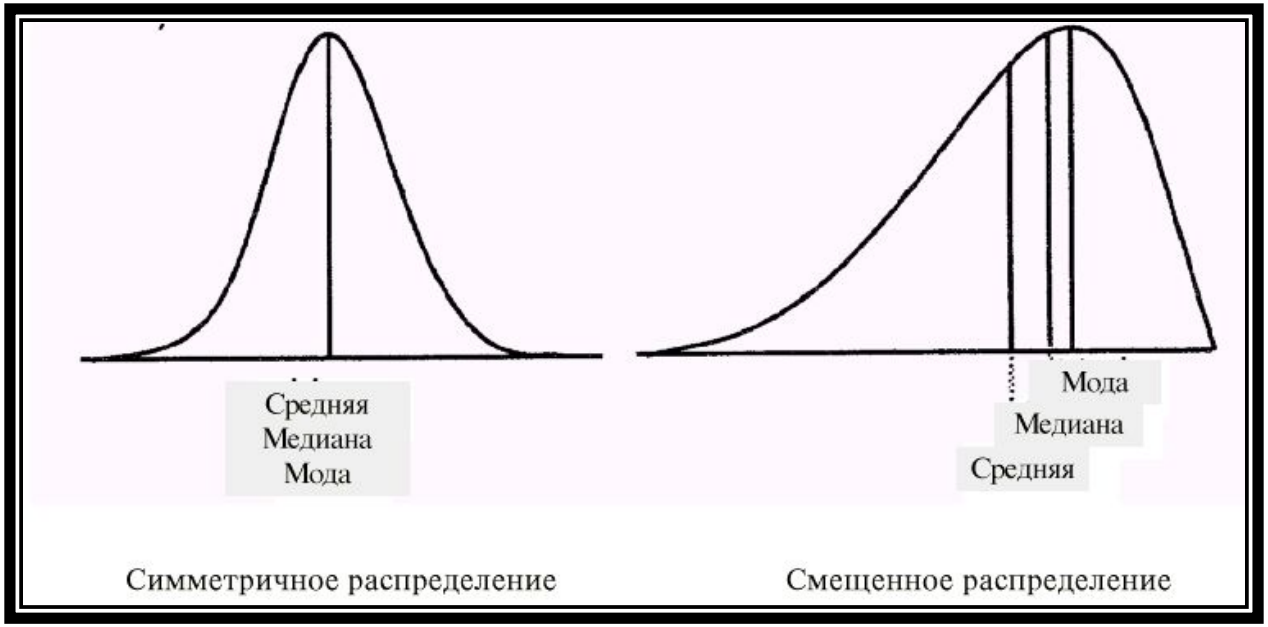
$$\overline{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}.$$

Используя математическую систему обозначения, мы можем сократить это выражение:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n},$$

МЕДИАНА (MEDIAN - Me)

Вид меры центральной тенденции. Простейшее деление набора измерений на две части: нижнюю и верхнюю половины. Точка на шкале, которая делит группу таким образом, называется медианой.



МОДА (MODE - Mo)

Вид меры центральной тенденции. Наиболее часто встречающееся значение среди набора наблюдений.

Мода — значение, которое встречается наиболее часто в наборе данных; если данные непрерывные, то мы обычно группируем их и вычисляем модальную группу. Некоторые наборы данных не имеют моды, потому что каждое значение встречается только один раз.

СРЕДНЕЕ ГЕОМЕТРИЧЕСКОЕ
(GEOMETRIC MEAN)

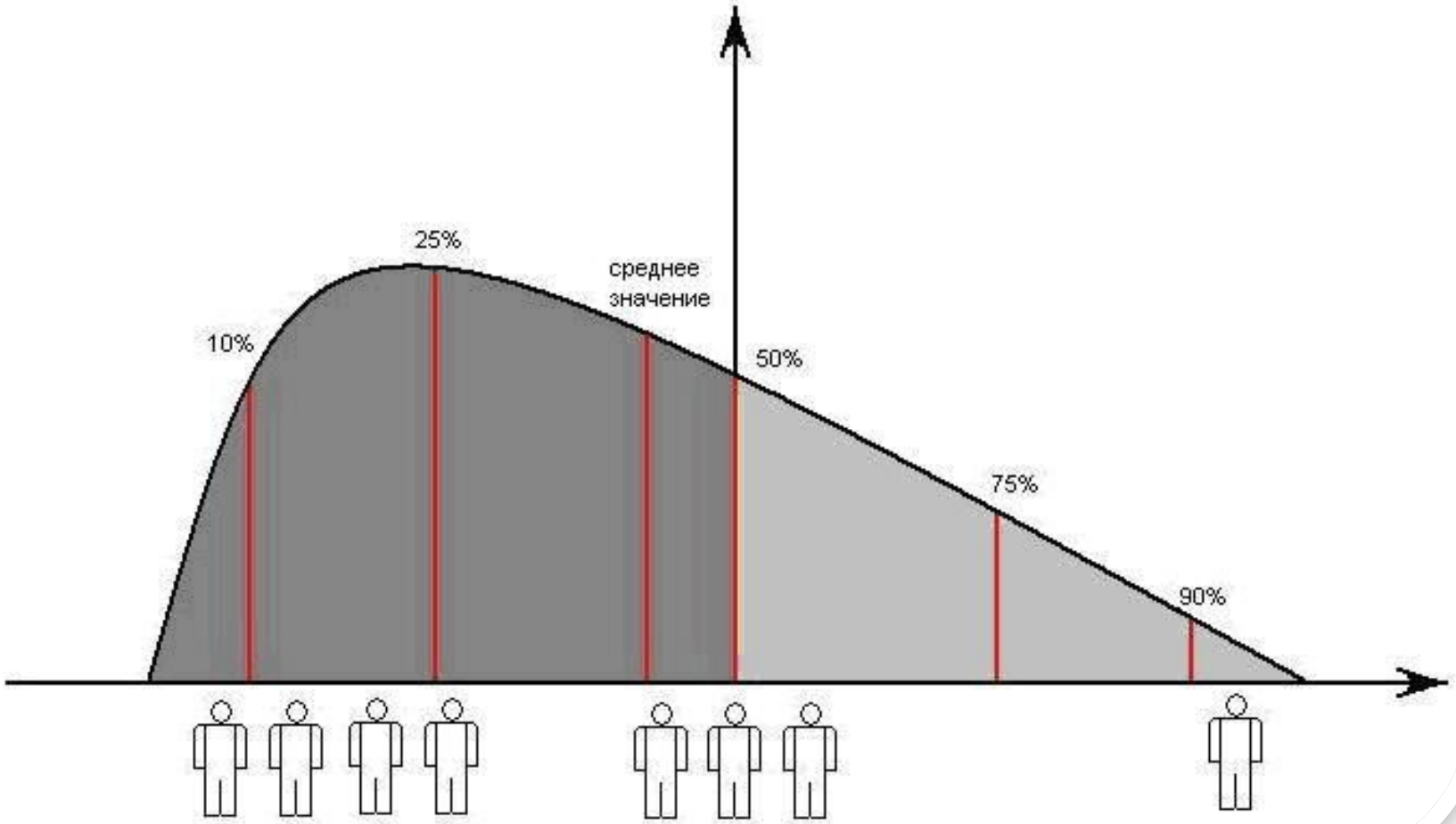
Одна из мер центральной тенденции. Вычисляется суммированием логарифмов всех величин в группе, вычислением средней арифметической, затем от полученного значения берут антилогарифм. Может быть найдена только в случае, если все величины в группе положительны.

ОПИСАНИЕ ДАННЫХ: «МЕРЫ РАССЕЯНИЯ»

РАЗМАХ (ИНТЕРВАЛ ИЗМЕНЕНИЯ)

Разность между максимальным и минимальным значениями переменной в наборе данных; вы найдете эти две величины, на которые ссылаются вместо их разности.

Размах, полученный из процентилей. Что такое процентили?



ПРИМЕНЕНИЕ ПРОЦЕНТИЛЕЙ

Межквартильный размах – разница между первым и третьим квартилем, т.е. между 25-м и 75-м процентилями. В него входят центральные 50% наблюдений в упорядоченном наборе, где 25% наблюдений находятся ниже центральной точки и 25% - выше.

Интердецильный размах содержит в себе центральные 80% наблюдений, т.е. те наблюдения, которые располагаются между 10-м и 90-м процентилями.

Часто используют размах, который содержит 95% наблюдений, т.е. он исключает 2,5% наблюдений снизу и 2,5% сверху. Можно применить этот интервал, осуществляя диагностику болезни. В этом случае он называется ***референтный интервал, референтный размах или нормальный размах.***

ДИСПЕРСИЯ

(ОТ ЛАТ. – *DISPERSES* – РАССЕЯННЫЙ, РАССЫПАННЫЙ)

Один из способов *измерения рассеяния* данных заключается в том, чтобы *определить степень отклонения* каждого наблюдения от *средней арифметической*. Очевидно, что чем больше отклонение, тем больше изменчивость, вариабельность наблюдений. Однако мы не можем использовать среднее этих отклонений как меру рассеяния, потому что положительные отклонения компенсируют отрицательные отклонения (их сумма тождественно равна нулю). Для того чтобы решить эту проблему, мы возводим в квадрат каждое отклонение и находим среднее возведенных в квадрат отклонений; эта величина называется вариацией, или дисперсией.

СТАНДАРТНОЕ ОТКЛОНЕНИЕ

Стандартное (среднее квадратичное) отклонение – положительный квадратный корень из дисперсии. На примере n наблюдений это выглядит так. Мы можем размышлять о стандартном отклонении как о своего рода среднем отклонении наблюдений от среднего. Его вычисляют в тех же самых единицах (размерностях), что и исходные данные.

Если разделить стандартное отклонение на среднее арифметическое и выразить этот показатель в процентах, получится коэффициент вариации. Это мера рассеяния которая не зависит от единиц измерения (безразмерная), но имеет некоторые теоретические неудобства, поэтому статистики её не всегда одобряют.

ПОНИМАНИЕ ВЕРОЯТНОСТИ

МОЖНО ВЫЧИСЛИТЬ ВЕРОЯТНОСТЬ, ИСПОЛЬЗУЯ РАЗЛИЧНЫЕ ПОДХОДЫ:

- **СУБЪЕКТИВНАЯ;**
- **ЧАСТОТНАЯ;**
- **АПРИОРНАЯ.**

РАСПРЕДЕЛЕНИЕ ВЕРОЯТНОСТИ: ТЕОРИЯ

Случайная величина – это величина, которая может принимать любое из набора взаимоисключающих значений с определенной вероятностью.

Распределение вероятности показывает вероятности всех возможных значений случайной переменной. Это теоретическое распределение, которое выражено математически и имеет среднее и дисперсию, являющиеся аналогами среднего и дисперсии в эмпирическом распределении.

НОРМАЛЬНОЕ (ГАУССОВСКОЕ РАСПРЕДЕЛЕНИЕ)

Одно из самых важных распределений в статистике – нормальное распределение. Его функция плотности распределения вероятности:

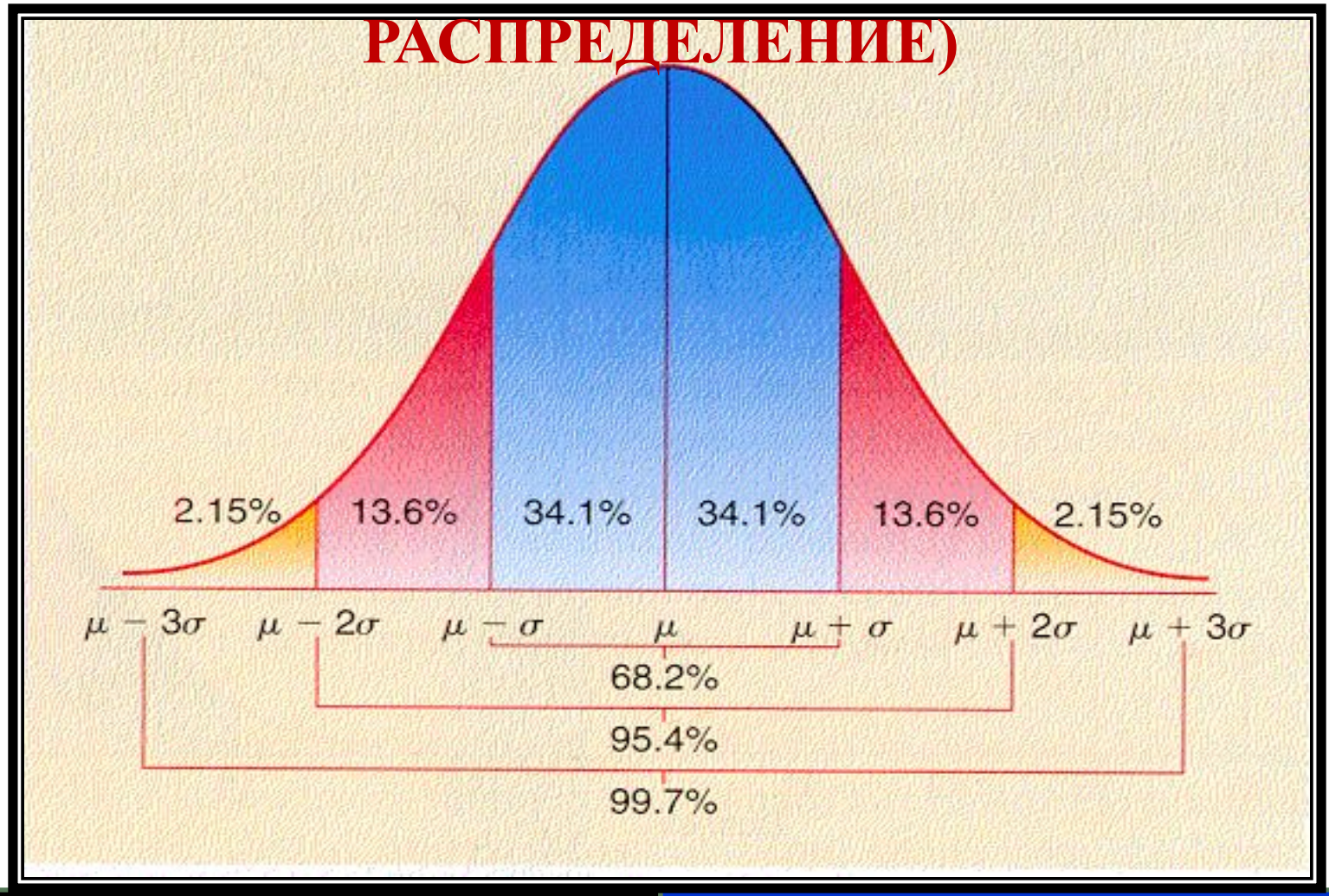
- полностью определяется двумя параметрами, среднее (μ) и дисперсия (σ^2);
- колоколообразна (унимодальна);
- симметрична относительно среднего;
- сдвигается вправо, если среднее увеличивается, и влево, если среднее уменьшается (при постоянной дисперсии);
- сплющивается, если дисперсия увеличивается, но становится более остроконечной, если дисперсия уменьшается (для постоянного среднего).

НОРМАЛЬНОЕ (ГАУССОВСКОЕ РАСПРЕДЕЛЕНИЕ)

ДОПОЛНИТЕЛЬНЫЕ СВОЙСТВА

- Среднее и медиана нормального распределения равны.
- Вероятность того, что нормально распределенная случайная переменная X, со средним μ и стандартным отклонением σ, находящаяся между:
- $(\mu-\sigma)$ и $(\mu+\sigma)$, равна 0,68;
- $(\mu-1{,}96\sigma)$ и $(\mu+1{,}96\sigma)$, равна 0,95;
- $(\mu-2{,}58\sigma)$ и $(\mu+2{,}58\sigma)$, равна 0,99.

НОРМАЛЬНОЕ (ГАУССОВСКОЕ РАСПРЕДЕЛЕНИЕ)

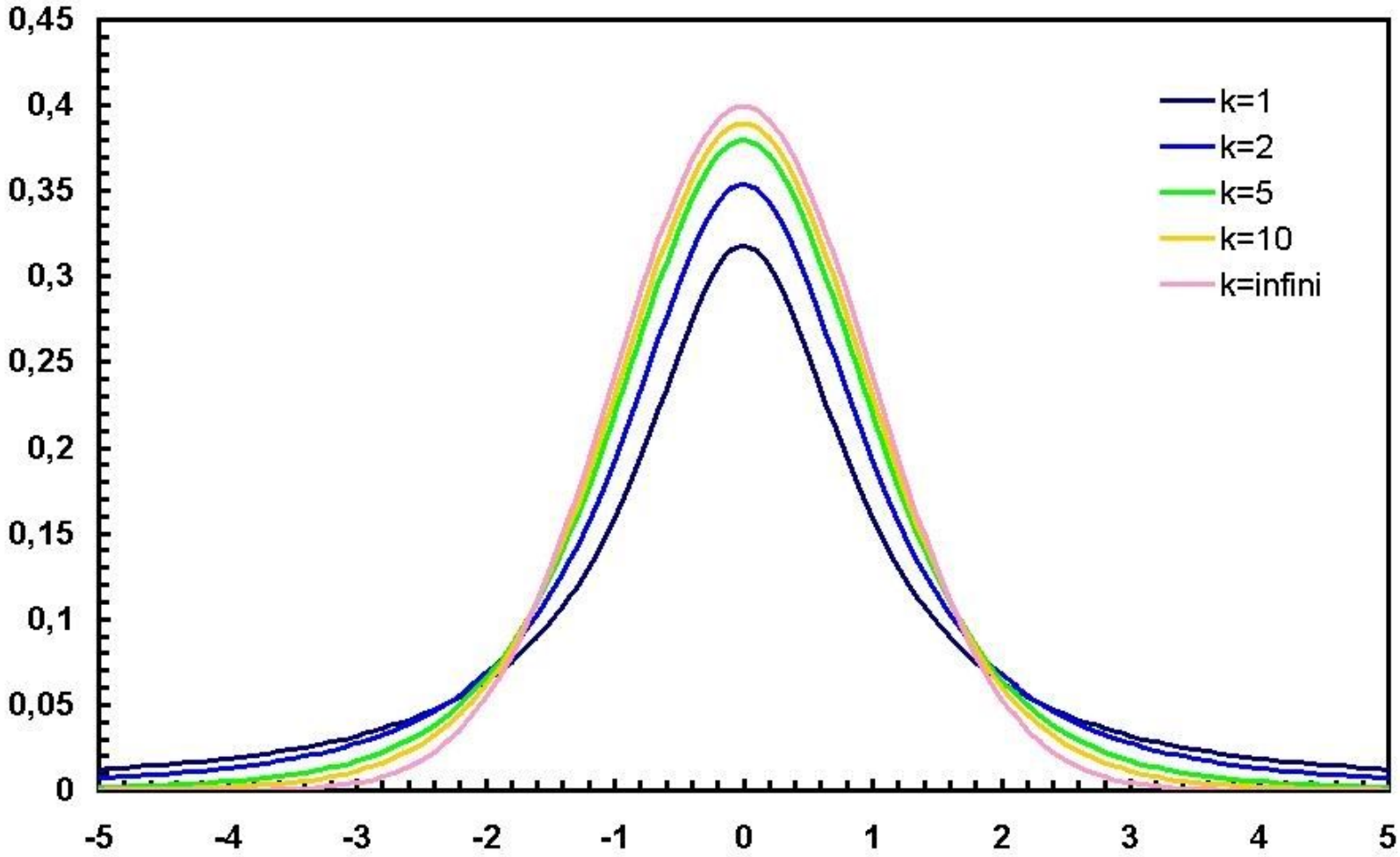


ТЕОРЕТИЧЕСКИЕ РАСПРЕДЕЛЕНИЯ:
ДРУГИЕ РАСПРЕДЕЛЕНИЯ

t-распределение

- Получено Вильямом Госсетом, который публиковался под псевдонимом Студент (Student), поэтому его часто называют t-распределением Стьюдента.
- Параметры, которые характеризуют t-распределение, - это степени свободы (df), так как мы сможем начертить функцию плотности распределения вероятности только в том случае, если мы будем знать уравнение t-распределения и степени свободы. Степени свободы часто выражаются через объем выборки.
- Форма подобна форме для стандартизованного нормального распределения, но более приплюснута и с более длинными хвостами. Форма приближается к нормальной кривой, по мере того как увеличиваются степени свободы.
- В частности, его применяют для вычисления доверительных интервалов и исследования гипотез с одной или двумя средними.

t-распределение



ТЕОРЕТИЧЕСКИЕ РАСПРЕДЕЛЕНИЯ:

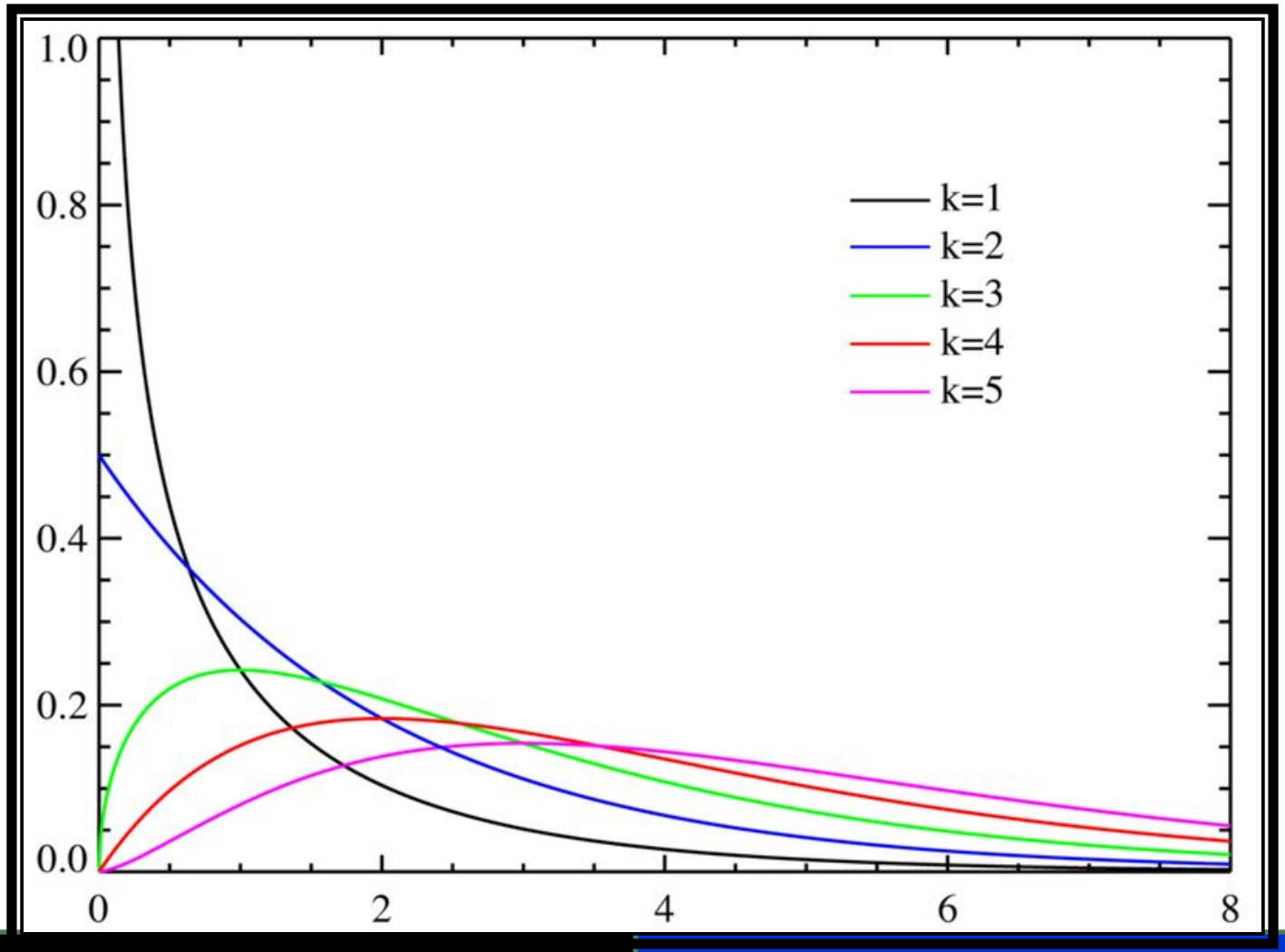
ДРУГИЕ РАСПРЕДЕЛЕНИЯ
НЕПРЕРЫВНОЕ РАСПРЕДЕЛЕНИЕ ВЕРОЯТНОСТЕЙ

Хи-квадрат

Хи-квадрат, (χ^2) или распределение Пирсона:

- скошено вправо и принимает только положительные значения;
- характеризуется степенями свободы;
- его форма зависит от числа степеней свободы – становится более симметричной и приближается к нормальной с их ростом;
- особенно часто используется для анализа категориальных данных.

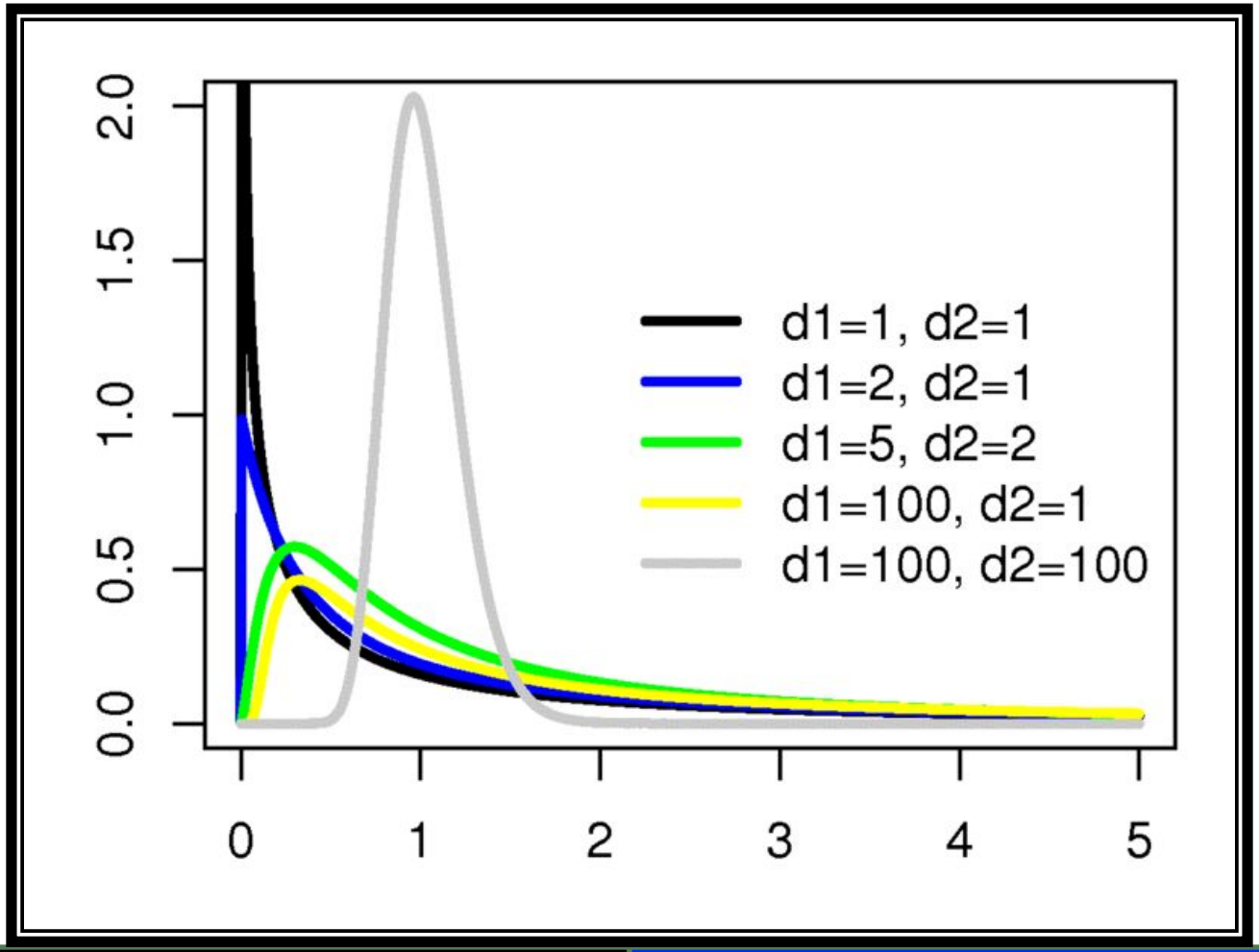
Хи-квадрат



F-распределение

- Скошено вправо.
- Определяется как отношение. Распределения отношения двух оценок дисперсий, вычисленных для нормально распределенных данных, аппроксимируется F-распределением.
- Два параметра, которые характеризуют его, - степени свободы числителя и знаменателя отношения.
- F-распределение особенной полезно для сравнения двух дисперсий и более чем двух средних при использовании дисперсионного анализа (ANOVA).

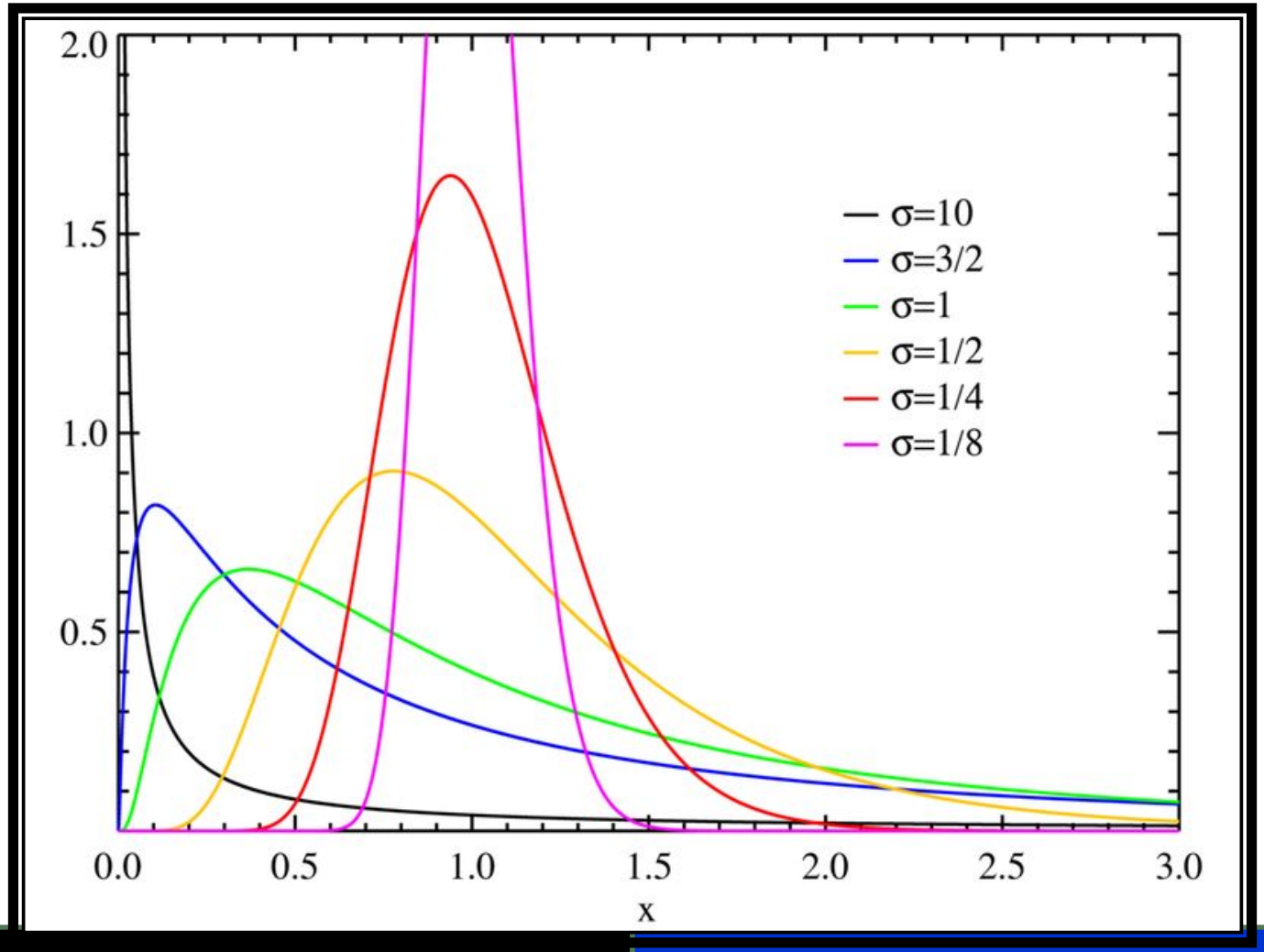
F-распределение



Логнормальное распределение

- Распределение вероятности случайной переменной, логарифм которого (по основанию 10 или более ***e*** – основание натурального логарифма) имеет нормальное распределение.
- Сильно скошено вправо.
- Если взять логарифмы исходных данных, которые скошены вправо, мы создадим эмпирическое распределение, которое почти нормальное и тогда данные соответствуют приближенно логнормальному распределению.
- Многие переменные в медицине имеют логнормальное распределение. Можно использовать свойства нормального распределения для того, чтобы сделать выводы относительно этих переменных после логарифмического преобразования данных.
- Если набор данных имеет логнормальное распределение, то используют среднее геометрическое как обобщающий показатель положения.

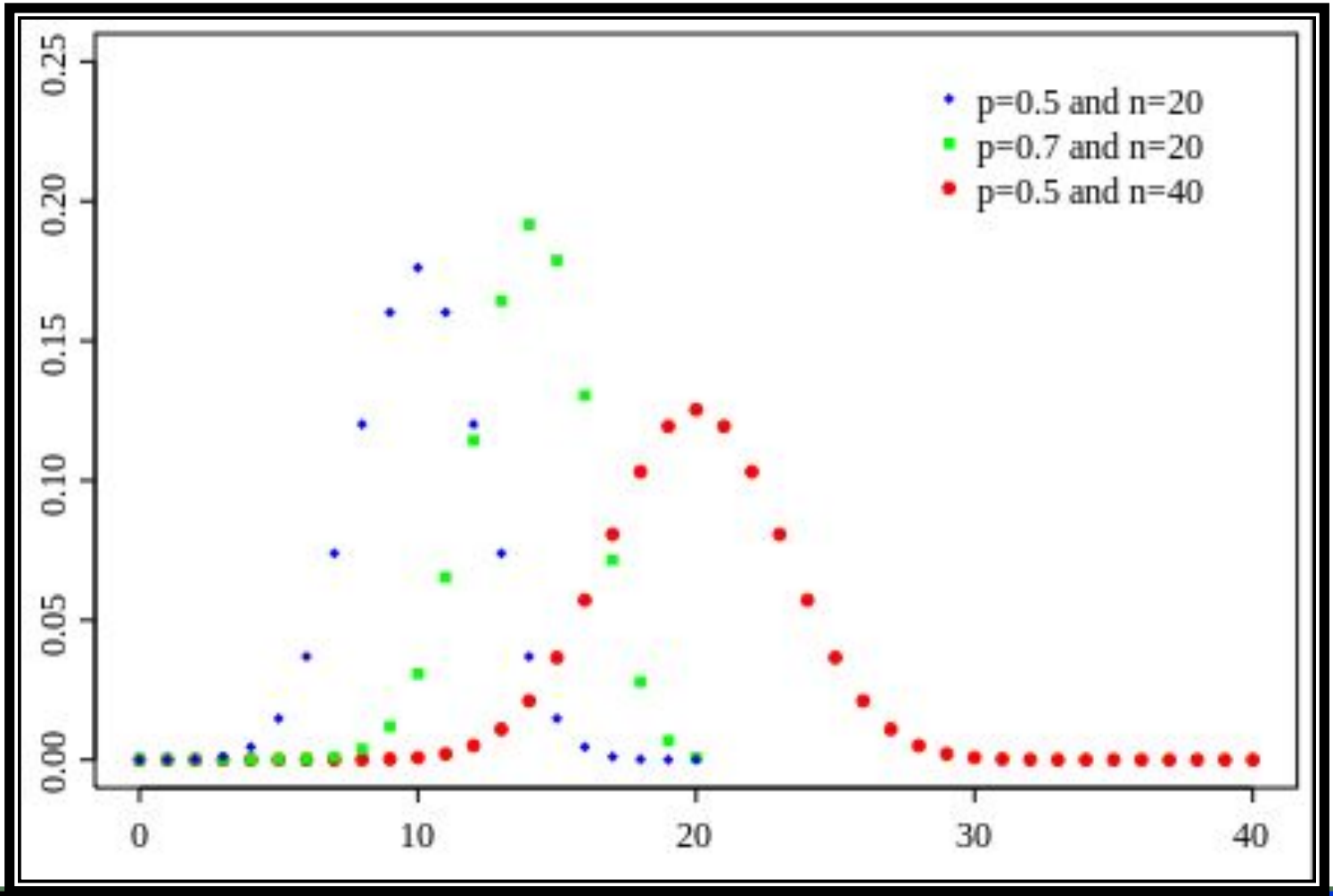
Логнормальное распределение



ТЕОРЕТИЧЕСКИЕ РАСПРЕДЕЛЕНИЯ:

ДРУГИЕ РАСПРЕДЕЛЕНИЯ
ДИСКРЕТНЫЕ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ

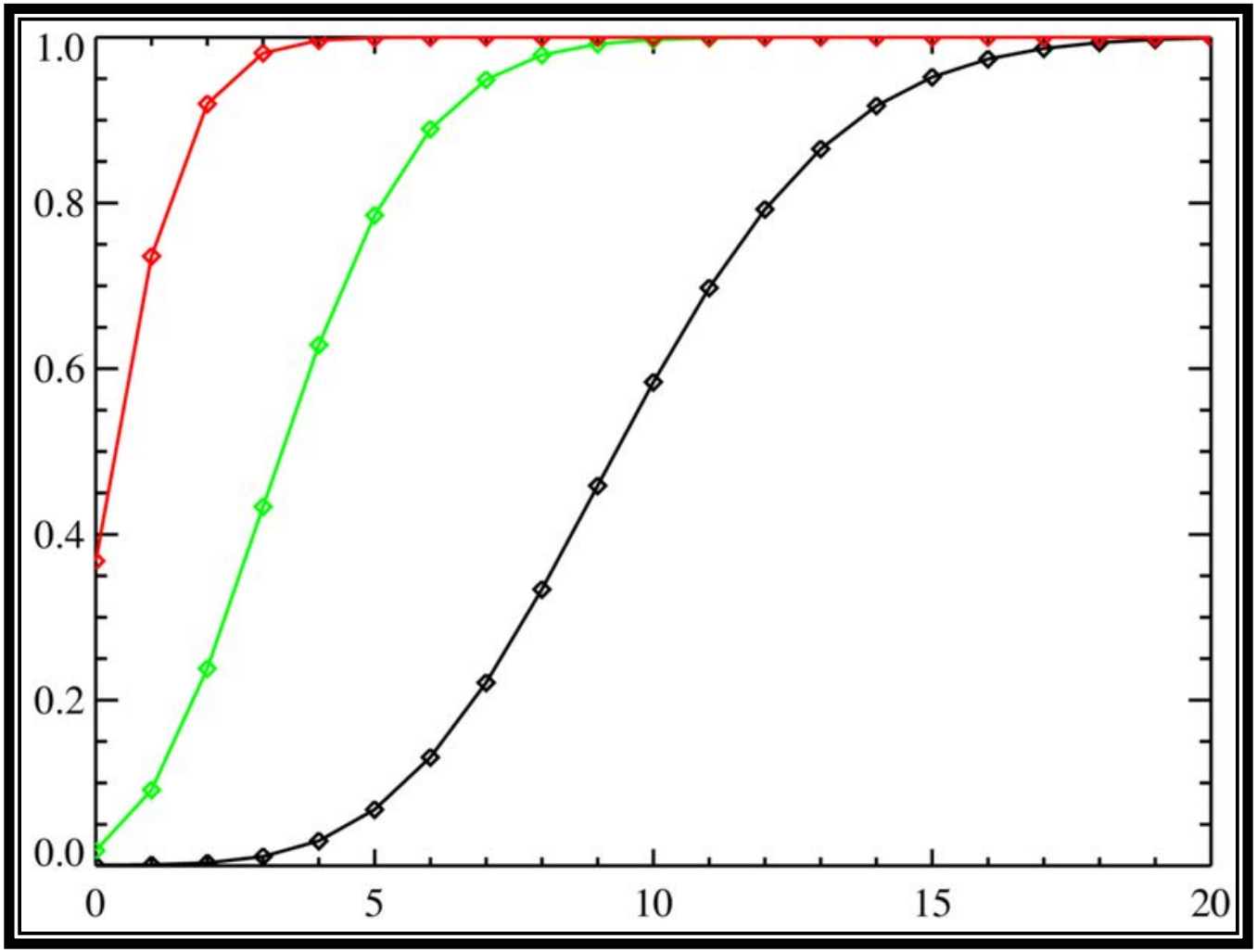
Биноминальное распределение



Распределения Пуассона

- Пуассоновская случайная переменная – число событий, которые происходят независимо и случайно во времени или пространстве с постоянной средней интенсивностью μ. Например, количество госпитализаций в день типично отвечает распределению Пуассона. Знание распределения Пуассона используют для того, чтобы вычислить вероятность конкретного количества госпитализаций в любой отдельный день.
- Параметр, которым описывают распределение Пуассона, - среднее, т.е. средняя интенсивность μ.
- Среднее равняется дисперсии в распределении Пуассона.
- Если среднее ближе к минимальному, то распределение будет скошено вправо и становится более симметричным по мере того, как среднее будет увеличиваться, оно приближается, по форме, к нормальному распределению.

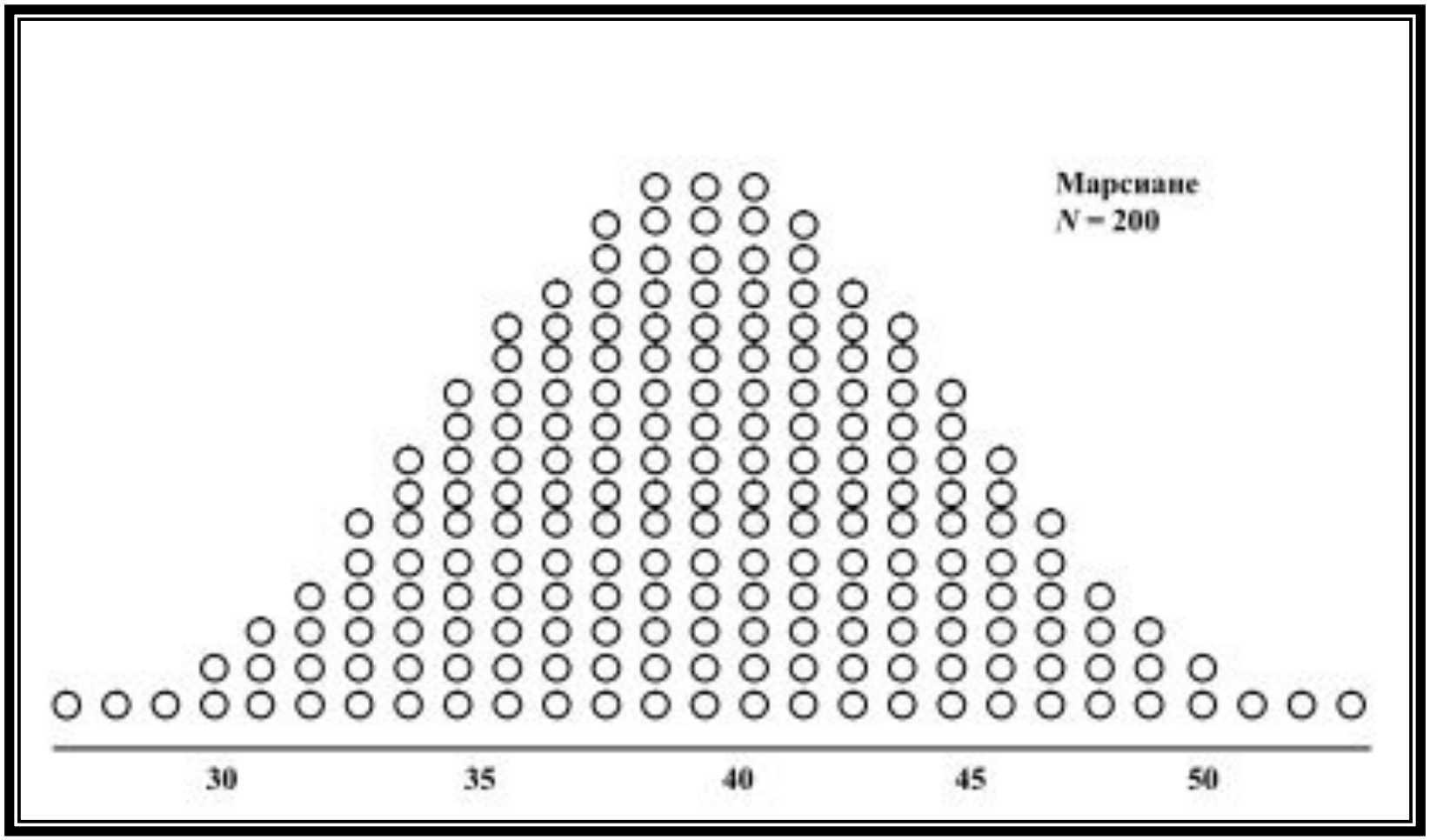
Распределения Пуассона



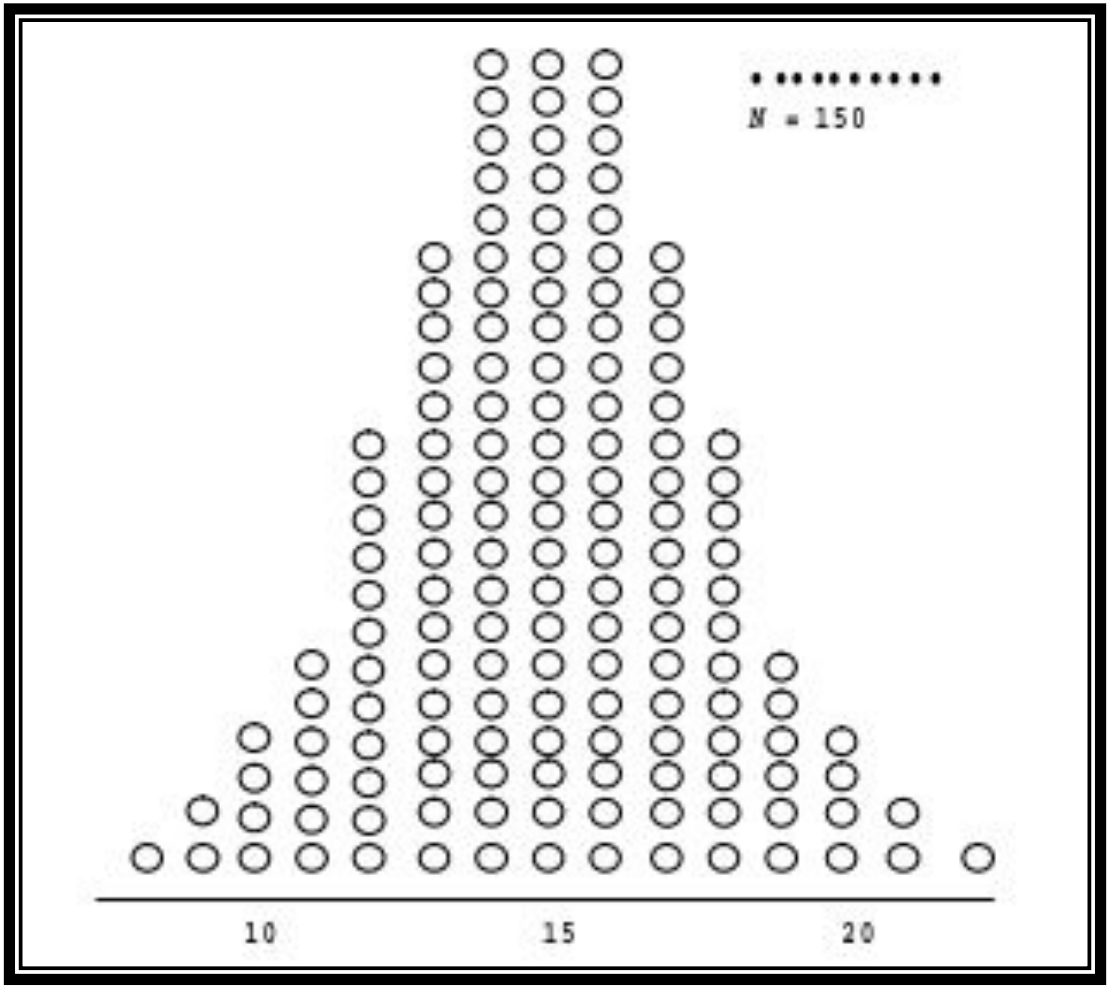
КАК ОПИСАТЬ ДАННЫЕ?

Если значения интересующего нас признака у большинства объектов близки к их среднему и с равной вероятностью отклоняются от него в большую или меньшую сторону, лучшими характеристиками совокупности будут само ***среднее значение и стандартное отклонение***. Напротив, когда значения признака распределены несимметрично относительно среднего, совокупность лучше описать с помощью ***медианы и процентилей.***

РАСПРЕДЕЛЕНИЕ МАРСИАН ПО РОСТУ

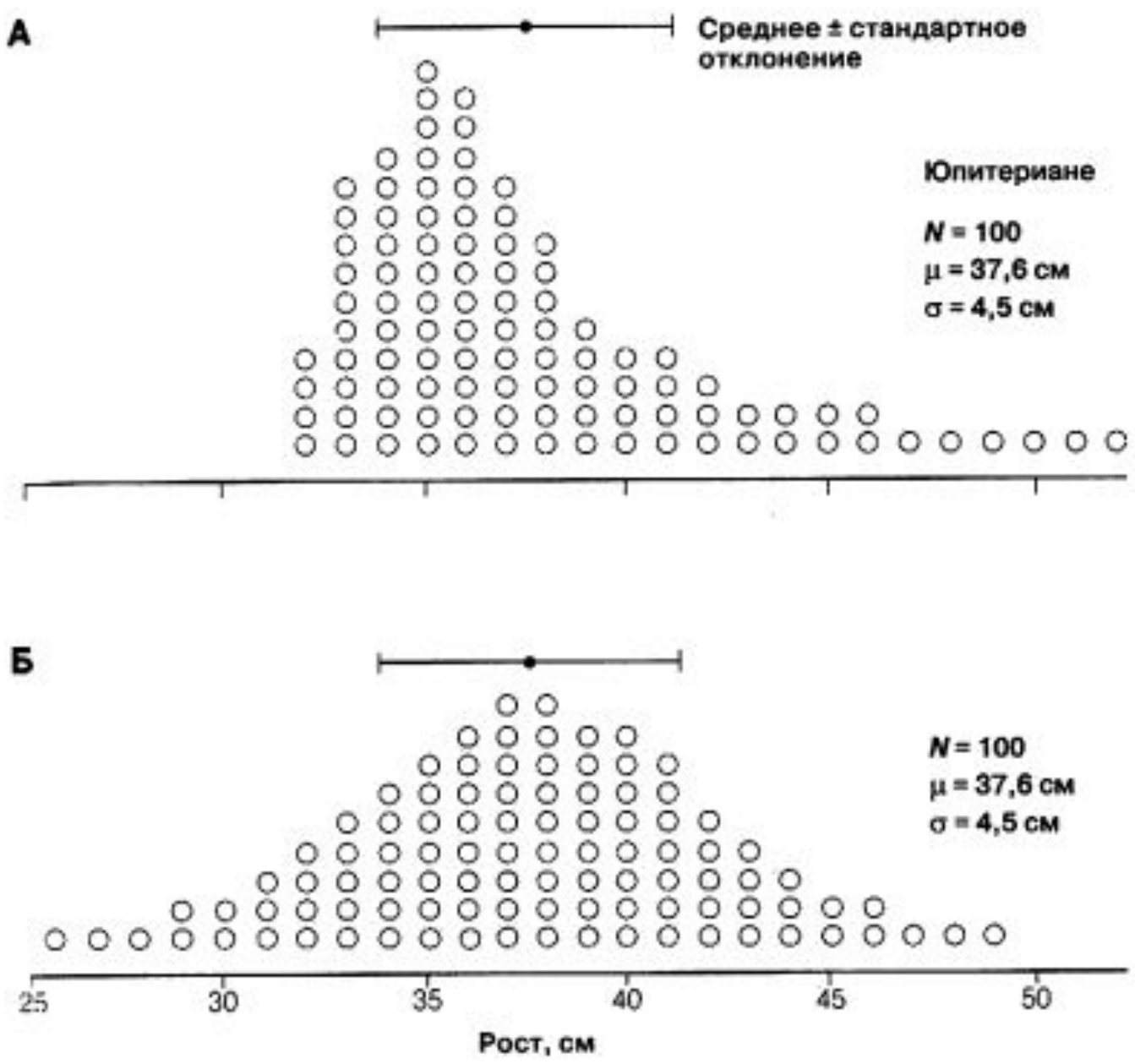


РАСПРЕДЕЛЕНИЕ ВЕНЕРИАЦЕВ ПО РОСТУ



ПАРАМЕТРЫ РАСПРЕДЕЛЕНИЯ МАРСИАН И ВЕНЕРИАЦЕВ

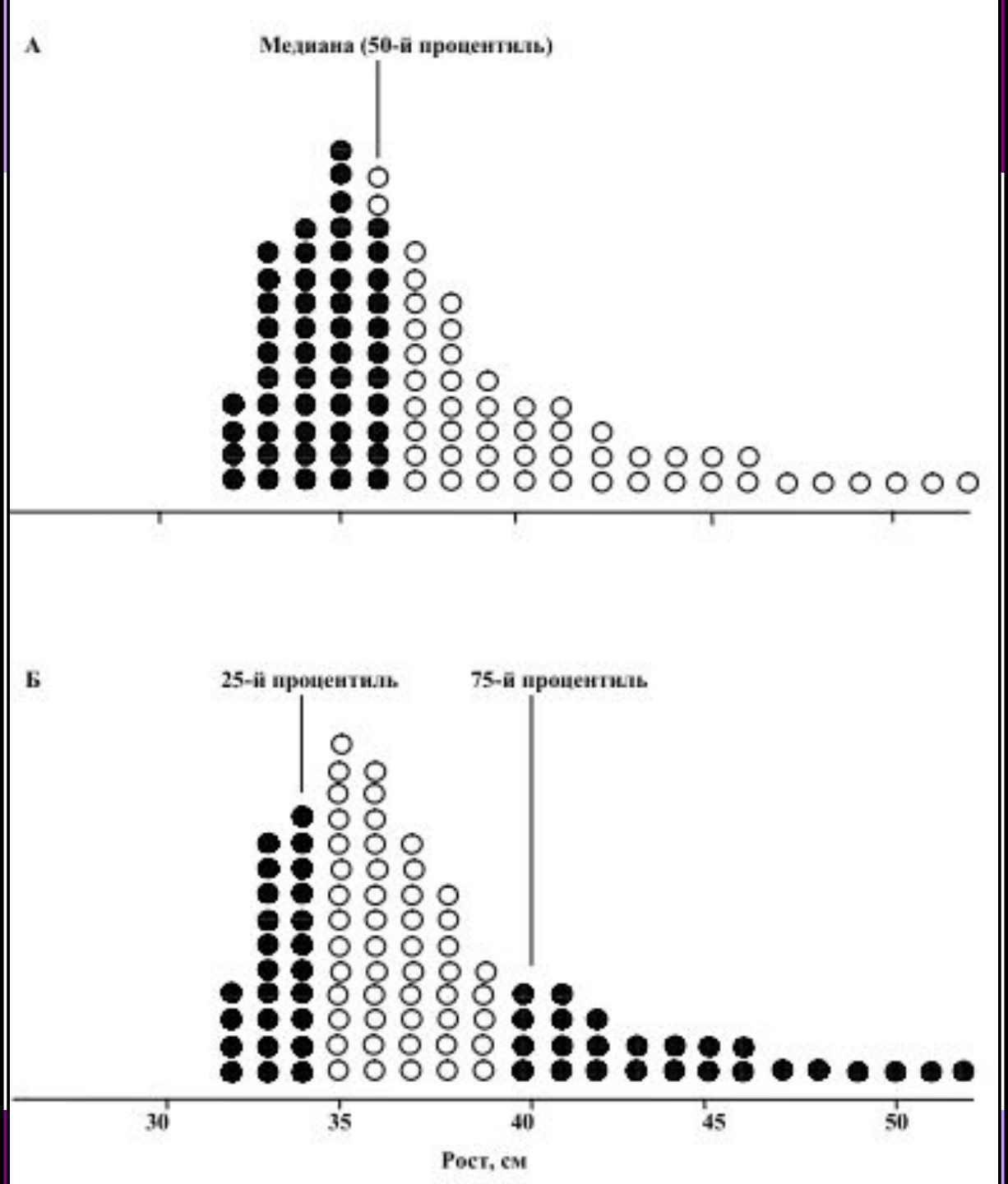
	Объем совокупности	Среднее, см	Стандартное отклонение, см
Марсиане	200	40	5
Венерианцы	150	15	2,5


А
Среднее ± стандартное отклонение
Юпитериане
N = 100
μ = 37,6 см
σ = 4,5 см
Б
N = 100
μ = 37,6 см
σ = 4,5 см
25
30
35
40
45
50
Рост, см

Если распределение асимметрично полагаться на среднее и стандартное отклонение нельзя.

А. Распределение юпитериан по росту.

Б. Нормальное распределение с теми же средним и стандартным отклонением, не смотря на тождественность параметров, оно ничуть не похоже на реальное распределение юпитериан.


А
Медиана (50-й процентиль)
Б
25-й процентиль
75-й процентиль
30
35
40
45
50
Рост, см

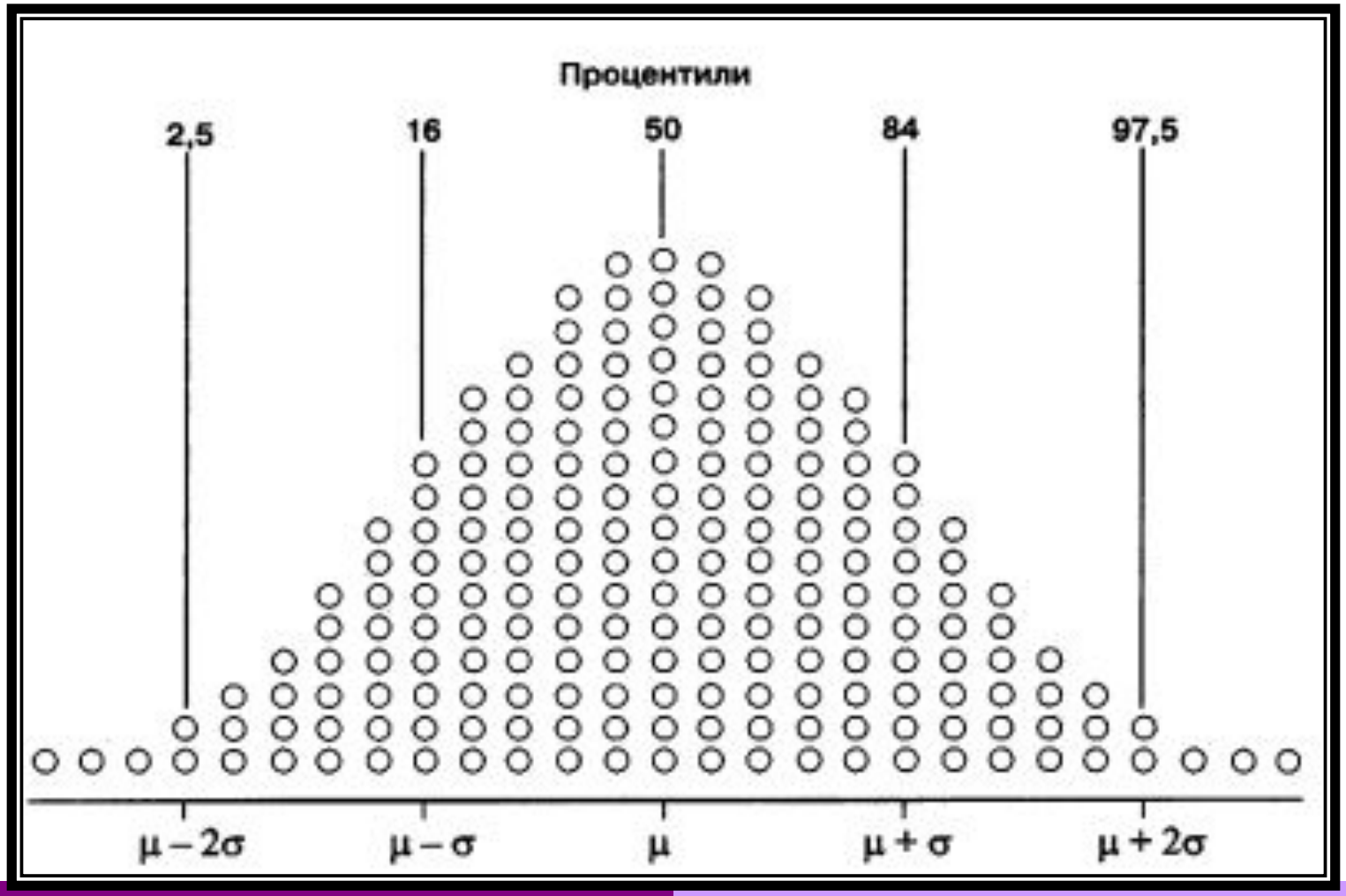
Для описания асимметричного распределения следует использовать медиану и процентили.

Медиана — это значение, которое делит распределение пополам.

А. Медиана роста юпитериан — 36 см.

Б. 25-й и 75-й процентили отсекают четверть самых низких и четверть самых высоких юпитериан 25-й процентиль ближе к медиане, чем 75-й — это говорит об асимметричности распределения.

НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ, СООТВЕТСТВИЕ МЕЖДУ ЧИСЛОМ СТАНДАРТНЫХ ОТКЛОНЕНИЙ ОТ СРЕДНЕГО И ПРОЦЕНТИЛЯМИ



А в чем проблема?

- Вариабельность
  - Случайная
  - Систематическая

(a) Large bias, small variability

(b) Small bias, large variability

(c) Large bias, large variability

(d) Small bias, small variability

Статистика

- Описательная
  - Графические методы
  - Суммирование данных
- Статистические выводы
- Статистические модели
  - Проверка гипотез
  - Поиск закономерностей (data mining)

Статистические выводы

- ✔ цель статистики: аппроксимация истины
- ✔ некоторые определения
- ✔ различия между *статистической* и *клинической* значимостью

Позиция #1. Статистика как *отражение* истины

- Статистическая значимость не истина, а "аппроксимация" истины
- Истина
  - Что мы можем сделать для людей, что бы они жили дольше или лучше
  - Исследования позволяют нам приблизиться к истине
- Наша цель: выяснить, насколько точно статистика отражает истину

Позиция #2. *Пользователи* статистики не должны быть профессиональными статистиками

- ✔ Вам не надо знать много о статистике, чтобы эффективно ее использовать
- ✔ Не концентрируйтесь на том, правильна ли *статистика*
  - ✔ Попытайтесь понять, что статистика пытается вам сказать

Священная P-оценка

P< .05

Алтарь статистики

P оценка

- "***P***robability"
- Вероятность того, что различия между двумя группами возникли **случайно**
- Искусственно фиксирована на уровне 5% (P = 0.05)

P оценка

- Зависит от нескольких факторов.
  - Насколько был **большим** эффект.
  - Насколько **одинаковым** был эффект у обследованных.
  - Как **много** пациентов было обследовано.
- Если все эти факторы растут, вероятность нахождения значимых различий увеличивается.
- После того, как мы решили, что различия не вследствие случайности, нам нужно решить значимы ли они клинически.

Извлечение информации из p-оценки

- **"Высоко значимая"** — $P < 0.001$
  - Если количество пациентов небольшое, p-оценка свидетельствует о том, что эффект был либо очень большим, либо униформным (либо и то и другое)
  - Если количество пациентов велико эффект может быть и не очень большим

Извлечение информации из p-оценки

- “**Не значимо**” P > 0.05 (например, 0.15)
  - Если количество пациентов мало, их может быть просто недостаточно для обнаружения реально существующих различий
  - Если количество пациентов достаточно велико, мы можем быть уверены в том, что нет различий между терапевтическими режимами или эффект лечения не стабильный

Извлечение информации из p-оценки

- **"Пограничная значимость"** — P = 0.08 — ????
  - *Могли* бы достичь значимости, если бы в исследовании было больше пациентов
  - Размер эффекта небольшой или нестабильный
  - Нельзя сделать никаких выводов кроме того, что нужны дополнительные исследования

Статистика в медицинских исследованиях

- ✔ Логика научного метода
  - ✔ Дедуктивная логика (выдвигается гипотеза, затем собираются факты) - от общего к частному
  - ✔ Индуктивная логика (от фактов к формулировке гипотезы
- ✔ Фальсификация (C.Popper)

Нулевая гипотеза

- ✔ Предполагаем, что различий *нет*
- ✔ Собираем данные и оцениваем *существующие* различия
- ✔ Если нулевая гипотеза справедлива, то какова вероятность получения подобных результатов в результате *случайного процесса*?
- ✔ Если вероятность достаточно мала, нулевая гипотеза отвергается

Альтернативная гипотеза

- ✔ Между группами существуют различия (но мы не можем сказать, какой величины)

Ошибки при статистическом выводе

- ✔ Альфа ошибка (вероятность отвергнуть нулевую гипотезу, если на самом деле она справедлива) - ошибка потребителя, ошибка первого типа
- ✔ Бета ошибка (вероятность отвергнуть альтернативную гипотезу, если на самом деле она верна) - ошибка спонсора, ошибка второго типа

Доверительные интервалы

- "Статистика статистики"
- статистические показатели - это оценки
  - Доверительные интервалы показывают нам границы нашей оценки

Доверительный интервал

- ✔ Интервал, в котором с заданной вероятностью (обычно 95%) находится популяционное среднее значение

ГЗТ у женщин с пролиферативными заболеваниями молочных желез в анамнезе

TABLE 4
Relative Risk of Invasive Breast Carcinoma Associated with Duration of Estrogen Replacement Therapy in Menopausal Women with a History of Premenopausal Benign Breast Disease

Estrogen replacement therapy	No. of patients	No. of woman-years	No. of breast carcinomas	Relative risk[a] (95% confidence interval)
Unknown	402	3952	18	1.44 (0.87-2.4)
Yes, duration	3383	39,509	107	0.91 (0.68-1.2)
1-12 mos	707	9221	26	1.00 (0.65-1.6)
1-5 yrs	888	14,028	29	0.78 (0.51-1.2)
>5 yrs	1779	16,063	52	0.98 (0.69-1.4)
Unknown	9	197	0	0.0
No	2028	28,154	88	1.0[b]
Total	5813	71,615	213	