

An abstract, colorful pattern is visible along the left edge of the slide, featuring various shapes and colors like red, yellow, blue, and purple.

Законы распределения и их применение для расчетов и анализа



Структура лекции

1. Первичная обработка информации
2. Законы распределения и их применение для расчетов и анализа
3. Эмпирические законы распределения и их содержательный смысл
4. Переход от эмпирических законов распределения к теоретическим
5. Алгоритм проверки гипотезы о законе распределения СВ
6. Нормальное распределение



Первичная обработка информации

Существует ли собрана информация (выборка) – данные, характеризующие анализируемый объект или ситуацию:

- 1) Какую полезную информацию предоставляет собранный материал?
- 2) Как обработать эти данные и установить существующие закономерности для того, чтобы в дальнейшем достаточно достоверно прогнозировать?
- 3) Какие значения показателя являются наиболее типичными (ожидаемыми)?
- 4) Как, например, используя имеющуюся статистику, оценить шансы того, что изучаемый показатель будет находиться в некотором конкретном диапазоне значений или не превысит некоторого критического уровня?

Как обработать первичную информацию

1. Для одномерных массивов группировка данных (разбиение на интервалы);
2. Рассчитать частоты (количество наблюдений);
3. Рассчитать относительные частоты (оценка вероятности событий, соответствующих созданным интервалам);
4. Построить гистограмму ряда распределения (по оси абсцисс – интервалы, по оси ординат – относительные частоты);
5. Также можно провести кластерный анализ рыночной информации;
6. Применить инструмент описательная статистика и провести анализ типичных значений случайной величины (средние, наиболее ожидаемые; различные меры вариации, с помощью которых оценивают диапазоны изменения возможных значений исследуемой величины; законы распределения, показывающие какие значения встречаются чаще, а какие реже).

Описательные статистики

- Гистограмма – это диаграмма частот, а не данных.
- Столбиковая диаграмма показывает частоту повторения каждого значения переменной.
- Если значение переменной измерены в номинальной шкале, то применяется столбиковая диаграмма, если в количественной шкале, то гистограмма.

Стандартное отклонение – число, описывающее насколько значения переменной обычно отличаются от среднего.

Если стандартное отклонение большое по отношению к среднему значению, то значения переменной чаще заметно отличаются от среднего значения.

Обращать внимание на выбросы: правило «трех сигм»: все наблюдения, отличающиеся от среднего больше, чем на 3 стандартных отклонения, либо вовсе отбрасываются, либо рассматриваются отдельно.

Ящиковые диаграммы

Процентилью уровня $p\%$ называется число, меньше которого $p\%$ наблюдений, а больше которого $(100-p)\%$.

Процентили уровня 25, 50 и 75% называются квартилями. Вся выборка делится ими на 4 части.

5%-ное усеченное среднее – из выборки исключают 2,5% самых маленьких наблюдений и 2,5% самых больших.

Ящиковая диаграмма - прямоугольник (межквартильный размах) представляет собой пространство между первым и третьим квартилями, т.е. от 25 до 75% данных попадают в него.

Линия внутри прямоугольника соответствует медиане, т.е. значению, левее и правее которого лежит половина всех значений.



Законы распределения и их применение для расчетов и анализа

Математическое ожидание, дисперсия, стандартное отклонение в сжатой форме отражают наиболее существенные особенности каждого конкретного распределения. Многие задачи анализа можно решать, оперируя только числовыми характеристиками с.в., без рассмотрения законов распределения.

Если использовать эмпирический закон распределения (гистограмму) и на его основе подобрать соответствующий ему теоретический закон распределения из тех, которые хорошо изучены и имеют аналитическое описание, то можно провести более глубокое исследование и получить более обоснованные результаты об изучаемом экономическом процессе.



Законы распределения и их применение для расчетов и анализа

Закон распределения дискретной С.В. – всякое соотношение, связывающее между собой ее возможное значение и соответствующие им вероятности, можно представить таблицей (рядом распределения).

Примеры дискретных С.В. – число покупателей, проданных единиц товара и т.д.

В эмпирическом аналоге закона распределения оценками вероятностей служат относительные частоты (доли), в качестве значений С.В. Используют значение середины интервала.

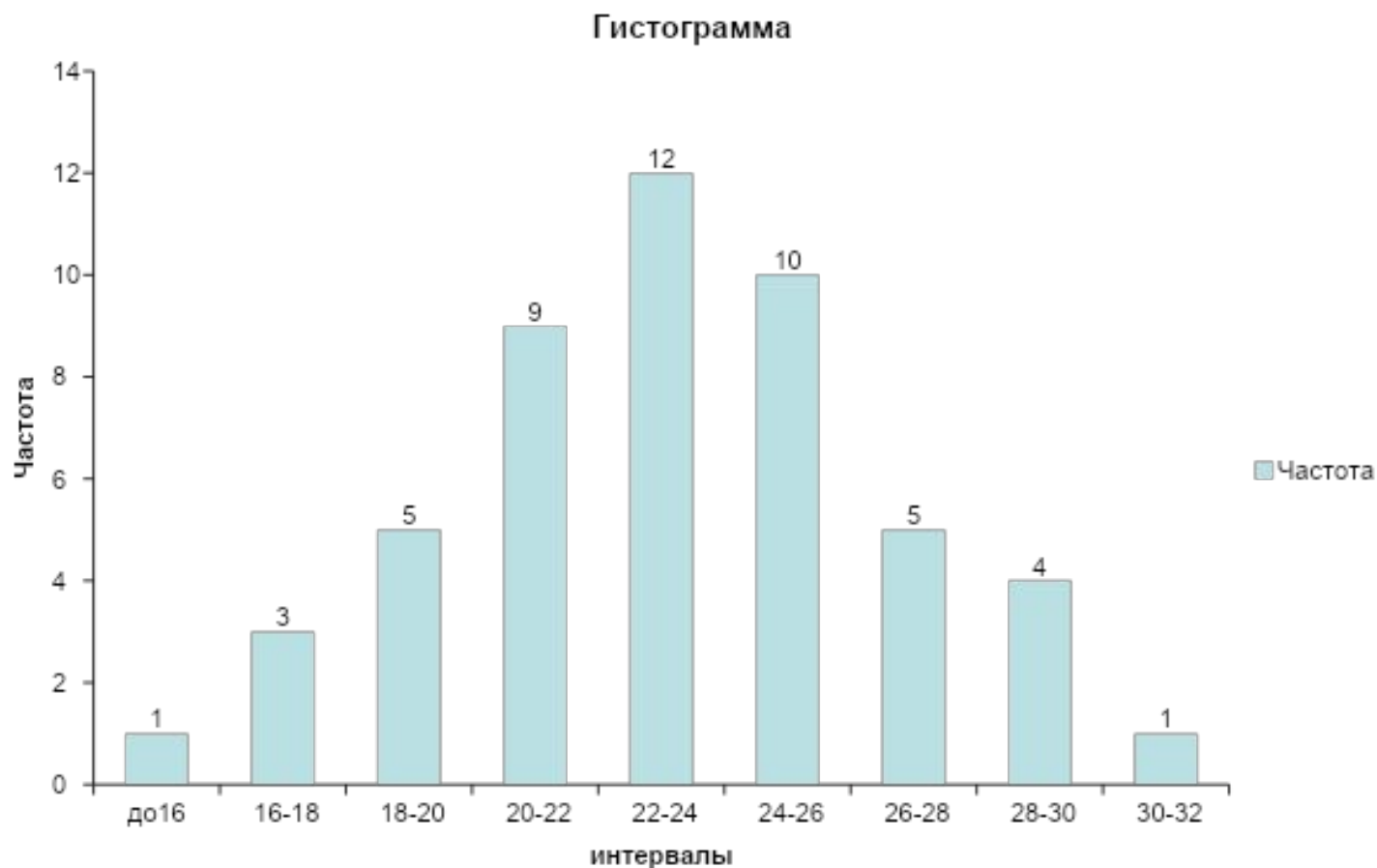
Непрерывные С.В. –экономические показатели такие, как сроки службы объекта, время обслуживания клиента, вес груза и т.д.

Для задания закона распределения таких величин используют не вероятность события $X=x$, а вероятность события $X<x$. Для этого вводят специальную функцию распределения непрерывной С.В.

$F(x)=P(X<x)$, а также функцию плотности вероятности $f(x)$.

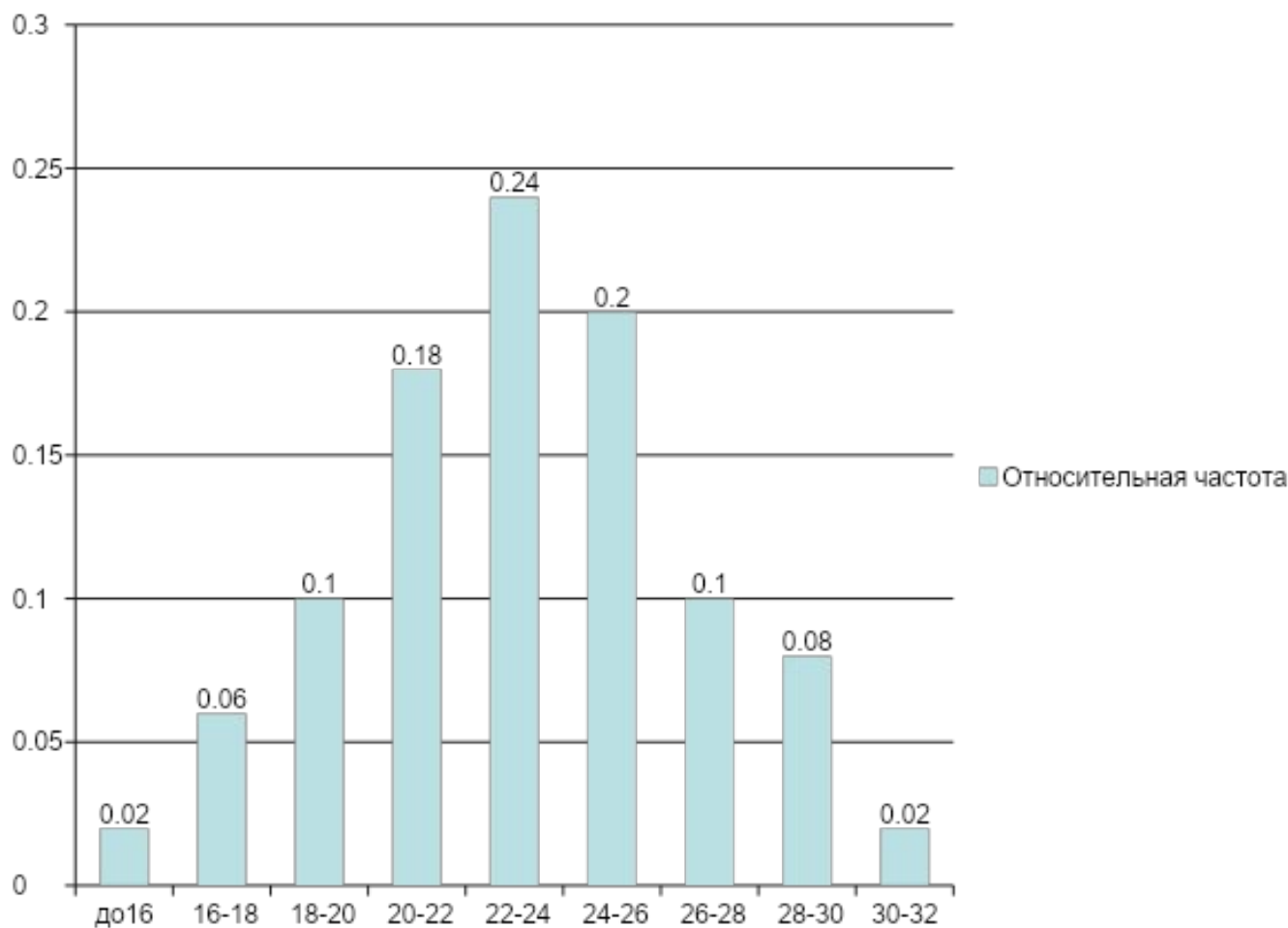
Эмпирические законы распределения и их содержательный смысл

Эмпирическое распределение используют для получения различных оценок (гистограмма частот)



Эмпирические законы распределения и их содержательный смысл

Эмпирическое распределение используют для получения различных оценок (гистограмма относительных частот)





Переход от эмпирических законов распределения к теоретическим

Установление вида закона распределения для исследуемой СВ на основе выборочных данных – задача математической статистики.

Для проведения аналитических исследований и расчетов, связанных с изучением характеристик реальной СВ, используют известные законы распределения.

Для проверки гипотез о предполагаемом законе распределения СВ используют **критерии согласия**. С их помощью устанавливают, достаточно ли хорошо опытные данные согласуются с предполагаемым законом распределения. Если да, то теоретический закон распределения используют для описания СВ.

Основанием для выдвижения гипотезы о том, каков вид (тип) закона распределения СВ, может служить ее эмпирический ряд распределения, полученный методом группировки данных, и его графическое представление (гистограмма).



Алгоритм проверки гипотезы о законе распределения СВ

1. Этап выдвижения нулевой гипотезы. Нулевая гипотеза утверждает, что различие между сравниваемыми эмпирическим и теоретическим распределениями отсутствуют, а наблюдаемые отличия объясняются лишь случайными отклонениями. Если выдвинутая гипотеза отвергается, то принимается альтернативная гипотеза.
Построить эмпирический ряд распределения (относительная частота); Построить гистограмму. Вид гистограммы – основание для выбора предполагаемого теоретического закона распределения.
2. Используя функцию плотности вероятности предполагаемого теоретического распределения, вычисляют вероятности попадания случайной величины в каждый из интервалов. Для их нахождения используют стандартные функции Excel.
3. Для проверки нулевой гипотезы применяют критерий согласия (Хи-квадрат), который измеряет расхождения между эмпирическим и теоретическим законом распределения. В качестве меры расхождения используют сумму квадратов отклонений относительных частот от теоретических вероятностей.

Алгоритм проверки гипотезы о законе распределения СВ

Статистика $\chi^2 = \sum_{i=1}^m \frac{N \left(\frac{n_i}{N} - p_i \right)^2}{p_i}$ имеет распределение хи-квадрат с $(m-k-1)$ степенями свободы, где m - число карманов, k - число параметров теоретического распределения.

4. По формуле вычисляют значение критерия согласия для рассматриваемой выборки (выборочное).

5. Для заданного уровня значимости α (уровня надежности $\gamma=1-\alpha$) и числа степеней свободы $(m-k-1)$ находят границу односторонней критической области χ_{α}^2 - квантиль уровня α (функция хи2.обр (α ; $m-k-1$)).

6. Сравнить наблюдаемое значение хи-квадрат с критическим значением и принять решение:

Если $\chi_{\text{наб}}^2 < \chi_{\alpha}^2$, то для отклонения нулевой гипотезы нет оснований. Считается, что выбранная теоретическая функция распределения согласуется с опытными данными.

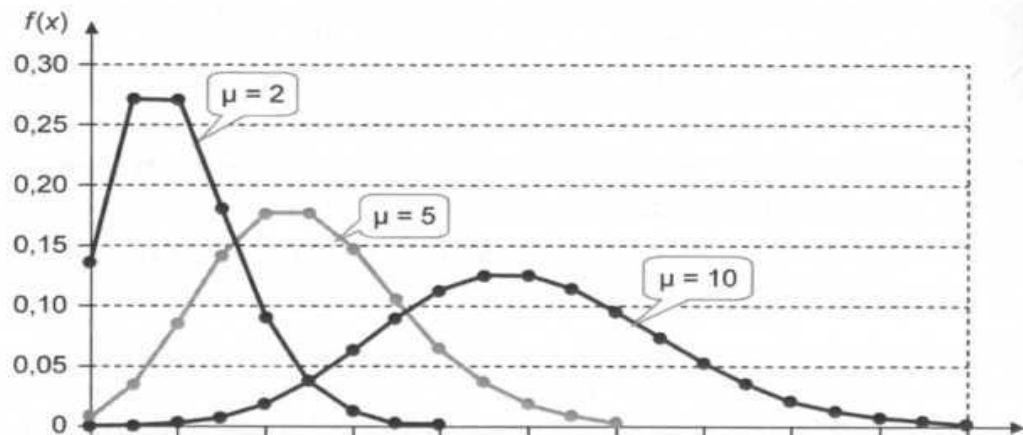
В противном случае нулевая гипотеза отвергается в пользу альтернативной.

Основные распределения СВ

Распределение Пуассона предназначено для описания дискретных случайных величин. Его удобно использовать при исследовании ситуаций, в которых необходимо оценить число появления событий (например, число обслуженных клиентов или посетителей за заданный промежуток времени).

$$P(X = n) = \frac{\mu^n e^{-\mu}}{n!}, \quad n = 0, 1, 2, \dots$$

Соотношение вычисляется вероятность того, что случайная величина X примет значение, точно равное n . В Excel ПУАССОН.РАСП (...)

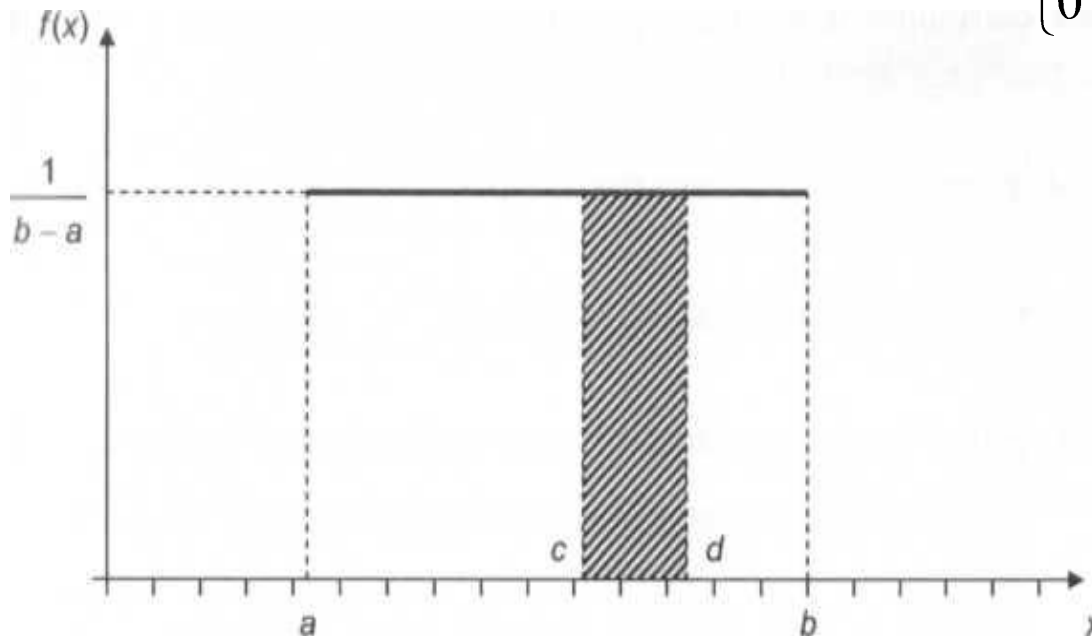


Основные распределения СВ

Равномерное распределение

Если значения случайной величины внутри определенных границ равновероятны, говорят, что случайная величина имеет **равномерное распределение**. Распределение полностью определяется двумя параметрами — границами интервала.

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{иначе } a \leq x \leq b, \\ 0 & \text{иначе } x < a, x > b \end{cases}$$

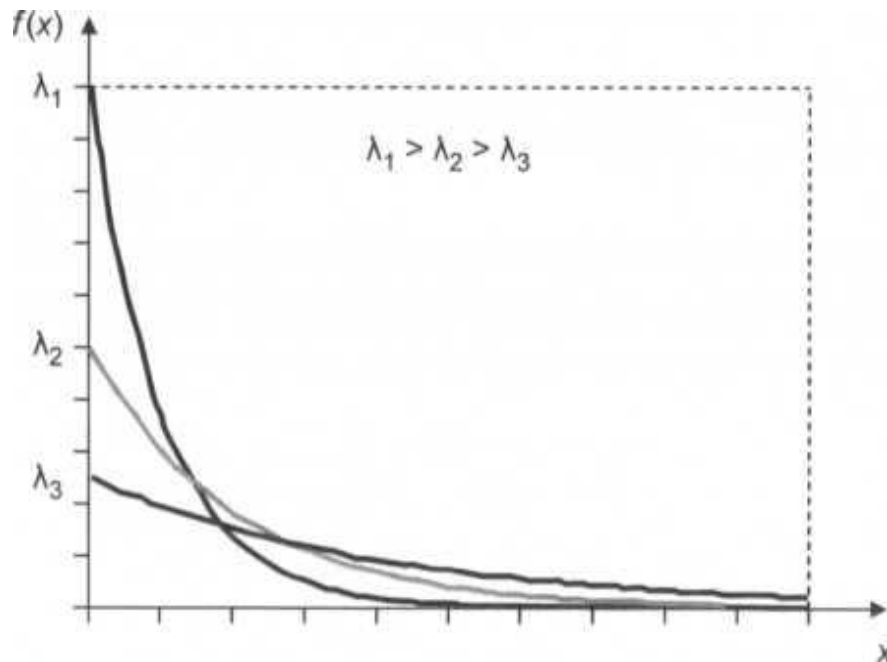


Основные распределения СВ

Экспоненциальное распределение

хорошо описывает процессы «ожидания в очереди», время обслуживания клиентов и заявок в системах, которые называют системами массового обслуживания

Для проведения расчетов и оценок случайных величин, распределенных по экспоненциальному закону, в *Excel* предусмотрена стандартная функция **ЭКСП.РАСП**



$$P(a < x < b) = \lambda \int_a^b e^{-\lambda x} dx = e^{-\lambda a} - e^{-\lambda b}$$

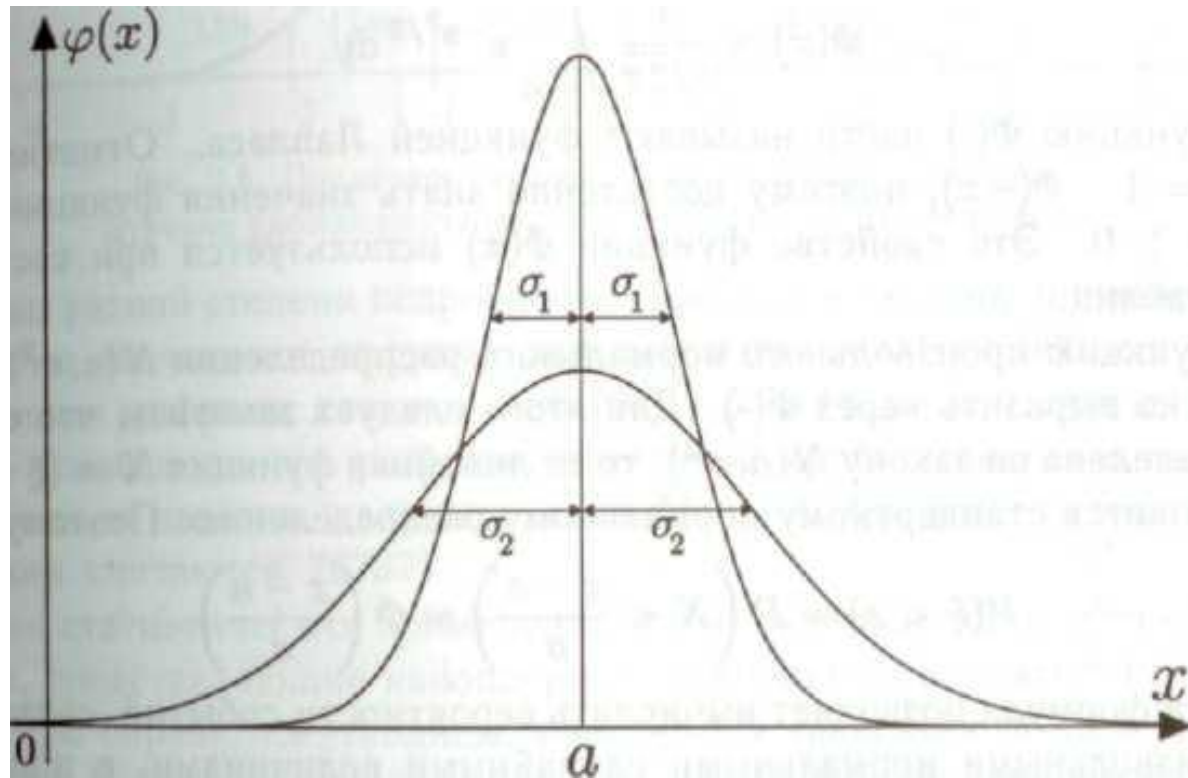


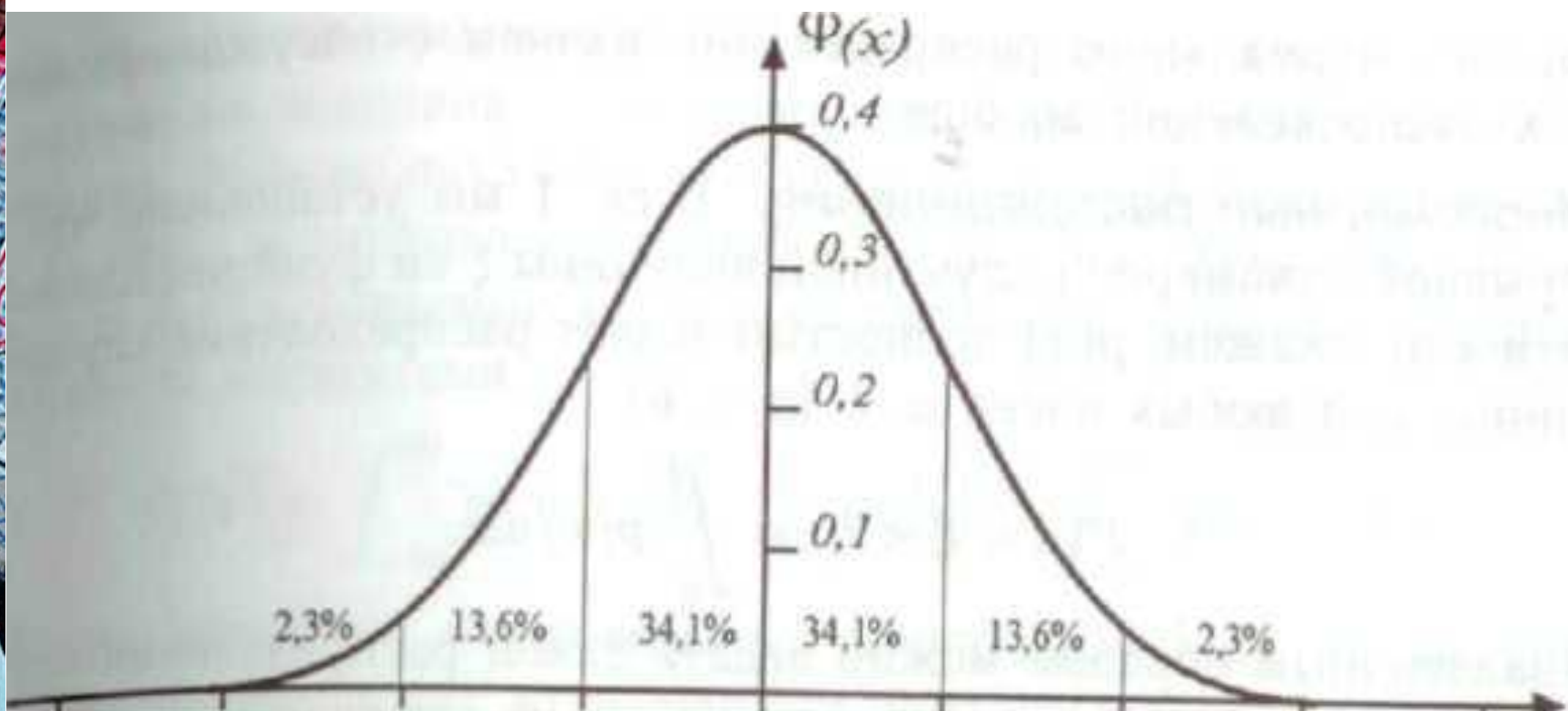
Нормальное распределение

- *Область применения.* Нормальное распределение используется для приближенного описания многих случайных явлений, в которых на интересующий нас результат воздействует большое количество независимых случайных факторов, среди которых нет сильно выделяющихся.
- *Замечание.* Использованию нормального распределения для приближенного описания распределении случайных величин не препятствует то обстоятельство, что эти величины обычно могут принимать значения только из какого-то ограниченного интервала, а нормальное распределение не сосредоточено ни на каком интервале. Дело в том, что вероятность больших отклонений нормальной случайной величины от центра распределения настолько мала, что ее практически можно считать равной нулю

Нормальное распределение

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





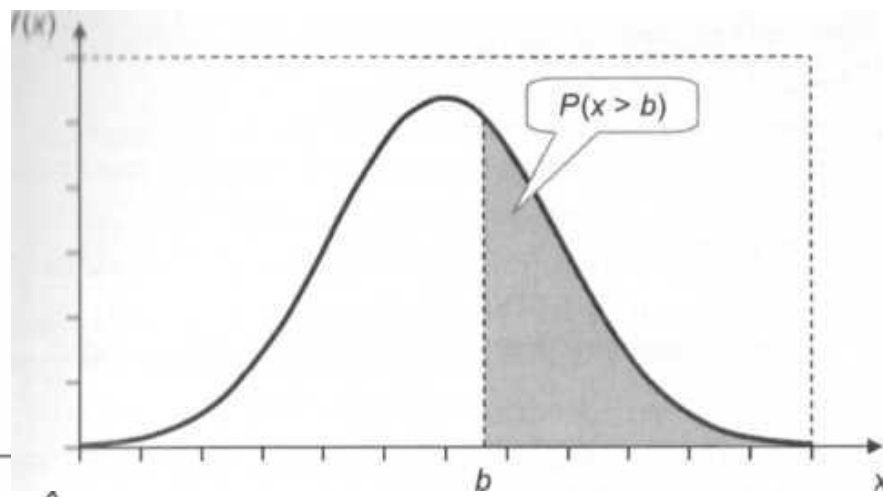
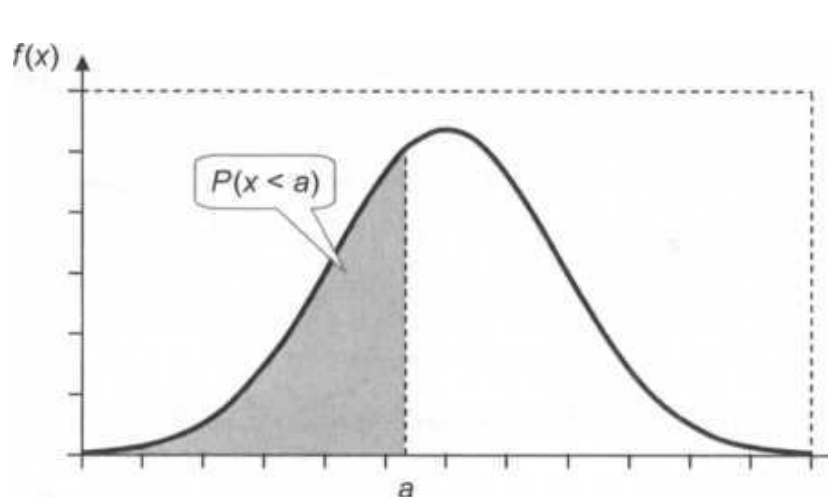
Примерное распределение площадей под кривой функции плотности стандартного нормального распределения

Нормальное распределение

Если для нормально распределенной случайной величины известны значения μ , σ или их оценки, вычисленные на основе имеющейся статистики, то с помощью стандартных функций *Excel* можно вычислить вероятность того, что случайная величина окажется в том или ином диапазоне значений.

Для вычисления в *Excel* необходимой вероятности в какой-либо из ячеек рабочего листа *Excel* надо записать правые части одной из следующих формул:

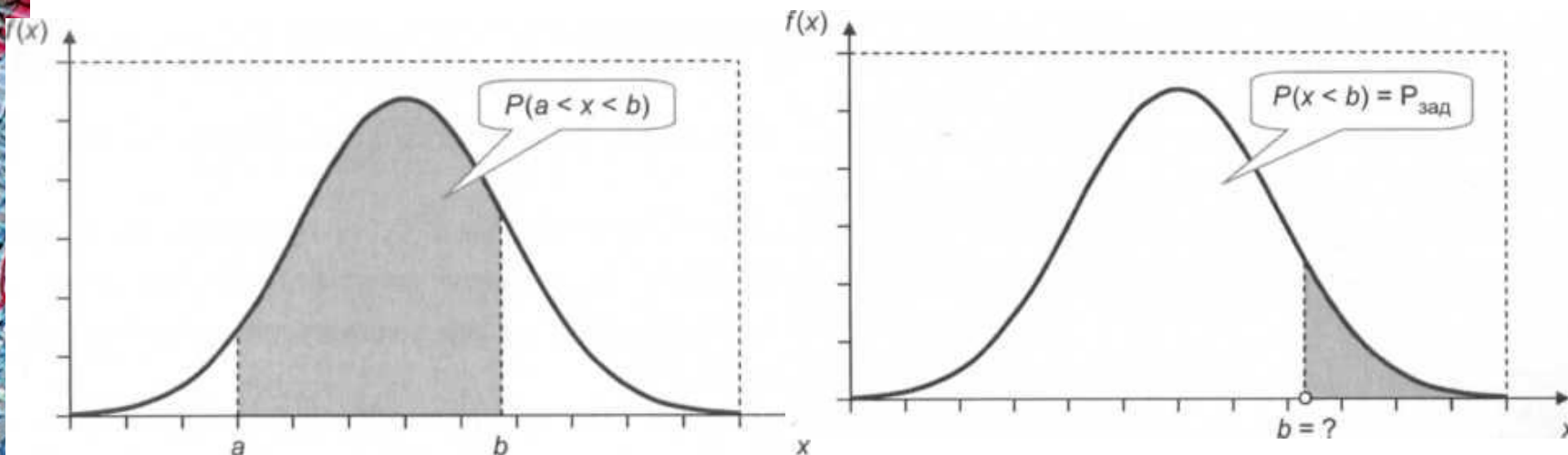
вероятность того, что значения случайной величины X примут значения, меньшие, чем a (рис.):
 $P(X < a) = \text{НОРМ.РАСП}(a; \mu; \sigma; \text{ИСТИНА});$



вероятность того, что значения случайной величины X примут значения, большие, чем b
 $P(X > b) = 1 - \text{НОРМ.РАСП}(b; \mu; \sigma; \text{ИСТИНА});$

Нормальное распределение

$P(a < X < b) = \text{НОРМ.РАСП}(b; ; \mu; \sigma; \text{ИСТИНА}) - \text{НОРМ.РАСП}(a; ; \mu; \sigma; \text{ИСТИНА})$.



Во многих задачах на практике весьма актуальной бывает обратная задача. А именно: требуется найти такое граничное — **критическое** — значение b , которое с заданной вероятностью $P_{\text{зад}} = \gamma$ не будет превышено. Критическое значение b называют **квантилем уровня γ** .

Для решения такой задачи в *Excel* предусмотрена специальная функция **НОРМ.ОБР (...)**. Вызов функции: *Формулы -> Вставить функцию* —> Категория: «*Статистические*» —> **НОРМ.ОБР (...)**



Нормальное распределение

Для начинающих рекомендуется количественные переменные, значительно отличающиеся от нормальных, анализировать теми же методами, что и переменные, измеренные в порядковой шкале.

Как отличить переменные, имеющие нормальное распределение?

1. Визуальный метод (анализ гистограмм). Гипотеза о нормальности распределения переменной не принимается, если есть выраженная ассиметрия гистограммы и наличие выбросов, очень выраженное отклонение от колоколообразной формы..
2. Ассиметрия и эксцесс. Практика показала (а теория подтвердила), что оба показателя неустойчивы (вариабельны). В настоящее время не рекомендуется пользоваться этими характеристиками.
3. Провести тест на нормальность. Часто тест отвергает гипотезу нормальности. Для переменных, имеющих нормальную форму, применяют параметрические методы. Для переменных, имеющих другую форму распределения применяют непараметрические методы.

Логарифмически нормальное распределение

В ряде экономических задач, связанных с изучением распределения доходов, заработных плат, сроков эксплуатации, долговечности устройств и некоторых других, встречаются **логарифмически нормальные** распределения (**логнормальные**).

Непрерывная случайная величина X имеет логарифмически нормальное распределение, если ее натуральный логарифм $\ln(X)$ подчинен нормальному закону распределения $\ln(X) \sim N(a; \sigma)$.

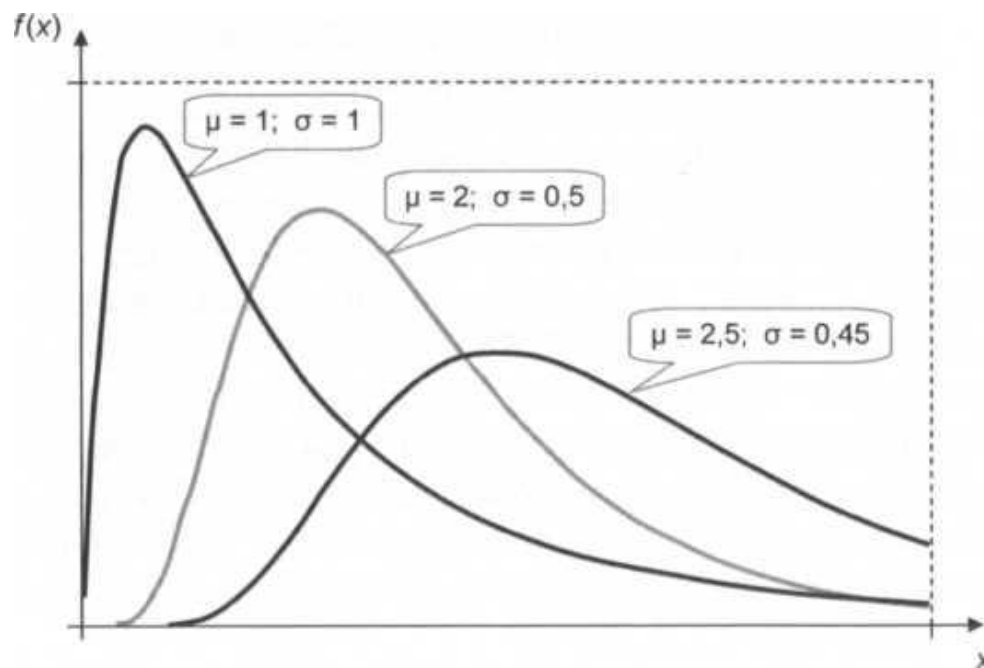
Функция плотности вероятности случайной величины X , имеющей логарифмически нормальное распределение, имеет вид:

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

Для проведения расчетов и оценок случайных величин, распределенных по логнормальному закону, в *Excel* предусмотрена стандартная функция **ЛОГНОРМ.РАСП (...)**.

Логарифмически нормальное распределение

В ряде экономических задач, связанных с изучением распределения доходов, заработных плат, сроков эксплуатации, долговечности устройств и некоторых других, встречаются **логарифмически нормальные** распределения (логнормальные).



Типовые распределения статистики

Основные распределения в статистике

При выполнении статистических вычислений, связанных с нахождением различных оценок на основе выборочных данных (например, выборочная средняя, выборочное стандартное отклонение, коэффициенты уравнения регрессии и т.д.) возникает следующая проблема.

Данные выборки — случайные величины, а конкретный набор выборочных значений (например, $x_1, x_2, \dots, x_N$) можно рассматривать как одну из многих реализаций некоторой многомерной случайной величины, компоненты которой независимы и имеют один и тот же закон распределения $f(x)$, соответствующий генеральной совокупности.

Любая из **выборочных оценок** — например, выборочная средняя или выборочная дисперсия S^2 является *функцией случайной величины* и, следовательно, также случайной величиной с некоторым законом распределения.

Если законы распределения выборочных данных известны (чаще всего предполагается, что это нормально распределенные случайные величины), то, зная аналитическую формулу, по которой



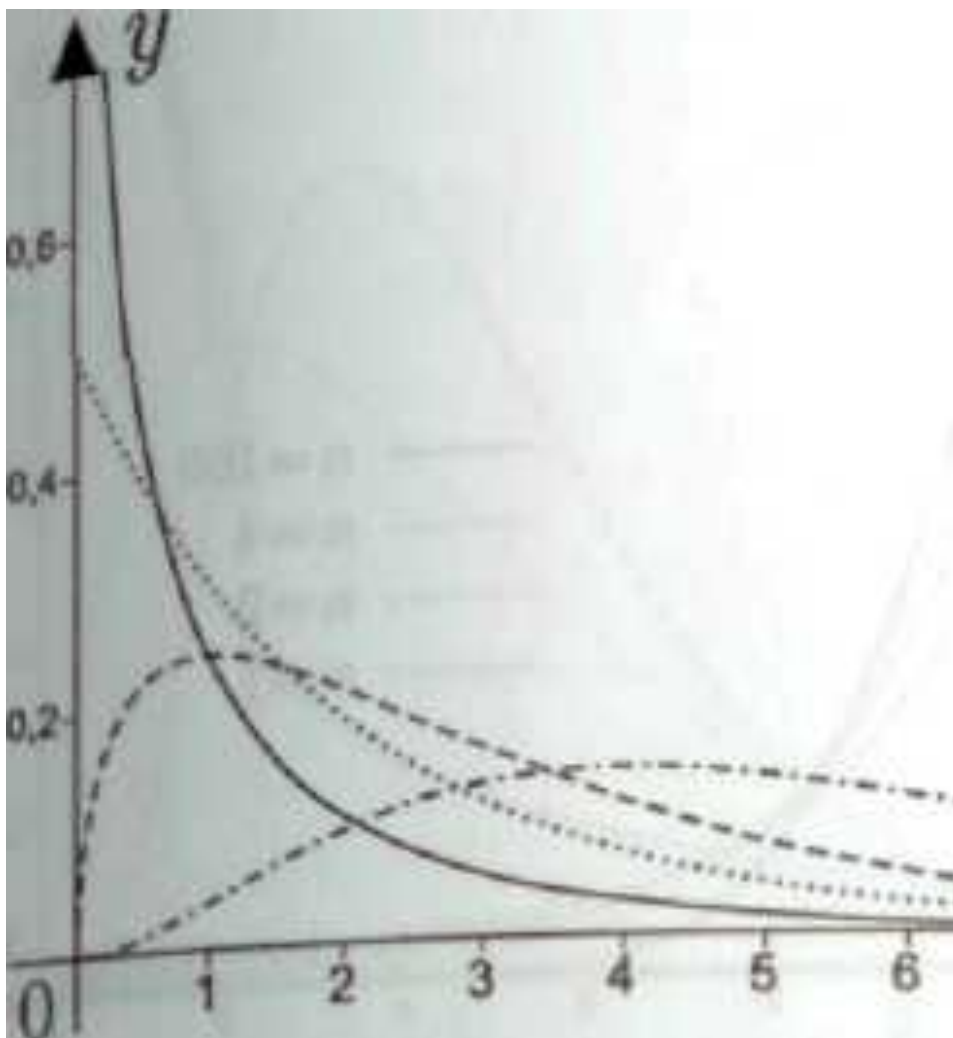
Типовые распределения статистики

по которой получается та или иная оценка, можно получить закон распределения и для результата, т.е. для самой оценки.

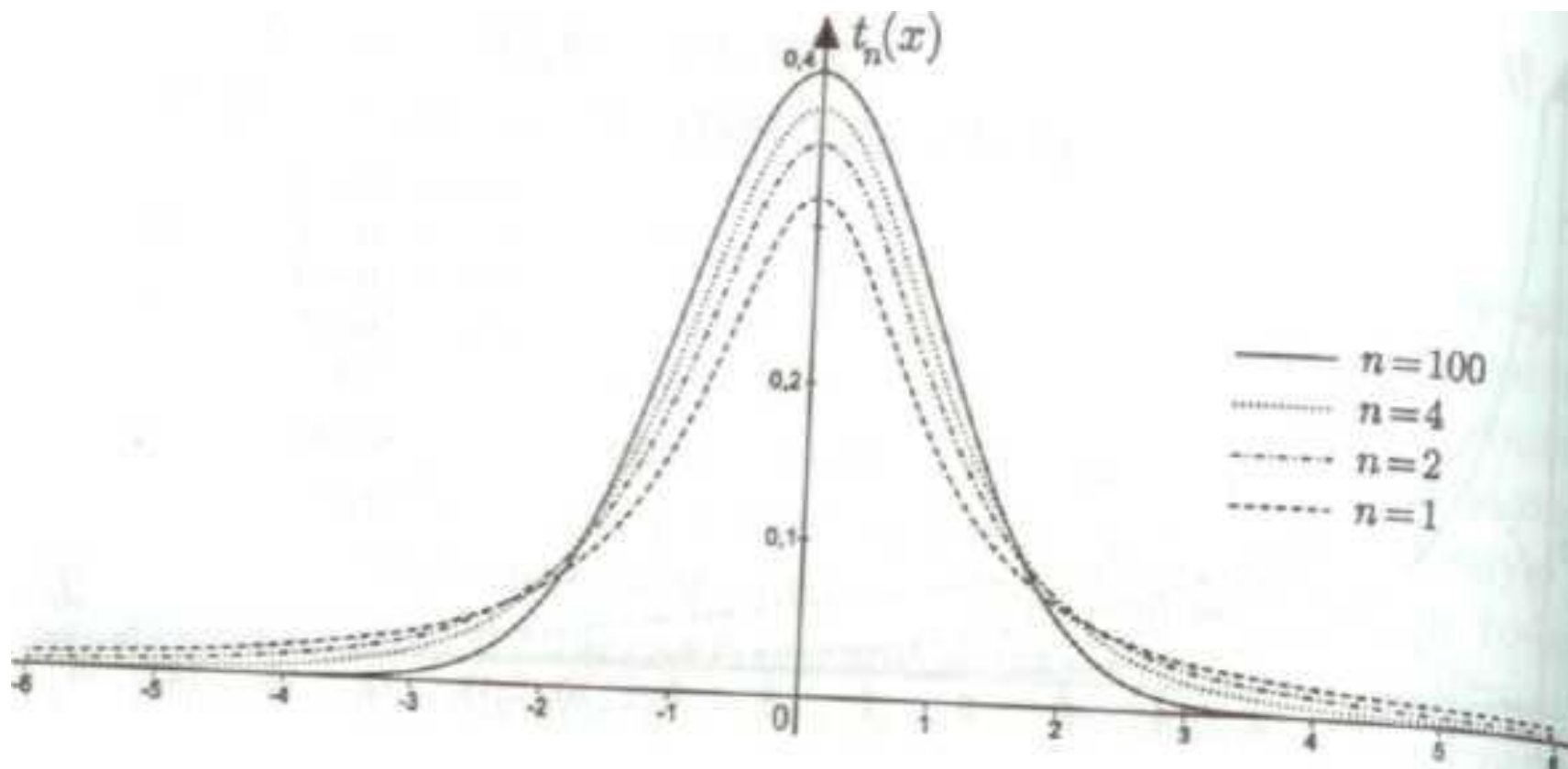
Затем на основе

Полученного закона распределения несложно определить доверительные области для истинной числовой характеристики, приближенное значение которой дает оценка, и решать много других задач, связанных с анализом точности и достоверности результатов, полученных на основе выборочных данных.

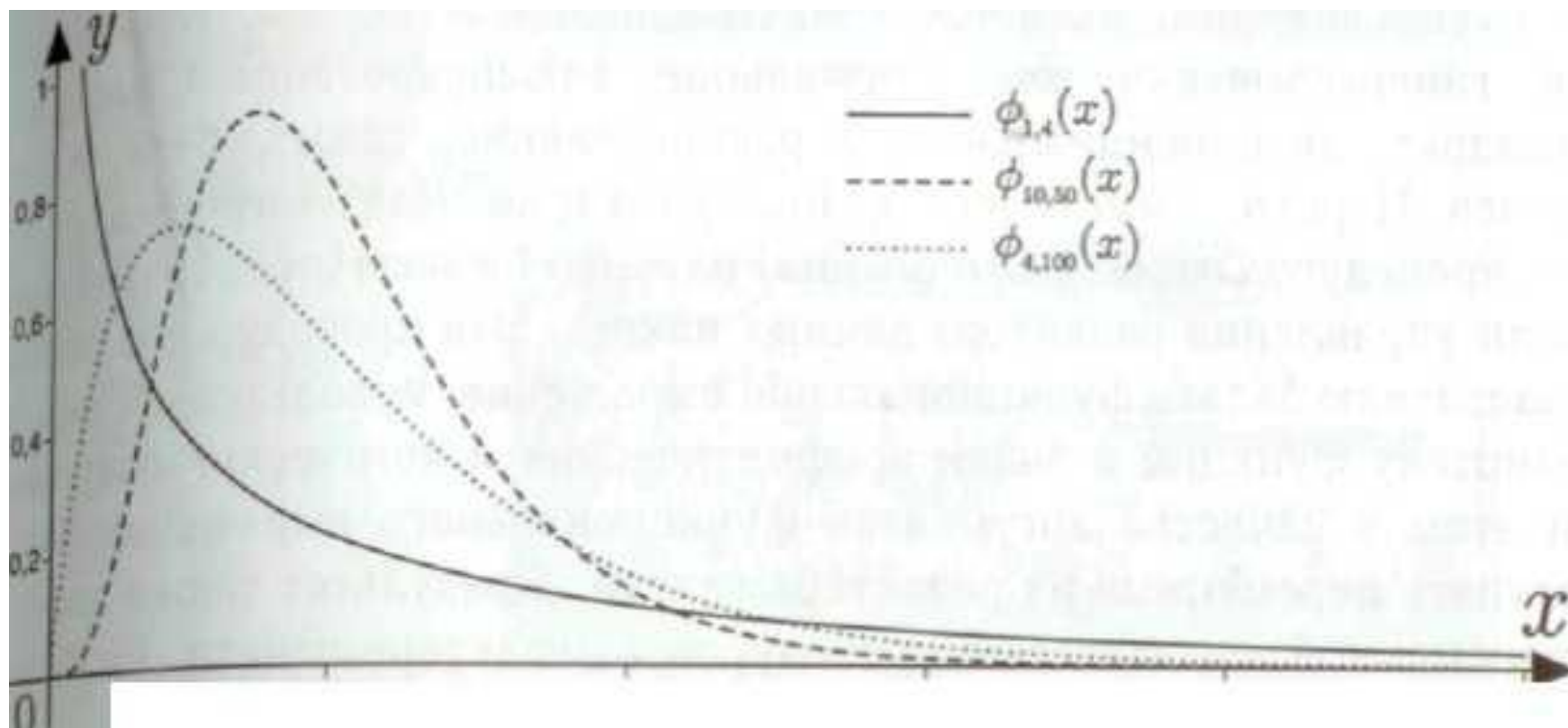
Большинство выборочных оценок связано с вычислением либо *суммы значений*, либо *суммы их квадратов*, либо *комбинаций суммы и суммы квадратов* случайных величин. В теории вероятностей для этих основных функций от случайных аргументов получены соответствующие распределения. К таким типовым распределениям относятся распределение хи-квадрат, распределение Стьюдента, распределение Фишера и ряд других. При этом, как правило, предполагается, что выборочные данные — независимые и нормально распределенные случайные величины.



Функции плотности распределения хи-квадрат с различным числом степеней свободы n



Функции плотности распределения Стюдента с различным числом степеней свободы n



Функции плотности F -распределения с различным числом степеней свободы