Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт космических и информационных технологий
Кафедра систем искусственного интеллекта

УТВЕРЖДАЮ
Заведующий кафедрой СИИ
_____ Г. М. Цибульский
подпись
«1» 06 2016 г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Информационно-аналитические методы численной обработки данных
в условиях неопределенности

09.04.02 – Информационные системы и технологии
09.04.02.01 – Информационно-управляющие системы

Научный руководитель _____ проф., д-р физ.-мат. наук Б. С. Добронец
подпись, дата

Выпускник _____ Д. И. Корчикова
подпись, дата

Рецензент _____ ст. науч. сотр., канд. физ.-мат. наук А.Н. Рогалев
подпись, дата

Нормоконтролер _____ (с замечаниями) М. А. Аникьева
подпись, дата 10.06.16

Красноярск 2016

Федеральное государственное автономное
образовательное учреждение
высшего профессионального образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт космических и информационных технологий
Кафедра систем искусственного интеллекта

УТВЕРЖДАЮ
Заведующий кафедрой СИИ
_____ П. М. Цибульский
подпись
«11» 06 2016 г.

ЗАДАНИЕ
НА ВЫПУСКНУЮ КВАЛИФИКАЦИОННУЮ РАБОТУ
в форме магистерской диссертации

Студенту Корчиковой Дарье Игоревне

Группа КИ14-05М-1М Направление (специальность) 09.04.02 Информационные системы и технологии

Тема магистерской диссертации: информационно-аналитические методы численной обработки данных в условиях неопределенности

Утверждена приказом по университету № 8168/С от 14.06.2016

Руководитель ВКР: Б. С. Добронец, доктор физико-математических наук, профессор кафедры «Системы искусственного интеллекта», Институт космических и информационных технологий Сибирского федерального университета.

Исходные данные для ВКР: диссертационные исследования, материалы с преддипломной практики, книги, научные журналы и статьи зарубежных и отечественных авторов, монографии по теме исследования.

Перечень разделов ВКР: введение, анализ проблемы исследования, анализ методов и подходов к обработке неопределенностей, разработка модуля для численного моделирования эмпирических данных, заключение.

Перечень графического материала: плакаты презентации, структурная схема программы.

Руководитель _____ Б. С. Добронец
подпись

Задание принял к исполнению _____ Д. И. Корчикова
подпись

«30» 01 2016 г.

Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт космических и информационных технологий
Кафедра систем искусственного интеллекта

УТВЕРЖДАЮ
Заведующий кафедрой
_____ Г. М. Цибульский
«____»_____ 2016 г.

ГРАФИК
НАПИСАНИЯ И ОФОРМЛЕНИЯ ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ
в форме магистерской диссертации

Студент: Корчикова Дарья Игоревна

Группа: КИ14-05-1М Направление (специальность): 09.04.02 Информационные системы и технологии

Тема выпускной квалификационной работы: Информационно-аналитические методы численной обработки данных в условиях неопределенности

График выполнения выпускной квалификационной работы (ВКР) приведен в таблице 1.

Таблица 1 – График выполнения этапов ВКР

Наименование и содержание этапа	Срок выполнения	Примечание
Подбор литературы, ее изучение и анализ. Составление списка литературы по основным источникам	До «15»февраля 2015 г.	
Составление плана ВКР и согласование его с руководителем	До «20» марта 2015 г.	
Разработка и представление на проверку первой главы	До «05» мая 2015 г.	
Накопление, систематизация, анализ практических материалов	До «05» сентября 2015 г.	
Разработка и представление на проверку второй главы	До «12» ноября 2015 г.	
Разработка и представление на проверку третьей главы	До «21» декабря 2015 г.	
Согласование с руководителем выводов и предложений	До «12» февраля 2016 г.	
Переработка (доработка) ВКР в соответствии с замечаниями и представление ее на кафедру	До «10» апреля 2016 г.	
Разработка тезисов доклада для защиты	До «28» мая 2016 г.	
Ознакомление с отзывом и рецензией	До «13» июня 2016 г.	
Завершение подготовки к защите с учетом отзыва и рецензии	До «20» июня 2016 г.	

Руководитель ВКР _____ Б. С. Добронец

Студент _____ Д. И. Корчикова

«30» 01 2016 г.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Темой магистерской диссертации является информационно-аналитические методы численной обработки данных в условиях неопределенности. При принятии управленческих решений в условиях, когда необходимая информация доступна, основная задача сводится к принятию рационального выбора оптимального варианта путем вычисления функции полезности. Однако, на практике чаще всего происходит, что получена только неопределенная информация, когда для принятия решений требуется применение более сложных методов представления и расчетов, при этом основой выбора оптимального решения в условиях неопределенностей могут стать функция полезности, функция плотности вероятности. На практике, при решении различных задач неопределенность в данных может препятствовать принятию оптимальных решений.

Для многих практических задач характерна, так называемая, неопределенность в данных, которая может существенно влиять на результаты вычисления различных показателей и характеристик, которые необходимо рассчитать в рамках исследуемой проблемы. В зависимости от предмета решаемой задачи неопределенности данных принимают отраслевой характер, например, экономическая неопределенность.

При решении различных задач, связанных с анализом экспериментальных данных, полученных в результате проведения каких-либо опытов и отображающих поведение параметров системы, перед аналитиками возникает вопрос, какой способ представления этих данных позволит провести наиболее эффективный анализ, который наглядно продемонстрирует возможные исходы и позволит принять правильные решения. Стоит отметить, что данные получаемые в ходе различных экспериментов, носят случайный характер.

Специфика сложности исследования таких систем обуславливается рядом факторов, которые сводятся к четырем группам: первая группа характеризуется внутренней сложностью системы; вторая – внешней сложностью явлений и процессов, влияющих на систему и взаимодействующих с ней; третья группа определяется недостаточностью информации и знаний о предыстории процесса функционирования системы; четвертая группа факторов связана с разрабатываемыми и применяемыми технологиями анализа систем, которые в настоящее время отличаются сложностью, расширением круга решаемых задач [8].

В связи с этим, целью магистерской диссертации является повышение эффективности обработки данных в условиях неопределенности на основе численных методов и алгоритмов.

Для достижения цели диссертации поставлены следующие задачи:

1) провести анализ проблемной области исследования;
2) провести анализ методов обработки неопределенных данных;
3) разработать программный модуль, предназначенный для обработки данных в условиях неопределенности.

1 Анализ проблемы исследования

1.1 Анализ способов представления неопределенных данных и численные операции над ними

Для повышения достоверности численных процедур в условиях неопределенных данных существуют различные подходы. Например, разрабатываются методы, которые составляют суть понятия «арифметики неопределенных данных».

1.2 Анализ неопределенностей в данных

Для анализа неопределенностей в данных важно рассмотреть различные типы неопределенностей. Анализ публикаций показал, что можно выделить три типа “неопределенных” данных: нечеткие, интервальные и случайные. Кроме того, следует различать неопределенности, которые заключаются в данных. Выделяют элиторную неопределенность, которая определяется изменчивостью процессов и явлений, и эпистемическую неопределенность, которая обусловлена недостатком знаний о системе и характеризуется неопределенностью самих вероятностных оценок [11, 16].

Данные, содержащие случайную неопределенность, задаются некоторыми вероятностными распределениями их возможных значений; «нечеткие» данные задаются лингвистически сформулированными распределениями их возможных значений; данные, содержащие интервальную неопределенность, задаются интервалами их возможных значений без указания какого-либо распределения возможных значений числа внутри заданного интервала [10]. Интервальная неопределенность — это состояние неполного (частичного) знания об интересующей нас величине, когда нам известна лишь ее принадлежность некоторому интервалу, т. е. мы можем указать границы возможных значений этой величины или пределы их изменения. Например, интервальная неопределенность присутствует в измерении объемов продаж, объемов производства, прибыли, оценке себестоимости и других экономических показателях. Это позволяет сделать вывод об интервальном характере ряда экономических данных, которые служат основой принятия экономических решений.

Важно отметить, что каждому типу неопределенности в данных соответствует своя арифметика. Так, для интервальных данных разработана интервальная арифметика, которая является частью интервального анализа. Основная идея интервального анализа состоит в замене арифметических операций и вещественных функций над вещественными числами интервальными операциями и функциями, преобразующими интервалы, содержащие эти числа [20]. Для нечетких данных в 1965 году Заде предложил теорию нечетких множеств, которая в настоящее время превратилась в детально разработанную область с широким спектром приложений к задачам практического характера. Нечеткая арифметика строится на таких понятиях,

как нечеткая величина, степень принадлежности, функция принадлежности [24]. Случайная неопределенность и арифметические операции над случайными числами рассматриваются в теории вероятностей.

Таким образом, неопределенность в данных можно классифицировать по нескольким группам. На рисунке 1 представлена классификация неопределенности в данных.

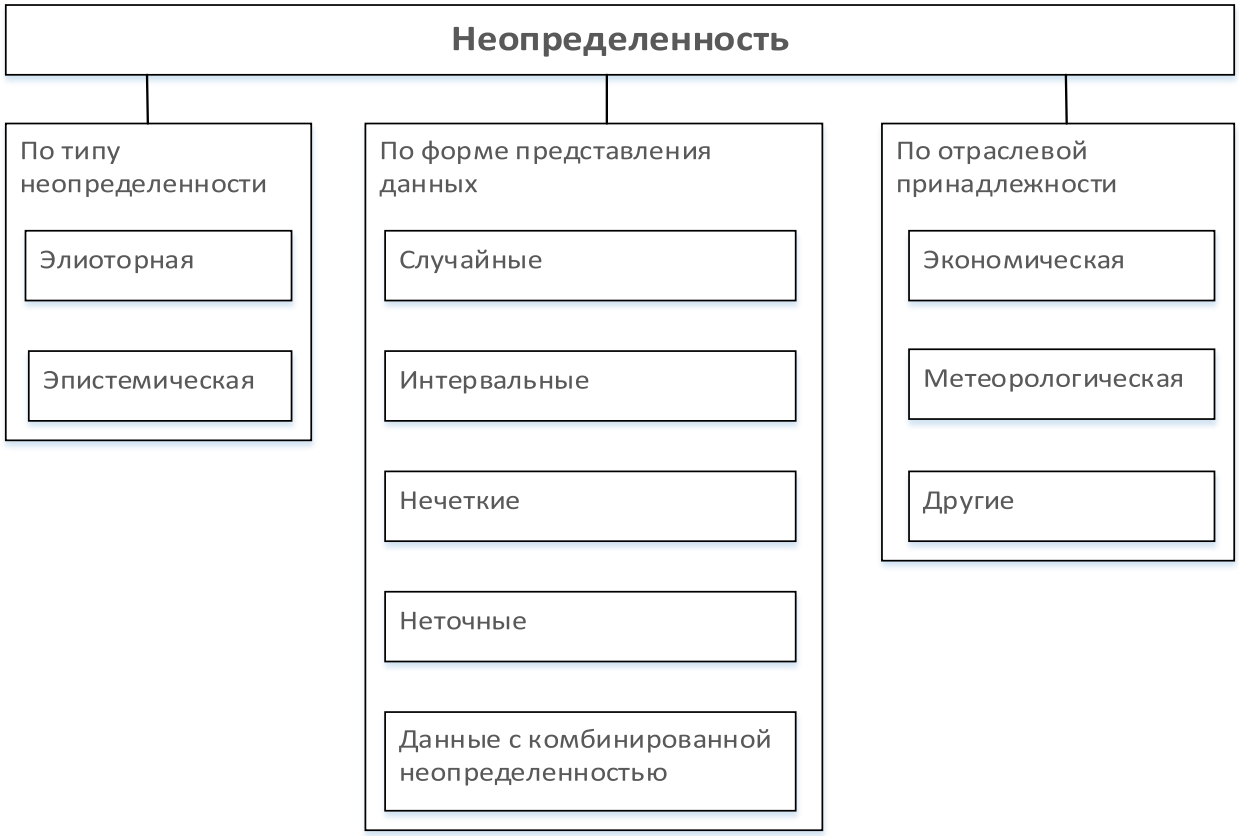

Рисунок 1 – Классификация неопределенности

1.3 Случайные данные. Арифметика над случайными величинами

Исследования в области теории вероятностей ведутся, начиная с семнадцатого века. К настоящему времени она нашла много применений в различных областях науки, техники и организационного управления. Один из важнейших инструментов в этих применениях является статистическое моделирование, которое может быть определено как метод осуществления имитационных экспериментов с моделями стохастических систем. Оно в значительной степени основывается на получении значений случайных величин из их вероятностных распределений.

Статистическое моделирование относится к классу приближенных методов и позволяет получать лишь статистические оценки анализируемых величин, но не их точные значения. К недостаткам статического моделирования можно отнести и то, что это медленный и достаточно дорогой способ решения проблем. Тем не менее, статистическое моделирование

является одним из мощных средств решения сложных задач в случаях, когда подходящие аналитические методы отсутствуют [7,20].

Пусть Ω – множество всех исходов случайного эксперимента. Пусть так же имеется некоторый непустой класс $Ą$ подмножеств (называемых событиями) множества Ω, который обладает следующими свойствами:

– $\Omega \in Ą$;
– если $A \in Ą$, тогда $A^c \in Ą$;
– если $A_n \in Ą$ - счетная последовательность событий, тогда $\cup_n A_n \in Ą$.

Такая система подмножества $Ą$ называется $\sigma-$алгеброй. Для каждого случайного события А существует неотрицательное число $Pr\{A\}$, называемое его вероятностью, такое, что $Pr\{\emptyset\} = 0$, $Pr\{\Omega\} = 1$ и $Pr\{\cup_n A_n\} = \sum_n Pr\{A_n\}$ для каждой счетной последовательности попарно несовместимых событий A_n. Тройка $(\Omega, Ą, Pr)$ называется вероятностным пространством, а функция Pr называется вероятностной мерой.

Стохастическая (случайная) величина на вероятностном пространстве $(\Omega, Ą, Pr)$ представляет собой определенную на Ω и принимающую значения на вещественной прямой R функцию ξ такую, что для борелевского множества O из R $\{\omega \in \Omega \mid \xi(\omega) \in O\} \in Ą$.

Это значит, что множество $\{\omega \in \Omega \mid \xi(\omega) \in O\}$ является событием.

Распределение вероятностей $\Phi: \Re \to [0,1]$ случайной величины ξ определяется как

$$\Phi(x) = \Pr\{\omega \in \Omega | \xi(\omega) \leq x\} \tag{1.1}$$

где $\Phi(x)$ – вероятность того, что случайная величина ξ примет значение меньшее или равное x.

Плотность вероятности (плотность распределения) $\phi: \Re \to \Re$ случайной величины ξ представляет собой кусочно-непрерывную функцию такую, что для всех x выполняется соотношение [24]:

$$\Phi(x) = \int_{-\infty}^{x} \phi(y)dy \tag{1.2}$$

Основа статистического моделирования – получение случайных чисел. Пусть x – случайная величина с кумулятивным распределением вероятностей Ф(•). Поскольку Ф(•) – неубывающая функция, на [0,1] определена обратная функция $\Phi^{-1}(\bullet)$. Предположим, что μ – случайная величина, равномерно распределенная на интервале [0,1]. Тогда имеем, что

$$\Pr\{\Phi^{-1}(\mu) \leq y\} = \Pr\{\mu \leq \Phi(y)\} = \Phi(y) \tag{1.3}$$

откуда следует, что величина $x = \Phi^{-1}(\mu)$ имеет функцию распределения Ф(•). Чтобы получить случайную величину x с распределением Ф(•), можно сформировать равномерно распределенную случайную величину μ, заданную на интервале [0,1], а затем присвоить величине x значение $\Phi^{-1}(\mu)$. Такой процесс носит название метод обратного преобразования.

Однако, для основных известных распределений вместо метода обратного преобразования можно воспользоваться процессами прямого формирования собственноручных случайных чисел. Есть различные вычислительные методы порождения случайных чисел для ряда распределений, таких как: равномерное распределение, распределение Бернулли, биномиальное распределение, распределение Коши, эмпирическое распределение, экспоненциальное распределение, распределение Эрланга, гамма-распределение, бета-распределение, распределение Вейбулла, геометрическое распределение, отрицательное биномиальное распределение, логистическое распределение, нормальное распределение, χ^2- распределение, *F*- распределение, *t*- распределение Стьюдента, логнормальное распределение, многомерное нормальное распределение, распределение Пуассона, треугольное распределение и равномерное распределение в сложной области [24].

Пусть осуществлено N независимых опытов, в результате которых получено N случайных чисел $\varepsilon_1, \ldots, \varepsilon_N$. Записав эти цифры (в порядке появления) в таблицу, получается то, что называется таблицей случайных цифр.

Способ употребления такой таблицы весьма прост. Если в ходе расчета некоторой задачи нам потребуется случайная цифра ε, то мы можем взять любую цифру ε_s из этой таблицы. Если понадобится случайное число γ, то можно взять из таблицы n очередных цифр и считать, что $\gamma = 0, \varepsilon_s \varepsilon_{s+1} \ldots \varepsilon_{s+n-1}$. Выбирать цифры из такой таблицы по порядку не обязательно. Их можно выбирать подряд. Но, конечно, можно начинать с любого места, читать в любом направлении, использовать любой заранее заданный алгоритм выбора, не зависящий от конкретных значений цифр таблицы [30].

Отмечаются некоторые трудности создания таблицы [5].

Во-первых, изготовление хорошей таблицы – весьма сложное дело, ибо в любом реальном опыте всегда возможны ошибки. Например, для изготовления таблицы, содержащей миллион случайных цифр, была построена и тщательно отлажена специальная «рулетка» (с использованием электроники). Тем не менее после некоторого периода хорошей работы она стала выдавать, как показала проверка, не равновероятные цифры. Таким образом, проверка «качества» таблицы абсолютно необходима. Никакие априорные соображения о тщательности постановки опыта не гарантируют нас от ошибок.

Вторая трудность связана с «незаконностью» многократного использования одной и той же таблицы. Вычислителей этот вопрос не очень беспокоит, так как интуитивно ясно, что таблицу можно повторно использовать при решении независимых задач. Возможность использования одних и тех же случайных величин для решения целых классов задач доказана.

Третья трудность заключается в том, что в большой таблице найдутся плохие участки. Например, в таблице, содержащей 10^{100} цифр, вполне

вероятно найти 100 нулей подряд. Очевидно, самостоятельное использование таких участков недопустимо.

Для проверки таблицы случайных цифр предложено использовать четыре теста. В каждом из них цифры классифицируются по некоторому признаку и эмпирические частоты сравниваются с их математическими ожиданиями при помощи критерия χ^2. Тесты эти:

– проверка частот. Проверяется частота различных цифр в таблице;

– проверка пар. Проверяется частота различных двузначных чисел среди пар цифр;

– проверка интервалов. Проверяется частота различных интервалов между двумя последовательными нулями;

– проверка комбинаций. Проверяется частота различных типов четверок среди четверок $\varepsilon_1\varepsilon_2\varepsilon_3\varepsilon_4, \varepsilon_2\varepsilon_3\varepsilon_4\varepsilon_5, \ldots$

С детерминистической точки зрения проверка частот и проверка независимых пар – важнейшие необходимые тесты.

1.4 Интервальные данные. Интервальная арифметика

Недавно было предложено использовать интервалы для представления неопределенности в связи с худшими и лучшими случаями оценки последствия технологического развития. Интервальный подход изначально разработан в 1962 году Муром, чтобы получить нижние и верхние оценки для точных результатов при проведении численных расчетов для цифровых вычислительных машин с «конечным числом значащих цифр» [7].

Следуя Муру, определяется число интервала, как упорядоченную пару $[a, b]$ реальных чисел $a \leq b$. Также она может быть определена как обычные множество действительных чисел x таких, что, $a \leq x \leq b$, или $[a, b] = \{x | a \leq x \leq b\}$.

Далее рассматривается применение интервалов для представления экономической неопределенности, благодаря интервалам можно представлять любую неопределенность экономических параметров для его нижнего и верхнего пределов для того, чтобы иметь возможность выявить экономические последствия этих неопределенностей [25].

Для примера, требуемое количество определенного товара в будущие периоды времени может быть представлено неопределенными интервалами с увеличением ширины в зависимости от времени и расстояния в будущем [33].

Если используются алгебраические операции: сложение, вычитание, умножение и деление, то они обозначаются символом #, можно определить операции на двух отрезках $I_1 = [a_1, b_1]$ и $I_2 = [a_2, b_2]$ на основе множества теоретических формулировок:

$$I_1 \# I_2 = \{x \# y | a_1 \leq x \leq b_1, a_2 \leq y \leq b_2\} \quad (1.4)$$

Вместо теоретико-множественного определения в (1.4) можно дать альтернативное определение с точки зрения конечных точек полученных интервалов [5]:

$$I_1 + I_2 = [a_1 + a_2, b_1 + b_2] \quad (1.5)$$

$$I_1 - I_2 = [a_1 - a_2, b_1 - b_2] \quad (1.6)$$

$$I_1 * I_2 = [\min(a_1a_2, a_1b_2, b_1a_2, b_1b_2), \max(a_1a_2, a_1b_2, b_1a_2, b_1b_2)] \quad (1.7)$$

$$\frac{I_1}{I_2} = [a_1, b_1] * \left[\frac{1}{b_3}, \frac{1}{a_2}\right], \text{если } 0 \notin [a_2, b_2] \quad (1.8)$$

Следует отметить, что в некоторых случаях прямое применение интервальной арифметики приводит к слишком пессимистичным (например, слишком длинные) промежуткам времени. Простым примером этого является расчет выражения $I * (1 - I)$, где I – интервал, $I = [0, 1]$.

Применение вышеуказанных формул дает результат [0, 1], который, очевиден в слишком широком интервале. В соответствии с основными определениями арифметических операций на интервалах, основанных на теории множеств (1.4), возможно получить более узкий интервальный результат $\left[0, \frac{1}{4}\right]$.

1.5 Нечеткие данные. Нечеткая арифметика

Нечёткое множество (иногда размытое, расплывчатое, туманное, путанное, пушистое) — понятие, введённое Лотфи Заде в 1965 году в статье «FuzzySets» в журнале Information and Control, в котором расширил классическое понятие множества, допустив, что функция принадлежности элемента множеству может принимать любые значения в интервале [0,1], а не только значения 0 или 1.

Л. Заде предложил теорию нечетких множеств, которая превратилась за прошедшее время в детально разработанную область с широким спектром приложений к задачам практического характера. Термин «нечеткая величина» впервые был введен Кофманом, а затем появился в работах Заде и Nahmias. Теория возможностей предложена Заде и разрабатывалась многими исследователями, в частности, Дюбуа и Прадом [24].

Нечеткое имитационное моделирование было разработано Лю и Ивамурой и определяется как метод экспериментирования с моделями нечетких систем. Многими численными экспериментами показано, что нечеткое имитационное моделирование прекрасно подходит для работы с нечеткими ограничениями и для оценки возможностей нечетких систем [17].

Обычное множество *A*, выделенное из универсального множества *U*, принято определять как коллекцию элементов $x \in U$. Каждый отдельный элемент может принадлежать либо не принадлежать множеству *A* такому, что

$A \subset U$. Это множество можно описать несколькими способами: можно перечислить те элементы, которые принадлежат множеству; можно описать множество аналитически, с помощью набора равенств и неравенств (ограничений); можно так же определить элементы, принадлежащие множеству, с использованием характеристической функции, которая принимает значение 1 для элементов, принадлежащих рассматриваемому множеству, и 0 – для элементов, не принадлежащих ему. Во многих случаях, однако, на вопрос о принадлежности элемента множеству ответить непросто. Например, для множеств, связанных с понятиями «старый человек», «уважаемый», «подобный», «большое число». Множества этого вида не поддаются истолкованию средствами классической теории множеств или теории вероятностей. Чтобы иметь возможность работать с объектами подобного рода, Заде предложил концепцию нечеткого множества [24].

Обозначим через U универсальное множество. Тогда нечеткое подмножество $\widetilde{\mathrm{A}}$ универсального множества U определяется с помощью функции принадлежности $\mu_{\widetilde{\mathrm{A}}}: U \to [0,1]$, которая ставит в соответствие каждому элементу $x \in U$ действительное число $\mu_{\widetilde{\mathrm{A}}}(x)$ из интервала [0,1], где значение $\mu_{\widetilde{\mathrm{A}}}(x)$ для x представляет собой степень принадлежности элемента x множеству $\widetilde{\mathrm{A}}$. Считается при этом, что чем ближе значение $\mu_{\widetilde{\mathrm{A}}}(x)$ к единице, тем выше степень принадлежности элемента x множеству $\widetilde{\mathrm{A}}$.

Множество элементов, принадлежащих нечеткому множеству $\widetilde{\mathrm{A}}$ и имеющих степени принадлежности не менее α, будем называть множеством уровня α.

$$\widetilde{A_\alpha} = \{x \in U | \mu_{\widetilde{\mathrm{A}}}(x) \geq \alpha\} \tag{1.9}$$

Намиас предложил теоретическую базу, которая позволяет построить аксиоматическую теорию для описания нечеткости. Дадим определение возможностного пространства (пространство паттернов).

Пусть $\Theta -$ непустое множество, а $\mathrm{P}(\Theta) -$ множество всех подмножеств для Θ. Для каждого $A \in \mathrm{P}(\Theta)$, существует некоторое неотрицательное число *Pos{A}*, называемое его возможностью, такое, что $Pos\{\emptyset\} = 0$, $Pos\ \{\Theta\} = 1$ и $Pos\{\cup_k A_k\} = sup_k Pos\{A_k\}$ для некоторого произвольного набора $\{A_k\}$ в $P(\Theta)$.

Тройка $(\Theta, P(\Theta), Pos)$ называется возможностным пространством, а функция *Pos* трактуется как мера возможности.

Нечеткая величина определяется как функция из возможностного пространства $(\Theta, P(\Theta), Pos)$ в вещественную прямую *R*.

Пусть ξ - нечеткая величина на возможностном пространстве $(\Theta, P(\Theta), Pos)$. Тогда ее функция принадлежности может быть получена из меры возможности следующим образом:

$$\mu_{\widetilde{\mathrm{A}}}(x) = \mathrm{Pos}\{\theta \in \Theta \mid \xi(\theta) = \mathrm{x}\} \tag{1.10}$$

Концепция произведения возможностных пространств. Предположим, что $(\Theta_i, P(\Theta_i), Pos_i)$ – возможностные пространства, $i = 1, 2, ..., m$. Тогда

$$\Theta = \Theta_1 * \Theta_2 * ... * \Theta_m. \quad (1.11)$$

Для любого $A \in \mathrm{P}(\Theta)$ введем определение меры возможности следующим образом:

$$Pos\{A\} = \sup_{(\theta_1, \theta_2, .., \theta_m) \in A} \min_{1 \le i \le m} Pos_i\{\theta_i\} \quad (1.12)$$

Пусть $(\Theta_i, P(\Theta_i), Pos_i)$, $i = 1, 2, ..., m$, – возможностные пространства. Произведение возможностных пространств определяется как $(\Theta, P(\Theta), Pos)$, где Θ и Pos определяются посредством (1.11) и (1.12) соответственно.

Пусть $\widetilde{a_i}$ – нечеткие величины, определенные на возможностных пространствах $(\Theta_i, P(\Theta_i), Pos_i)$, $i = 1, 2, ..., m$, соответственно. Тогда функции принадлежности для них, полученные из мер возможности, запишутся как

$$\mu_{\widetilde{a_i}}(x) = Pos_i\{\theta \in \Theta_i | \widetilde{a_i}(\theta) = x\}, i = 1,2. \quad (1.13)$$

Сумма $\tilde{a} = \widetilde{a_1} + \widetilde{a_2}$ представляет собой нечеткую величину, определенную на произведении возможностных пространств $(\Theta, P(\Theta), Pos)$ как

$$\tilde{a}(\theta_1, \theta_2) = \widetilde{a_1}(\theta_1) + \widetilde{a_2}(\theta_2) \quad \forall(\theta_1, \theta_2) \in \Theta \quad (1.14)$$

с функцией принадлежности, определяемой выражением

$$\mu_{\widetilde{a_i}}(x) = \sup_{x_1, x_2 \in R} \{\mu_{\widetilde{a_1}}(x_1) \wedge \mu_{\widetilde{a_2}}(x_2) \big| x = x_1 + x_2\} \quad (1.15)$$

для любого $x \in R$. То есть, возможность того, что нечеткая величина $\tilde{a} = \widetilde{a_1} + \widetilde{a_2}$ достигнет значения $x \in R$, равняется по величине наиболее возможной комбинации действительных чисел x_1, x_2 таких, что $x=x_1+x_2$, где значения переменной $\widetilde{a_i}$ есть x_i, i = 1, 2, соответственно. Произведение $\tilde{a} = \widetilde{a_1} * \widetilde{a_2}$ представляет собой некоторую нечеткую величину, определенную на произведении возможностных пространств $(\Theta, P(\Theta), Pos)$ следующим образом:

$$\tilde{a}(\theta_1, \theta_2) = \widetilde{a_1}(\theta_1) * \widetilde{a_2}(\theta_2) \quad \forall(\theta_1, \theta_2) \in \Theta\,, \quad (1.16)$$

а функция принадлежности для этой величины задается выражением

$$\mu_{\widetilde{a_i}}(x) = \sup_{x_1, x_2 \in R} \{\mu_{\widetilde{a_1}}(x_1) \wedge \mu_{\widetilde{a_2}}(x_2) \big| x = x_1 * x_2\} \quad (1.17)$$

для любого $x \in R$. В более общем случае получаем следующую арифметику.

Пусть $f: R^n \to R$ – непрерывная функция, а ξ_i– нечеткие величины на возможностных пространствах (Θ_i, $P(\Theta_i)$, Pos_i), $i = 1, 2, ..., m$, соответственно. Тогда $\xi = f(\xi_1, \xi_2, \dots, \xi_n)$ представляет собой нечеткую величину, определенную на произведении возможностных пространств (Θ, $P(\Theta)$, Pos) следующим образом:

$$\xi(\theta_1, \theta_2, \dots, \theta_n) = f\big(\xi_1(\theta_1), \xi_2(\theta_2), \dots, \xi_n(\theta_n)\big) \tag{1.18}$$

для любого $(\theta_1, \theta_2, \dots, \theta_n) \in \Theta$.

Теорема. Пусть $\widetilde{a_1}, \widetilde{a_2}, \dots, \widetilde{a_n}$- нечеткие величины и $f: R^n \to R$ - некоторая непрерывная функция. Тогда, функция принадлежности $\mu_{\tilde{a}}$ от $\tilde{a} = f(\widetilde{a_1}, \widetilde{a_2}, \dots, \widetilde{a_n})$ строится по функциям принадлежности $\mu_{\widetilde{a_i}}, \mu_{\widetilde{a_2}}, \dots, \mu_{\widetilde{a_n}}$ следующим образом:

$$\mu_{\tilde{a}}(x) = \underset{x_1, x_2, \dots, x_n \in R}{sup} \left\{ \min_{1 \le i \le n} \mu_{\widetilde{a_i}}(x_i) \,|x = f(x_1, x_2, \dots, x_n) \right\}. \tag{1.19}$$

Эта теорема совпадает с принципом расширения, сформулированным Заде. Проиллюстрируем теперь операции с нечеткими величинами. Под трапецоидальнымми нечеткими величинами будем понимать нечеткие величины, полностью определяемые четверкой (r_1, r_2, r_3, r_4) обычных ("четких") чисел таких, что $r_1 < r_2 \le r_3 < r_4$, а функции принадлежности таких величин определяются выражениями вида [24]:

$$\mu(x) = \begin{cases} \frac{x - r_1}{r_2 - r_1}, & \text{если } r_1 \le x \le r_2 \\ 1 & \text{если } r_2 \le x \le r_3 \\ \frac{x - r_4}{r_3 - r_4}, & \text{если } r_3 \le x \le r_4 \\ 0, & \text{в других случаях} \end{cases} \tag{1.20}$$

Стоит отметить, что трапецоидальная нечеткая величина будет представлять собой треугольную нечеткую величину, если $r_2 = r_3$, т.е. если она полностью представима с помощью тройки (r_1, r_2, r_4).

Основываясь на понятии бинарной операции, можно получить сумму трапецоидальных нечетких величин $\tilde{a} = (a_1, a_2, a_3, a_4)$ и $\tilde{b} = (b_1, b_2, b_3, b_4)$ как

$$\mu_{\tilde{a}+\tilde{b}}(z) = sup\{\min\{\mu_{\tilde{a}}(x), \mu_{\tilde{b}}(y) \,\big|\, z = x + y\} =$$
$$= \begin{cases} \dfrac{z - (a_1 + b_1)}{(a_2 + b_2) - (a_1 + b_1)}, & \text{если } a_1 + b_1 \le z \le a_2 + b_2, \\ 1, & \text{если } a_2 + b_2 \le z \le a_3 + b_3, \\ \dfrac{z - (a_4 + b_4)}{(a_3 + b_3) - (a_4 + b_4)}, & \text{если } a_3 + b_3 \le z \le a_4 + b_4, \\ 0, & \text{в друих случаях} \end{cases} \tag{1.21}$$

Таким образом, сумма двух трапецоидальных нечетких величин также будет трапецоидальной нечеткой величиной:

$$\tilde{a} + \tilde{b} = (a_1 + b_1, a_2 + b_2, a_3 + b_3, a_4 + b_4). \quad (1.22)$$

Рассмотрев произведение трапецоидальной нечеткой величины и скалярной величины λ. Получим:

$$\mu_{\lambda * \tilde{a}}(z) = sup\{\mu_{\tilde{a}}(x) | \, z = \lambda x\}, \quad (1.23)$$

откуда следует, что

$$\lambda * \tilde{a} = \begin{cases} (\lambda a_1, \lambda a_2, \lambda a_3, \lambda a_4), \text{если } \lambda \geq 0 \\ (\lambda a_1, \lambda a_2, \lambda a_3, \lambda a_4), \text{если } \lambda < 0 \end{cases} \quad (1.24)$$

Итак, произведение трапецоидальной нечеткой величины и скалярной величины представляет собой снова трапецоидальную нечеткую величину.

1.6 Вывод по главе 1

Анализ публикаций показал, что неопределенность в данных можно классифицировать по нескольким признакам, а именно: по типу неопределенностей можно выделить элиторную неопределенность, которая характеризуется изменчивостью процессов и состояний систем, эпистемическую неопределенность, характеризующуюся неопределенностью самих вероятностных оценок и недостаточностью знаний о системе; по видам неопределенных данных выделяют случайные, нечеткие, интервальные данные. Данные, содержащие случайную неопределенность, задаются некоторыми вероятностными распределениями их возможных значений; «нечеткие» данные задаются лингвистически сформулированными распределениями их возможных значений; данные, содержащие интервальную неопределенность, задаются интервалами их возможных значений без указания какого-либо распределения возможных значений числа внутри заданного интервала [6]. Следует отметить, что для каждого вида неопределенных данных разработана своя арифметика.

2 Анализ методов и подходов к обработке неопределенностей

Для оценки различных видов неопределенностей следует применять разные методы и подходы.

2.1 Ядерное восстановление функции плотности вероятности

Ядерная оценка плотности вероятности - метод, который основывается на методе ядерного сглаживания [10]. Этот метод прост в применении, не требует дополнительных математических сведений и понятен на интуитивном уровне. Ядерное сглаживание во многих случаях является подходящим средством. Существуют разнообразные альтернативные методы сглаживания такие, например, как сплайны, но в асимптотическом смысле они эквивалентны ядерному сглаживанию [22].

Ключом к проведению качественного непараметрического оценивания является выбор подходящей ширины окна для имеющейся задачи. Хотя ядерная функция K остается важной, ее главная роль состоит в обеспечении дифференцируемости и гладкости получающейся оценки. Ширина окна h, с другой стороны, определяет поведение оценки в конечных выборках, что ядерная функция сделать просто не в состоянии.

Принцип, использующий идейно простой подход к представлению последовательности весов $\{W_{mi}(x)\}_{i=1}^{m}$ состоит в описании формы весовой функции $W_{mi}(x)$ посредством функции плотности со скалярным параметром, который регулирует размер и форму весов около x. Эту функцию формы принято называть ядром K.

Полученные таким образом веса далее используются для представления величины $a(x)$ в виде взвешенной суммы значений y_i обучающей выборки.

Ядро — это непрерывная ограниченная симметричная вещественная функция K с единичным интегралом $\int K(u)du = 1$.

Последовательность весов для ядерных оценок (для одномерного x) определяется как :

$$W_{mi}(x) = \frac{K_{h_m}(x - X_i)}{\widehat{f}_{h_m}(x)} \tag{2.1}$$

где

$$\widehat{f}_{h_m}(x) = \frac{1}{m}\sum_{i=1}^{m} K_{h_m}(x - X_i) \tag{2.2}$$

$$K_{h_m}(u) = \frac{1}{h_m} K\left(\frac{u}{h_m}\right) \tag{2.3}$$

где (2.3) представляет собой ядро с параметром h_m. Этот параметр принято называть шириной окна. Подчеркнув зависимость $h = h_m$ от объема выборки m, условимся сокращенно обозначать последовательность весов $W_{mi}(x)$.

Функция $\widehat{f}_{h_m}(x)$ является ядерной оценкой плотности Розенблата — Парзена для (маргинальной) плотности переменной x. Как следствие, оценка ожидаемой величины восстанавливаемой зависимости $E(y|x)$:

$$\widehat{m}_h(x) = \frac{\frac{1}{m}\sum_{i=1}^{m} K_{h_m}(x-X_i)Y_i}{\frac{1}{m}\sum_{i=1}^{m} K_{h_m}(x-X_i)} \tag{2.4}$$

часто называют оценкой Надарая—Ватсона. Ширина окна определяет, насколько быстро убывают веса $W_{mi}(x)$ по мере удаления объектов x_i от x. Характер убывания определяется видом ядра K. Нормализация весов $\widehat{f}_{h_m}(x)$ гарантирует, что сумма весов равна единице.

Замечание. При ряде условий имеет место сходимость по вероятности данной оценки к $E(y|x)$.

Ядерная оценка плотности распределения непрерывной случайной величины определяется по следующей формуле:

$$f_n(t) = \frac{1}{n\sigma}\sum_{i=1}^{n} V\left(\frac{t-\xi_i}{\sigma}\right) \tag{2.5}$$

На практике используется несколько видов ядерных функций. Чаще всего используется квартическая ядерная функция

$$K(u) = (15/16)(1-u^2)^2 I(|u| \le 1) \tag{2.6}$$

Также используется ядро Епанечникова, обладающее некоторыми свойствами оптимальности. Это функция параболического типа:

$$K(u) = 0.75(1-u^2) I(|u| \le 1) \tag{2.7}$$

Другими примерами являются ядро Гаусса,

$$K(u) = (2\pi)^{-1/2}\exp(-u^2/2) \tag{2.8}$$

треугольное ядро:

$$K(u) = (1-|u|) I(|u| \le 1), \tag{2.9}$$

и прямоугольное ядро:

$$K(u) = (1/2) I(|u| \le 1) \tag{2.10}$$

На рисунке 2 представлены виды описанных ядер.

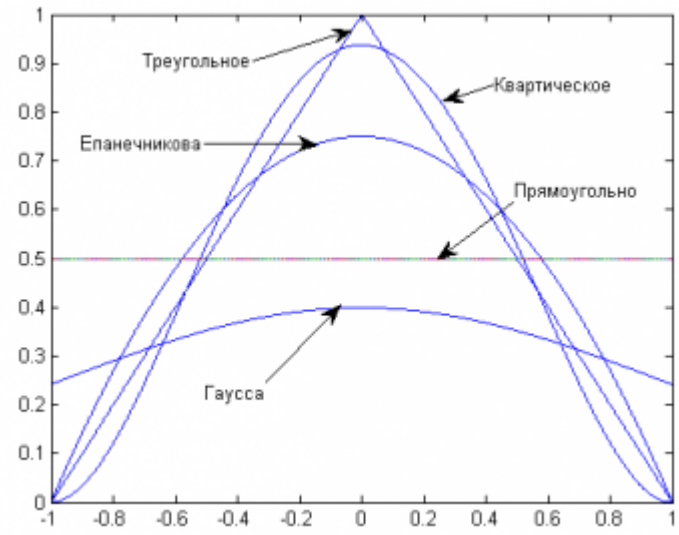

Рисунок 2 – Примеры ядер

Замечание. Точность восстанавливаемой зависимости мало зависит от выбора ядра. Ядро определяет степень гладкости функции $a(x)$.

Выбор окна решающим образом влияет на точность восстанавливаемой зависимости. При чересчур малых значениях h кривая $a(x)$ стремится пройти через каждую точку выборки, остро реагируя на шумы и претерпевая резкие скачки, поскольку в этом случае оценка опирается только на небольшое число наблюдений из узкой окрестности точки x. Наоборот, если ширина окна велика, функция чрезмерно сглаживается и в пределе при $h \to \infty$ вырождается в константу – усреднённое значение величин y_i. В этом случае сглаженная функция не даёт возможности определить характерные особенности искомой зависимости $y^*(x)$ [22].

2.2 Метод Монте –Карло

Для исследования систем в условиях элиторной неопределенности используются различные методы, которые опираются на теорию вероятностей, в том числе подходы, основанные на описании законов распределения случайных величин. В этом случае важно отметить, что вероятностные оценки законов распределения носят детерминированный характер [8].

Одним их самых распространенных и широко применяемых методов для решения задач в условиях неопределенностей является метод Монте-Карло.

В общем случае имитационное моделирование Монте-Карло – это процедура, с помощью которой математическая модель определения какого - либо показателя подвергается ряду имитационных прогонов с помощью компьютера [27]. В ходе процесса имитации строятся последовательные сценарии с использованием исходных данных, которые по смыслу проекта являются неопределенными, и потому в процессе анализа полагаются случайными величинами [29]. Процесс имитации осуществляется таким

образом, чтобы случайный выбор значений из определенных вероятностных распределений не нарушал существования известных или предполагаемых отношений корреляции среди переменных [16].

Лукашов А.В. в своей работе [23] на практических экономических примерах демонстрирует применение метода Монте-Карло, а именно, при анализе привлекательности инвестиционного проекта, вычисление и оптимизация NPV проекта, анализ по методу Монте-Карло с коррелированными параметрами.

Пример. В данном примере рассматривается применение метода Монте Карло для анализа привлекательности весьма простого инвестиционного проекта.

Описание проекта: фармацевтическая компания рассматривает вопрос о приобретении для последующего производства патента нового лекарственного препарата. Лекарство примечательно тем, что не имеет побочных эффектов. Стоимость патента составляет $3,4 млн.

Необходимо подготовить финансовый анализ приобретения данного патента методом дисконтированных денежных потоков, рассчитать NPV и IRR проекта. Горизонт расчетов составляет три года. Стандартная финансовая модель приводится на рисунке 3.

	Год 0	Год 1	Год 2	Год 3
Цена упаковки		$6,00	$6,05	$6,10
Количество проданных, шт.		802000	967000	1132000
Выручка		$4 812 000	$5 850 350	$6 905 200
Себестоимость		$2 646 600	$3 217 693	$3 797 860
Валовая прибыль		$2 165 400	$2 632 658	$3 107 340
Операционные издержки		$324 810	$394 899	$466 101
Чистый доход до налогов		$1 840 590	$2 237 759	$2 641 239
Налоги		$588 989	$716 083	$845 196
Стартовые инвестиции	–$3 400 000			
Чистый доход	–$3 400 000	$1 251 601	$1 521 676	$1 796 043
NPV (3 года)	$344 796			
IRR (3 года)	15%			

Рисунок 3 – Финансовая модель для проекта по покупке патента на изготовление препарата

Согласно прогнозам аналитиков, компания в первый, второй и третий год проекта продаст соответственно 802 тыс., 967 тыс. и 1132 тыс. упаковок лекарства по цене $6, $6,05 и $6,10 за упаковку.

Ставка налога на прибыль равна 32 %, ставка дисконтирования равна 10 %, себестоимость составляет 55 %, а операционные издержки – 15 % от цены препарата. Для вычисления NPV и IRR проекта в Excel использовались функции ЧПС («Чистая приведенная стоимость») и ВСД («Внутренняя ставка

доходности»). По результатам расчетов IRR проекта составляет 15 %, а NPV – $344,8 тыс. Поскольку NPV > 0, то компании следует принять проект.

Несмотря на положительные результаты стандартного анализа все равно в полученных прогнозах нельзя быть полностью уверенными. Рынок лекарственных препаратов является весьма конкурентным. Конкуренция со стороны других препаратов может привести к снижению цены ниже прогнозируемой. Также из-за влияния конкуренции трудно точно предсказать объем продаж препарата (количество упаковок). Помимо цены и объема продаж не поддаются точному прогнозу будущая себестоимость препарата и операционные издержки. Очень часто себестоимость и издержки превышают запланированные. Кроме того, они могут колебаться год от года [15].

В данном случае мы имеем дело с высоким уровнем непрерывной (рыночной) неопределенности, поэтому стандартная финансовая модель по методу DCF не может дать достаточных для принятия решения результатов. Для одновременного учета неопределенности в цене, продажах, себестоимости и издержках применяется анализ по методу Монте Карло.

Основные параметры финансовой модели — цена, объем продаж — моделируются как случайные переменные, имеющие вероятностное распределение. Анализ по методу Монте Карло предоставит необходимую информацию для ведения более обоснованных переговоров о покупке патента на изготовление лекарства, а также позволит понять, какие факторы в наибольшей степени повлияют на финансовые результаты проекта [11].

Для моделирования цены продажи (в первый, второй и третий год проекта отдельно) используется треугольное распределение. Треугольное распределение имеет три параметра — минимальное значение, максимальное значение и наиболее вероятное значение. Его, как правило, используют для моделирования параметров, которые менеджеры в значительной степени могут контролировать. Цена продажи в первый год имеет минимальное значение $5,90, максимальное значение — $6,10 и наиболее вероятное значение — $6,00, представленное на рисунке 4. Аналогично, цена продажи во второй год имеет треугольное распределение с параметрами $5,95; $6,05; $6,15. Цена продажи на третий год имеет треугольное распределение с параметрами $6,00; $6,10; $6,20.

В отличие от цены, которая колеблется, но находится под контролем менеджеров компании, объем продаж зависит от не контролируемых фирмой факторов. Как правило, объем продаж моделируется как случайная переменная с нормальным распределением [18].

Объем продаж в первый год имеет нормальное распределение со средним значением (математическим ожиданием) $802 тысяч и стандартным отклонением $25 тысяч, представленным на рисунке 5. Аналогично, объем продаж во второй год имеет нормальное распределение с ожиданием $967 тысяч и стандартным отклонением $30 тысяч. Наконец, объем продаж в третий год имеет нормальное распределение с ожиданием $1132 тысяч и стандартным отклонением $25 тысяч.

Себестоимость (процент от продаж), как предполагается, имеет треугольное распределение с минимальным значением 50 %, максимальным значением 65 % и наиболее вероятным значением 55 %. Следует отметить, что в данном случае треугольное распределение имеет не симметричную форму, а немного скошено вправо, т. е. имеется большая вероятность того, что себестоимость будет завышена, а не занижена по сравнению с наиболее вероятным значением. Операционные издержки (процент от продаж) моделируются как нормальное распределение с ожиданием 15 % и стандартным отклонением 2 %.

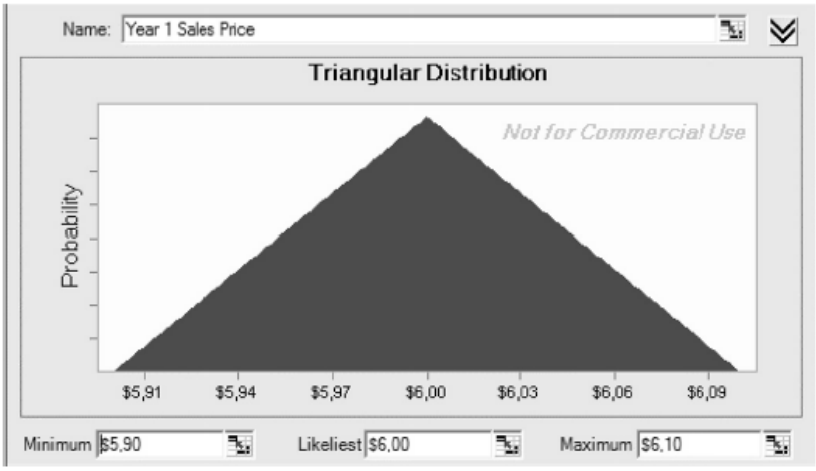

Рисунок 4 – Треугольное распределение для моделирования цены продаж в первый год проекта

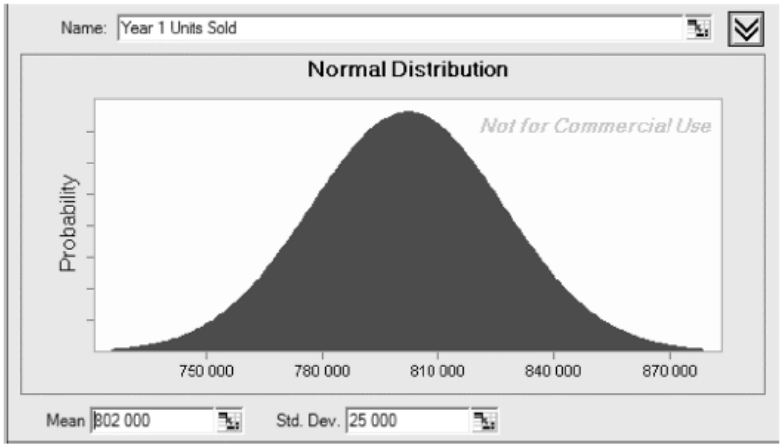

Рисунок 5 – Нормальное распределение для объема продаж в первый год проекта

Всего в ходе анализа по методу Монте Карло было сделано 10 тысяч повторов. При каждом повторе программа генерировала новые значения для случайных переменных (параметров финансовой модели) и вычисляла значение NPV и IRR проекта [31]. Результаты анализа в виде гистограммы показаны на рисунке 6 и рисунке 7, и обобщены на рисунке 8.

Рисунок 6 – Результаты анализа по методу Монте-Карло: гистограмма для внутренней ставки доходности (IRR) проекта

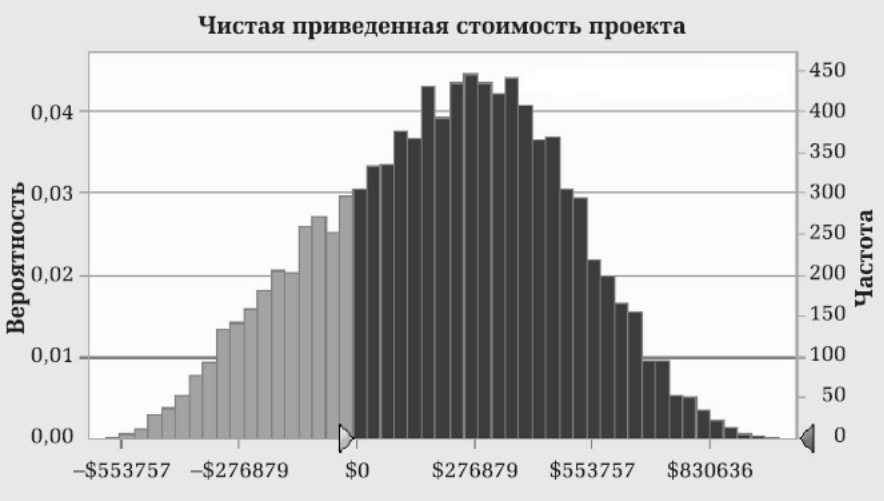

Рисунок 7 – Результаты анализа Монте-Карло: чистая приведенная стоимость (NPV) проекта

Как видно из рисунка 7, средняя NPV проекта составляет 202 тысяч, что значительно меньше, чем NPV стандартной модели (344,8 тысяч) Это результат скошенного вправо распределения себестоимости. Анализ по методу Монте Карло показывает: вероятность того, что NPV проекта будет положительной, не является стопроцентной. Как видно из гистограммы, существует вероятность (почти 25 %) того, что NPV проекта окажется отрицательным. Таким образом в одной четвертой всех случаев при определенной комбинации факторов компания понесет потери. В то же время при благоприятном стечении факторов NPV проекта может превышать $1 млн. [23].

Не смотря на широкие возможности, метод Монте-Карло имеет ряд недостатков, которые приводит Scott Ferson в [2]:

– Монте-Карло требует большого объема данных;

– Монте-Карло не распространяет частичное незнание ни при какой частотной интерпретации;

– Монте-Карло не может оценить превышение риском определенного уровня;

– Монте-Карло не может быть применен для эффекта несворачиваемости при решении задач обратных расчетов. Scott Ferson приводит 10 примеров, отражающих эти недостатки.

Рассмотрим следующие проблемные задачи-примеры:

1) Предположим, нам нужно оценить произведение двух недостаточно известных входов. Мы знаем, что первый, A, не может быть меньше, чем 0,2 и больше, чем 0,4. Мы также знаем, что второй, B, не может быть меньше, чем 0,3 и больше, чем 0,5. Но никакой дополнительной информации о A или B не имеется. Как следует охарактеризовать произведение AB?

2) Предположим, мы знаем, что A не меньше, чем 0,2 и не больше, чем 0,3 и что B логарифмически нормально распространяется с медианной 5 ($\mu = ln$ (медиана) $\approx$ 1,6) и коэффициентом вариации 0,2 ($б=\sqrt{\ln(GV^2+1)} \approx 0{,}2$). Что можно сказать о произведении AB?

3) Предположим, мы знаем, что A логарифмически нормально распространяется с медианой 2 и коэффициентом вариации 0,1 ($\mu \approx 0{,}69$, $б \approx 0{,}1$), и что B логарифмически нормально распространяется с медианой 5 и коэффициентом вариации 0,2 ($\mu \approx 1{,}6$, $б \approx 0{,}2$). Насколько велика частота, с которой произведение AB превышает 14?

4) Учитывая A и B, как указано выше, и корреляция между A и B равно нулю, насколько большой является частота, с которой произведение AB превышает 14?

5) Учитывая A и B, как указано выше, и корреляция между A и B составляет 0,3, насколько большой является частота, с которой произведение AB превышает 14?

6) Учитывая A и B, как указано выше, и корреляции Спирмена (ро) между A и B составляет 0,3, насколько большой является частота, с которой произведение AB превышает 14?

7) Предположим, AB =C, где A, B, и C являются случайные величины, где A и B независимы. Если A имеет логарифмически нормальное распределение, с медианой 2 и коэффициентом вариации 0,1($\mu \approx 0{,}69$, $б \approx 0{,}1$), и C имеет логарифмически нормальное распределение, с медианой 5 и коэффициентом вариации 0,2($\mu \approx 1{,}6$, $б \approx 0{,}2$), каково распределение B?

8) Предположим, что мы хорошо знаем маргинальные распределения для случайных величин A и B, но не можем точно указать их совместное распределение. Как мы можем охарактеризовать произведение AB?

9) Предположим, мы не можем точно указать маргинальные распределения для случайных величин A и B. Как мы можем характеризовать произведение AB?

10) Предположим, что мы знаем полное совместное распределение для случайных величин A и B, но не уверены в точной функции, которая должна

использоваться, чтобы объединить их. Как мы можем сделать обоснованную характеристику функции A и B?

Каждый из этих примеров по форме похож на общие проблемы анализа рисков, которые регулярно обращались к методам Монте-Карло. Тем не менее, ни одна из базовых проблем не может быть решена с помощью прямого применения методов Монте-Карло. Причины, почему методы Монте Карло не могут быть применены (или дают неправильные ответы) рассматриваются ниже для каждой задачи в частности. Большинство проблем имеют решения, однако, они также рассматриваются в следующих разделах.

Важно отметить, прежде всего, что конкретные данные были включены в первые семь проблем, чтобы сделать обсуждение конкретным. Например, некоторые проблемы относятся к логнормальному распределению с заданными медианами и коэффициентами вариации, или относятся к вопросам о произведении случайных величин. Понятно, что аналогичные проблемы могут быть сформулированы в условиях распределений с другими параметрами или другими формами и другими математическими операциями над ними. Там нет уточнений о параметрах или распределениях форм, которые делают проблему особенно сложной. Подробности позволяют рассчитать правильный ответ, который позволит нам сразу понять, как метод Монте-Карло дает сбой в каждом конкретном случае.

Первая из приведенных выше проблем спрашивает о произведении двух параметров, о которых известны только минимумы и максимумы. Около двух третей из пятидесяти пяти респондентов дали решение: AB = [0.2, 0.4] x [0.3, 0.5] = [0.2 x 0.3, 0.4 x 0.5] = [0.06, 0.2], ответ которого получен с помощью интервального анализа. Однако, почти треть респондентов предположили, что в данной задаче должен быть использован вероятностный подход. Большинство решений, которые они предложили явно описывал метод Монте-Карло, который моделировал A и B с равномерным распределением по их соответствующим диапазонам. На рисунке 8 представлена функция плотности вероятности, которая является результатом свертки этих двух равномерных распределений вместе в предположении независимости, которая является допущением относительно зависимости в большинстве программных пакетах, включая CrystalBall и RISK.

Ответ интервального анализа и распределение Монте-Карло сходятся в том смысле, что они оба говорят, что ответ должен лежать где-то в диапазоне от 0,06 до 0,2. Тем не менее, распределение вероятностей говорит немного больше, чем это. Оно утверждает, что вероятность того, что произведение стремится к одному из крайних значений гораздо меньше, чем вероятность того, что оно имеет более центральное значение. Но где в постановке задачи мы можем найти оправдание для этой концентрации вероятности в центре диапазона? В самом деле, конечно, любое распределение вероятности в диапазоне от 0,06 до 0,2 может быть истинным распределением произведения. В постановке задачи ничего не дано, что мы могли бы использовать, чтобы сузить этот набор. Даже распределение дельта в 0,06 или 0,2 не может быть

исключено.

Идея о том, что целесообразно предположить вероятностную однородность только когда диапазон информации доступных значений относится к Лапласу. Идея была известна как «принцип недостаточного основания». Хотя это было оправдано и обобщено сложной теоретической разработкой в разделе максимальной энтропии, идея широко рассматривается с некоторым скептицизмом, особенно теми, кто приводит к частотному виду теорию вероятности [1].

Есть утверждение, что, в некоторых деталях, интервальный анализ обеспечивает разумное решение только первой приведенной в качестве примера проблемы, по крайней мере, с точки зрения анализа рисков. В контексте поиска представление для одного базового числа, подход с использованием одинакового (или максимально энтропийного) распределения может быть разумным. В контексте анализа рисков, однако, подход дает ответы, которые более уверенны. Это более доверительно генерирует результаты, которые, по крайней мере, потенциально не защищены. Например, это может быть случай, когда большие (или меньшие) значения будут происходить гораздо чаще, чем это следует из полученного определенного распределения. Подход Монте-Карло и любой классический вероятностный подход не могут всесторонне распространяться на нестатистическую неопределенность, по крайней мере, при частотной интерпретации требуемого анализа рисков.

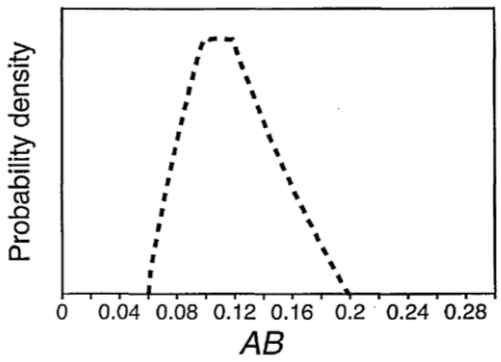

Рисунок 8 – Оценка MonteКарло для распределения произведения в первом примере проблемы.

Незначительное меньшинство респондентов предположили, что решение проблемы должно быть равномерно распределено в интервале [0.06, 0.2]. Конечно, это решение также завышает то, что обоснованно выводится о произведении AB, но в том смысле, что равномерное распределение представляет собой интервал, это, возможно, более разумно, чем треугольная форма, показанная на рисунке. Тем не менее, с точки зрения анализа риска, оно по-прежнему потенциально недооценивает хвосты распределения таким образом, что не может быть обоснованно обращение к каким-либо эмпирическим фактам или предположениям, изложенным в

задаче.

Некоторые могут критиковать первый пример проблемы, как нереальную, и утверждать, что, в большинстве реальных случаях, мы, на самом деле, имеем больше информации о параметре, чем просто его минимум и максимум. Конечно, это верно, что такая дополнительная информация часто доступна. Тем не менее, эта проблема, конечно, не исключена в реальных условиях и, в силу своей кажущейся простоты, заслуживают убедительное и обоснованное решение, даже если такая проблема редко встречается. Также, можно утверждать, что всегда остается некоторая степень незнания о параметре, который не является результатом изменчивости практически во всех реальных проблемах. Поскольку эта проблема дает представление об этом вопросе, она также заслуживает обсуждения. Более общий случай, когда другие сведения о распределении доступны, но само по себе распределение задано неточно, рассматривается ниже, в обсуждении девятого примера проблемы [2].

Второй пример проблемы стремится объединить интервал и распределение вероятностей. Опять же, использование равномерного распределения в качестве замены интервала позволит использовать для получения решения методы Монте-Карло. Но сделав так, это даст результат, который не может быть обоснован, учитывая указанное предположение. Рисунок 9 отображает оба результата: анализ Монте-Карло и пределы истинного решения, полученного путем прямого анализа границы вероятности. Учитывая информацию, изложенную в задаче, эти границы точечно оптимально узки. Другими словами, они не могут быть более узкими и все же заключить все вероятностные распределения, которые могут, на самом деле, возникать как произведение A и B. На рисунке мы можем увидеть, что результат Монте-Карло предполагает частоту, что произведение меньше 1.0 составляет около 15 %. В самом деле, эта частота может быть как 50 %, так и больше, в зависимости от распределения или значения, которое A принимает на самом деле в заданном интервале. Можно отметить, что подход Монте-Карло не может всесторонне распространяться на нестатистическую неопределенность [2].

Границы произведения интервала и распределения вероятностей могут быть также получены с помощью альтернативной стратегии с использованием набора интервалов для представления распределения. Если участки взяты систематически из логнормального распределения для B и умноженные на интервал [0.2, 0.3] в соответствии с элементарными правилами интервального анализа, произведения все будут интервалами. Суммирование левых границ этих произведений даст левую границу области, изображенной на рисунке, суммирование правых границ произведений будет приближаться к правой границе.

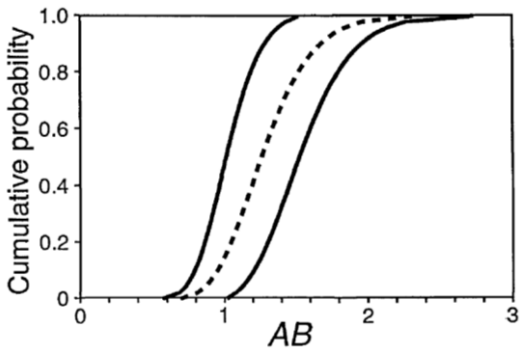

Рисунок 9 – Границы (сплошные линии) суммарной функции распределения произведения во втором примере проблемы и оценка Монте-Карло (пунктирная линия), при предположении равномерного распределения для A.

Превышение рисков - это частота, с которой случайная величина может превышать некоторое заданное значение. Превышение рисков есть фокальная проблема для аналитиков, делающих оценки рисков для окружающей среды или общественного здоровья, такие как моделирование облучения или заболеваемости раком, иногда, так как существует критический уровень, превысив который, последствия становятся недопустимыми. Одной из центральных задач в области анализа рисков является оценка, как часто могут возникать такие последствия. С третьего по шестой примеры проблем сформулированы вопросы о вычислении превышения рисков.

Пунктирная линия на рисунке 10 представляет результат моделирования по методу Монте-Карло третьего примера проблемы, в которой логнормально распределенные случайные отклонения выборки независимо от распределений A и B перемножаются. Результат показан в качестве кумулятивной функции распределения (которая также известна как функция выживания). От этой кривой мы можем считывать оценочную частоту произведения AB, являющегося большим, чем 14, как 6,5 %. Но эта оценка зависит от предположения о независимости A и B. Задача опускает какое-либо упоминание о зависимости между A и B. Однако их зависимость может оказывать существенное воздействие на результат распределения. Например, если зависимость очень высокая, то превышение риска вполне может быть в два раза больше. Если зависимость между A и B очень низкая, превышение риска может быть практически нулевым.

Уильямсон и Даунс описывают численный метод для вычисления возможного диапазона результатов при заданных маргинальных распределениях, когда их зависимость неизвестна. Их схема представления использует верхние и нижние дискретные приближения к квантильной функции (квазиобратной функции распределения) в качестве границ распределения. Метод основан, главным образом, на классических Фреше неравенствах

$$max(0, Pr(E) + Pr(F) - 1) \leq Pr(E \; and \; F) \leq min(Pr(E), Pr(F)) \quad (2.11)$$

$$max(Pr(E), Pr(E) \leq Pr(E \text{ or } F) \leq min(1, Pr(E) + Pr(F)) \qquad (2.12)$$

которые дают наилучшие оценки вероятностей конъюнкций или дизъюнкций событий, когда дана только общая вероятность событий *Pr(E)* и *Pr(F)* и нет никакой дополнительной информации о зависимости между событиями *E* и *F*. Хотя, переход от событий к случайным величинам требует вызов связок, которые отражают отношения зависимости, заключенные в совместных распределений, алгоритмы, разработанные Уильямсоном и Даунсом достаточно просты в использовании при анализе рисков. Применяя их метод это дает результат, обозначенный сплошными линиями, изображенными на рисунке.

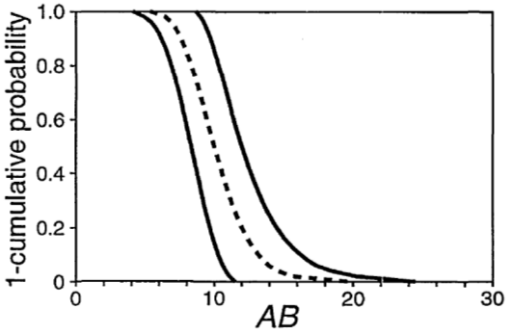

Рисунок 10 – Границы (сплошные линии) комплементарной интегральной функции распределения и оценка Монте-Карло (пунктирная линия), предполагающая независимость произведения двух логнормальных случайных величин A и B, определенных параметрами в третьем примере задачи.

Сплошные линии описывают границы истинного превышения риска, заданного только информацией о маргинальных распределениях A и B. Из рисунка мы видим, что риск того, что произведение AB больше, чем 14, может быть где угодно в пределах от 0 до 25 %. Эти оценки, как известно, являются оптимальным и в том смысле, что они не могут быть уже, без более конкретной информации о зависимости между A и B. В таком случае сложно интерпретировать оценку Монте-Карло 6,5 %.

В некоторых случаях, можно исследовать спектр возможного превышения рисков путем изменения коэффициента корреляции по всем возможным значениям между -1,0 и 1,0. Несколько исследователей предложили эту стратегию, которую можно назвать «дисперсионные выборки Монте-Карло», потому что ее отношение к дисперсии максимизировано. Однако, эта стратегия не работает вообще. Даже изменение коэффициента корреляции по всем возможным значениям не может дать весь спектр возможных значений превышения риска. Это потому, что корреляция очень ограниченный вид линейной зависимости, и такая стратегия исследует лишь малую часть пространства возможных зависимостей между двумя переменными [1].

Не зная полное совместное распределение двух переменных, которое выражает их зависимость, классические методы теории вероятностей, в том числе методы Монте-Карло, не могут вычислить оценку распределения произведения, даже в принципе. К сожалению, этот факт не остановил многих аналитиков, которые привычно полагают независимость среди всех переменных в выражении риска. Такие предположения иногда имеют место быть, хотя часто это не так.

Четвертый пример проблемы предполагает, что корреляция между двумя факторами равна нулю. Но, нулевая корреляция не есть то же самое, что независимость. Хорошо известно, например, что две нормальных случайных величины могут иметь сумму, которая даже близко не является нормальной, хотя их корреляция равна нулю. Точно так же, как произведение двух логнормальных случайных величин не должно давать логнормальное распределение, даже если они имеют нулевую корреляцию. Есть бесконечное множество способов достижения нулевой корреляции, но есть только один способ быть независимыми.

Аналогичные трудности сохраняются в пятом и шестом примерах. Наличие определенной (ненулевой) корреляции не подразумевает конкретное совместное распределение. То же самое верно для ранговой корреляции. Даже при том, что программные пакеты Монте-Карло могут имитировать такие корреляции, они не могут описать такие корреляции, в смысле обхвата разнообразия распределений, которые могут получиться, когда переменные арифметически объединены. Это означает, что даже если измерить корреляцию и включить ее в моделирование Монте-Карло, невозможно быть уверенным в том, что оценка вероятностей превышения не полностью отражает, возможно, существенное занижение. Иначе говоря, риски неблагоприятных последствий могут быть определенно выше, чем можно было бы предсказать по Монте-Карло, даже при необычайно хороших обстоятельствах, хорошо зная оба предельные распределения и структуру корреляции, даже с бесконечным числом повторений. Эти факты, кажется, не получили широкого признания в сообществе анализа рисков. Сколько высших истинных частот могут быть все еще оставаться открытым вопросом для исследования, разрешение которого заслуживает серьезного внимания.

Седьмой пример связан с обратными вычислениями. Если, как говорится в задаче, $A * B = C$, и необходимо оценить B из A и C, аналитик может попытаться вычислить B, перестроив уравнение, чтобы получить $B = C/A$ и оценить ответ по Монте Карло. Результат такого подхода показан на рисунке 11 в виде пунктирной линии. Этот ответ является неправильным, это можно легко проверить, вернув уравнение в исходный вид и вычислить C. Правильный ответ, то есть, распределение, которое позволит получить наблюдаемое распределение для C, изображено на рисунке в виде сплошной линии.

Причина расхождения в том, что была попытка вычислить B, предполагая независимость между A и C. Но, конечно, A и C не могут быть

независимы друг от друга. В самом деле, это представляет собой другую функцию, поэтому предположение о независимости явно неправильное.

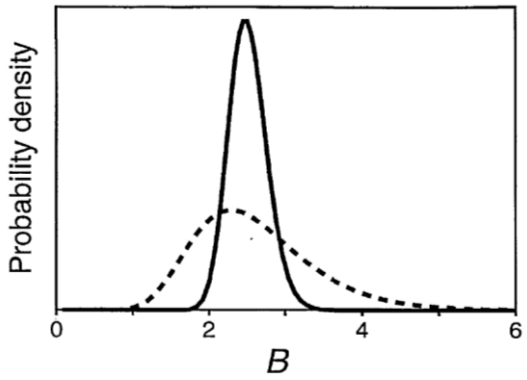

Рисунок 11 – Несоответствие истинного распределения (сплошная линия) и оценки, вычисленной с помощью Монте-Карло (пунктирная линия) в седьмом примере задачи.

К этому моменту, должно быть понятно, что методы Монте-Карло требуют большое количество информации, чтобы получить ответы. Есть три следствия этого общего наблюдения, воплощенные в последних трех примерах задач. Они могут быть выражены как:

– методы Монте-Карло не могут дать ответ, когда статистические зависимости между переменными неизвестны или неопределенны;

– методы Монте-Карло не могут дать ответ, когда входные распределения неизвестны или неопределенны;

– методы Монте-Карло не могут дать ответ, когда модель структуры неизвестна или неопределенна.

Восьмой пример проблемы, в которой совместное распределение величин неизвестно или неопределенно, очень часто возникает в реальных рисках и оценке безопасности. Действительно, случай, для которого были собраны достаточные данные для хорошей оценки совместного распределения вероятностей очень необычен. Существуют методы для вычисления оценки распределений арифметических комбинаций с использованием только маргинальных распределений, не делая никаких предположений о зависимости между переменными. Во многих случаях эти границы будут оптимально узкими, а переменные, которые, как известно, не зависят, так же могут быть вовлечены в выражении риска. Как упоминалось выше, однако, легкость, в которой частичная информация о зависимости между переменными известна, не решена. В настоящее время там нет способа вычисления оптимального предела, когда некоторая, но не вся, указанная информация говорит о зависимости, или, когда корреляция принимает определенное значение или лежит в пределах некоторого интервала.

Девятый пример проблемы обобщает первую проблему в том, что входные распределения неизвестны или неопределенны. В первой задаче было доступно очень мало информации о входных распределениях. Более общая

ситуация заключается в том, что доступна некоторая информация, и, возможно значительная, которая характеризует входные распределения, но этого недостаточно, чтобы точно определить распределения. Поскольку стандартная вероятностная оценка, включая Монте Карло, требует выбор, в частности, четко определенных вероятностных распределений в качестве входных, даже если имеющиеся эмпирические данные не оправдывает такой специфичности, для решения этой проблемы различные стратегии были предложены.

Один из подходов к проблеме заключается в использовании критерия максимальной энтропии для определения распределения, используя при этом любой имеющийся объем информации о переменной. Стратегия заключается в определении распределения, имеющего наибольшую энтропию, которая согласуется с имеющимися знаниями, которые образует ограничения на возможную фигуру. Таким образом, этот подход, еще не дав никаких предположений о фигуре, позволяет выбирать входное распределение оптимальным образом, используя лишь ограниченную информацию о переменными. Например, критерий выбирает равномерное распределение, когда известны только верхние и нижние границы значений, экспоненциальное распределение, когда известны только нижняя граница и среднее, и нормальное распределение, когда известно только среднее значение и стандартное отклонение.

Десятый пример проблемы, в которой сама модель неизвестна или неопределена, представляет собой мощную, повсеместную трудность в анализе рисков. Примеры оценки экологического риска включают анализ риска исчезновения находящихся под угрозой исчезновения видов, у которых лучшая модель зависимости плотности, касающейся изменения в демографических показателей численности популяции, неизвестна. В обоих случаях, есть несколько функций с различными видами нелинейности, которые традиционно используются, но без тщательного и конкретного изучения базовой биологии, аналитик не может быть уверен, в том, какую модель следует использовать. Как и следовало ожидать, выбор может иметь большое значение. Каллен исследовал последствия использования различных моделей в оценке рисков на конечный результат. Mosleh и Bier рассмотрели несколько более абстрактную задачу выбора уровня агрегации, который использовать для создания модели в анализе риска.

Существует очень мало соглашений о том, как должны быть обработаны неопределенности относительно вида модели. Некоторые утверждают, что неопределенность должна быть усреднена в процессе подобно тому, как рассматривается параметрическая неопределенность. Например, Голланд и Sielken утверждают, что усреднение результатов различных моделей весьма разумно, взвешенные по соответствующим данным они все имеют свои требования в качестве истины. Иногда это взвешивание вырождается подсчетом сторонников моделей, или считая количество работ,

опубликованных по обе стороны дебатов, популярность становится мерой истины.

Другие не согласны с таким подходом, полагая, что это бессмысленно усреднять результаты взаимоисключающих теорий. Оба метода: вероятностный анализ границ и моделирование второго порядка Монте-Карло имеют потенциал, чтобы комплексно обрабатывать модели неопределенности если есть конечный ряд моделей на выбор, хотя последний возрастает в комбинаторной сложности, поскольку количество вариантов примерных моделей форм увеличивается.

Как и любой аналитический инструмент, метод Монте-Карло может привести к неверным или необоснованным результатам, всякий раз, когда его предположения ложны или не оправданы эмпирически. Различные проблемы, примеры которых приведены выше, являются принципиальным результатом двух основных видов ошибок:

– использование точных распределений без эмпирического обоснования;

– ненадлежащее моделирование зависимостей между переменными.

Хотя очевидно, что неправильное использование точных распределений и независимости предположений это вина аналитика, а не анализа, также справедливо сказать, что анализ Монте-Карло, и сама теория вероятностей в общем, не будет иметь большого практического использования без таких предположений. Эти вопросы уже давно признаны. Как отметил Джейнес, что "эта проблема спецификации вероятностей в случаях, когда имеющейся информации мало или она вообще отсутствует, так же стара, как теория вероятности". Уиттакер объяснил, "Независимость цементируется в самом фундаменте теории вероятностей: тема, которая повторяется в дальних рубежах исследований, и она пронизывает все приложения вероятности в научном исследовании".

Расхождения между результатами Монте-Карло и более комплексными решениями примерных проблем не следует сбрасывать со счетов, как тривиальные или неважные. В некоторых случаях результаты Монте-Карло могут быть приближенно верны; в других случаях, они могут быть решительно далеко от истины. Но в любом случае, это важно для анализа риска, по крайней мере знать формально правильное решение проблемы, даже если она используется не каждый раз. Кроме того, возникает вопрос о хорошей профессиональной практике. Как правило, аналитики не должны делать необоснованных предположений просто ради удобства вычислений. Оценка риска не должна быть обратной стороной расчетов. Все, что известно эмпирически, должно быть указано явно и, в максимально возможной степени, дальнейшие предположения не должны делаться только ради облегчения расчетов. Всякий раз, когда дальнейшие предположения необходимо применить, если иначе проблема становится неразрешимой, они должны обсуждаться непосредственно в документации об оценке, где следует подчеркнуть их обоснование.

Несмотря на привлечение возможности использовать теорию вероятности в гораздо более широких контекстах, кажется, что есть некоторые проблемы, которые требуют исключительно частотной интерпретации. Например, в применении анализа риска, особенно для задач в области общественного здравоохранения и окружающей среды, это не кажется разумным, свернуть субъективные ощущения аналитика в анализ, независимо от того, насколько хорошо может быть обучен и благородно мотивирован аналитик. Знания и неопределенность одного аналитика может, и должно, отличаться от другого. Какова актуальность для анализа рисков может быть во мнении аналитика, в отличие от каких-либо доказательств, которые он собрал. Манипуляции таким мнением может привести лишь к формальным расчетам подозрений, а не истинных частот побочных эффектов.

Если теория вероятностей не математическая наука частот, то, возможно, анализу риска следует уделить эту обязанность. Не нужно утверждать, что все формы неопределенности могут быть выражены с точки зрения точно известных частот. Тем не менее, пессимизм Моргана и Генриона по этому вопросу кажется взвинченным. Просто потому, что количественные аспекты вероятностных событий неизвестны, не означает, что мы должны использовать вместо них субъективные оценки. Мы можем объективно связать наше незнание, используя обычные методы науки. Понятно, что подход основан, главным образом, на частотах, но обобщенное признание ограниченного незнания о них может служить математическим подкреплением анализа рисков с широким спектром применения. Это может представлять практический интерес в использовании мнения во вторичном анализе «что-если», и это использование вполне разумно. Важно, однако, тщательно различать случаи, когда субъективные оценки используются теми, кто претендует на обобщение состояния эмпирического знания.

2.2 Численный вероятностный анализ

2.2.1 Определение и основные задачи численного вероятностного анализа

Численный вероятностный анализ (ЧВА) представляет собой новый раздел вычислительной математики, который предназначен для решения разнообразных задач с неопределенными входными данными. Новизна ЧВА и основное его отличие от рассмотренных выше направлений исследования неопределенностей, в том числе интервального анализа, анализа вероятностных границ, анализа чувствительности, анализа на основе нечетких множеств, теории свидетельств, состоит главным образом в том, что в рамках ЧВА на основе использования нового понятия вероятностное расширение функций удалость разработать общий подход для выполнения разнообразных численных операций над случайными величинами, который в отличие, например, от интервального анализа и анализа на основе вероятностных границ, позволяет определять не только границы области

возможных значений исследуемых характеристик информационного потока, но получить их вероятностное описание внутри этого множества, что является крайне важным для процедур извлечения знаний из имеющихся данных. Построенные на основе ЧВА численные операции над данными в условиях неопределенности могут быть использованы для различных типов неопределенности [14]. В отличие от существующих подходов, ЧВА оперирует с плотностями случайных величин, представленных гистограммами, дискретными и кусочно-полиномиальными функциями [11, 13].

2.2.2 Гистограммы, полиграммы, полигоны

Гистограммой называется случайная величина, плотность распределения которой представлена кусочно-постоянной функцией.

Интервальная гистограмма. Зачастую в прикладных задачах нет возможности получить точную функцию распределения случайной величины. В таких случаях задаются оценки плотности распределения сверху и снизу. Такие оценки удобно аппроксимировать интервальными гистограммами. Гистограмму будем называть интервальной, если ее функция распределения P(x) - кусочно-интервальная функция [6].

Гистограмма второго порядка - в случае эпистемической неопределенности, наряду с интервальными гистограммами, возможно использование гистограмм второго порядка, т.е. таких гистограмм, каждый столбец которой – гистограмма [26].

Обозначим через R - множество $\{x\}$ случайных величин, заданными своими плотностями вероятностями $p_{x,}$ соответственно R^n - пространство случайных векторов.

Наряду с общими представлениями случайных величин своими плотностями в виде непрерывных функций, будем рассматривать случайные величины, плотность распределения которых представляет гистограмму. Гистограмма P - кусочно-постоянная функция определяется сеткой $\{x_i | i = 0, \dots, n\}$, на отрезке $\{x_{i-1}, x_i\}$ гистограмма принимает постоянное значение p_i.

Рассмотрим вопрос о построении гистограммы P по некоторой p_x. Тогда значения p_i на отрезке $\{x_{i-1}, x_i\}$ определится как среднее

$$p_i = \int_{x_{i-1}}^{x_i} p_x(\xi) d\xi / (x_i - x_{i-1}) \tag{2.13}$$

В работе [28] рассмотрена непараметрическая оценка плотности, основанная на утверждении

$$P(|F(x_{(i+k)}) - F(x_{(i)}) - \frac{k}{N+1}| > \varepsilon) \to 0, N \to \infty \tag{2.14}$$

где $F(x)$ — истинная (и неизвестная) непрерывная функция распределения, $x_{(s)}s$ – порядковая статистика выборки $x_1, x_2,..., x_n$.

По определению

$$f(x) = \lim_{h \to 0} \frac{F(x+h) - F(x)}{h} \tag{2.15}$$

следовательно, можно воспользоваться в качестве оценки неизвестной плотности в интервале $(x_{(i)}, x_{(i+k)}]$ величиной

$$\frac{k}{(N+1)(x_{(i+k)} - x_{(i)})} \tag{2.16}$$

Для удобства обозначим порядковые статистики, ранги которых кратны K, через $\xi_j = x_{(jk)}$.

Введя обозначение $m = \frac{K}{N}$ и введя функцию-индикатор интервала с помощью ступенчатой функции $c(t) = \{1 : t \geq 0; 0 : t < 0\}$, результирующую оценку плотности можно записать как

$$f_N(x) = \frac{1}{m} \sum_{j=1}^{m} \frac{c(x - \xi_j) - c(x - \xi_{j+1})}{\xi_{j+1} - \xi_j} \tag{2.17}$$

Будем называть оценку (2.17) полиграммой K-го порядка. Построение полиграммы сводится к упорядочению выборки и построению прямоугольников площади $\frac{K}{N+1}$.

Полиграмма K-го порядка является асимптотически несмещенной оценкой $f(x)$ и имеет конечные моменты лишь при $r \leq K$ [28,32].

Полиграмма обладает особенным преимуществом, заключающимся в том, что ее можно построить по малым выборкам [3], вплоть до 5-7 значений, при этом каждый столбец полиграммы может отображать лишь одно значение случайной величины. На рисунке 12 и рисунке 13 приведены примеры полиграмм для случайной величины объемом выборки 5 и 7 соответственно.

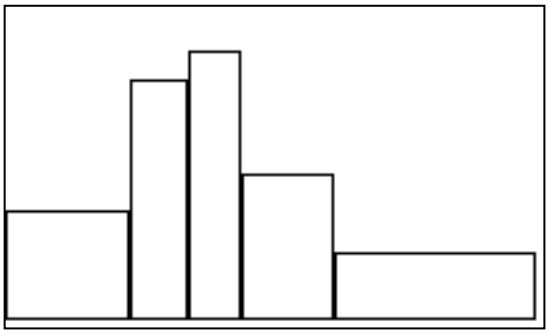

Рисунок 12 – Полиграмма (n =5)

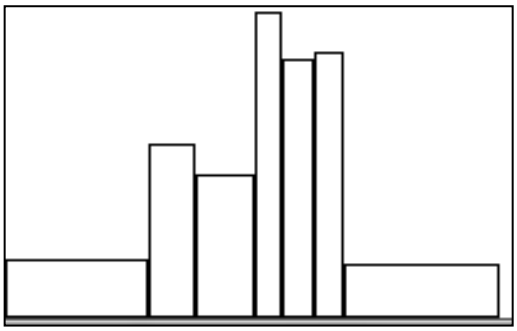

Рисунок 13 – Полиграмма (n =7)

Частотный полигон (ЧП) является непрерывной оценкой плотности на основе гистограммы, стой или иной формой линейной интерполяции.

Скотт исследовал теоретические свойства одномерных и двумерных частотных полигонов и обнаружил, что у них есть удивительные улучшения по сравнению с гистограммами. Фишер не одобряет частотный полигон, по иронии судьбы, из-за причин графического отображения:

Преимуществом является наглядность, что не только форма кривой, указанной подобным образом несколько вводит в заблуждение, но с особой тщательностью всегда следует отличать бесконечно большое гипотетическое множество, из которого отображается наша выборка наблюдений, из фактических результатов наблюдений, которыми мы обладаем; концепция непрерывной кривой частоты применима только к прежнему, и в иллюстрировании последнего не пытайтесь затушевать это различие.

Фишер не знал о каких-либо теоретических различиях между гистограммами и частотными полигонами и думал только об одномерных гистограммах, когда писал этот раздел. Его возражение использовать непрерывную непараметрическую оценку плотности больше не обосновано, но его заботу об использовании методов, которые полностью затеняют статистический шум с математической сложностью стоит подчеркнуть. Наконец, в качестве вопроса терминологии, различие между гистограммой и частотным полигоном в научной литературе стирается, гистограммная метка применяется для обеих случаев.

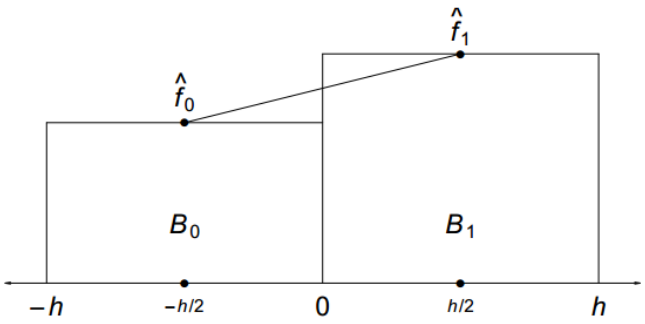

Рисунок 14 – Частотный многоугольник в области ($-$ h/2, h/2), которая является производной от двух соседних областей гистограммы.

Одномерный частотный полигон является линейным интерполянтом середин равных промежутков областей гистограммы. Как таковой, частотный полигон выходит за пределы гистограммы в пустую область с каждого края. Частотный полигон легко проверить с помощью истиной функции плотности, то есть неотрицательная с интегралом, равным 1.

Асимптотическая MISE легко вычисляется на основе подхода область-к-области, рассматривая типичную пару областей гистограммы, отображенных на рисунке 14. Частотный полигон соединяет два смежных значения гистограммы $\widehat{f_0}$ и $\widehat{f_1}$, между центрами областей, как показано на рисунке. Частотный полигон описывается уравнением

$$\hat{f}(x) = \left(\frac{1}{2} - \frac{x}{h}\right)\hat{f}_0 + \left(\frac{1}{2} + \frac{x}{h}\right)\hat{f}_1, \qquad -\frac{h}{2} \le x < \frac{h}{2}. \tag{2.18}$$

Случайность в частотном полигоне приходит полностью из случайности в уровнях гистограммы, $\widehat{f_i} = \nu_i/(nh)$. "$x$" в $\hat{f}(x)$ не является случайной, но она фиксирована.

Как и прежде, используя ряд Тейлора

$$f(x) = f(0) + xf'(0) + \frac{1}{2}x^2 f''(0) + \cdots, \tag{2.19}$$

приближения для p_0 и p_1может быть получено:

$$p_0 = \int_{-h}^{0} f(s)\, ds \approx hf(0) - h^2 f'(0)/2 + h^3 f''(0)/6$$
$$p_1 = \int_{0}^{h} f(s)\, ds \approx hf(0) + h^2 f'(0)/2 + h^3 f''(0)/6. \tag{2.20}$$

Смещения вычисляются, отмечая, что точечное математическое ожидание частотного полигона является линейной комбинацией математических ожиданий двух значений гистограммы. Как $E\{\widehat{f_i}\} = p_i/h$, то из (2.18) и (2.20), и снова отметим, что "x" не является случайным, мы имеем

$$\mathrm{E}\hat{f}(x) = \left(\frac{1}{2} - \frac{x}{h}\right)\frac{p_0}{h} + \left(\frac{1}{2} + \frac{x}{h}\right)\frac{p_1}{h} \approx f(0) + xf'(0) + h^2 f''(0)/6. \tag{2.21}$$

Вычитание (2.19) дает смещение $\hat{f}(x) \approx (h^2 - 3x^2)f''(0)/6$. Интеграл от квадрата смещения (ИКС) по области частотного полигона $(-h/2,\ h/2)$ равен $[49h^4 f''(0)^2/2{,}880] \times h$, с аналогичным выражением для других областей частотного полигона. Суммируя по всем областям и используя стандартное Риманово приближение, получаем

$$\text{ISB} \approx \sum_k \frac{49}{2{,}880} h^4 f''(kh) \times h = \frac{49}{2{,}880} h^4 R(f'') + O(h^6). \tag{2.22}$$

Очевидно, квадрат смещения частотного полигона имеет значительно более низкий порядок, чем порядок $O(h^2)$ гистограммы. Для целей перекрестной проверки, смещение отображается неизвестной шероховатостью $R(f'')$, а не $R(f')$. Частотный полигон расширяет хорошее свойство гистограммы на своих центрах областей, устранение $O(h)$ эффектов, для всей оценки. Смещения являются функцией кривизны в функции плотности, а не наклоном с гистограммой.

Вычисление отклонений аналогично. Из определения частотного полигона в (2.18), дисперсия $\hat{f}(x)$ равна

$$\left(\frac{1}{2} - \frac{x}{h}\right)^2 \text{Var}\hat{f}_0 + \left(\frac{1}{2} + \frac{x}{h}\right)^2 \text{Var}\hat{f}_1 + 2\left(\frac{1}{4} - \frac{x^2}{h^2}\right) \text{Cov}(\hat{f}_0, \hat{f}_1). \tag{2.23}$$

Для дисперсии и ковариации точки, только самое тривиальное приближение $hf(0)$ требуется для p_0 и p_1. Поскольку подсчет областей является биномиальной случайной величиной,

$$\text{Var}(\hat{f}_i) = \frac{np_i(1-p_i)}{(nh)^2} \approx \frac{f(0)(1-hf(0))}{nh} \tag{2.24}$$

и

$$\text{Cov}\hat{f}_0, \hat{f}_1) = \frac{-np_0p_1}{(nh)^2} \approx -\frac{f(0)^2}{n}. \tag{2.25}$$

Подстановка этих аппроксимаций в (2.23) дает

$$\text{Var}\hat{f}(x) = \left(\frac{2x^2}{nh^3} + \frac{1}{2nh}\right) f(0) - \frac{f(0)^2}{n} + o(n^{-1}). \tag{2.26}$$

Интегрирование по области ЧП $(-h/2, h/2)$ получаем $[2f(0)/(3nh) - f(0)^2/n] \times h$. Суммирование соответствующего выражения для всех областей и учитывая, что $\int f = 1$, дает

$$\text{IV} \approx \sum_k \left[\frac{2f(kh)}{3nh} - \frac{f(kh)^2}{n}\right] \times h = \frac{2}{3nh} - \frac{1}{n} R(f) + o(n^{-1}). \tag{2.27}$$

Если оптимальная ширина области гистограммы была использована в частотном полигоне, асимптотический эффект будет устранять смещение всецело относительно дисперсии в MISE, порядки будут $O(n^{-4/3})$ и $O(n^{-2/3})$

соответственно. Поскольку ИКС включает треть Mise для гистограммы, сокращение будет существенным. Но частотный полигон может быть построен лучше. Улучшенный порядок в смещении предполагает, что может быть использована большая ширина области для уменьшения дисперсии, но все еще с меньшим смещением, чем гистограммы. На самом деле, ширина области $h = O(n^{-1/5})$ оказывается в самый раз. Улучшение является существенным, как показывает следующая теорема.

Теорема: Пусть f'' абсолютно непрерывна и $R(f''') < \infty$. Затем

$$\text{AMISE}(h) = \frac{2}{3nh} + \frac{49}{2{,}880}h^4 R(f''); \tag{2.28}$$

следовательно,

$$h^* = 2[15/(49R(f''))]^{1/5} n^{-1/5} \tag{2.29}$$

$$\text{AMISE}^* = (5/12)[49R(f'')/15]^{1/5} n^{-4/5}. \tag{2.30}$$

Например, при 800 нормальных данных точках, оптимальная ширина области для ЧП – на 50 % шире, чем соответствующая ширина области гистограммы. По-видимому, для того, чтобы прерывистая гистограмма приближала непрерывную плотность, гистограмма должна быть довольно грубо отслеживать функцию плотности в тех областях, где ее уровень быстро меняется. ЧП изначально непрерывный и может приблизить непрерывную плотность лучше с кусочно-линейной подстановкой на широких областях. ЧП наиболее плохо работает вблизи пиков, где вторая производная и плотность являются крупными по величине. Улучшение Mise отражается не только понижением константы в передней части $n^{\frac{-2}{3}}$, но также по реальному уменьшению показателя.

Есть ситуация, при которой ЧП находится в невыгодном положении, когда основная плотность скачкообразная. Гистограмма не зависит от таких точек, если они известны и размещены на границах области. ЧП не может избежать дублирования таких точек, и асимптотическая теория, изложенная выше, не применяется.

Чтобы понять практические последствия и чтобы увидеть, где ЧП подходит среди параметрических оценок и гистограмм по отношению к размеру выборки, рассмотрим таблицу 1. Очевидно, что частотный полигон не простое теоретическое любопытство. ЧП даже более эффективен по отношению к гистограмме, когда размер выборки растет. Конечно, обе непараметрические оценки будут все больше и больше уступать правильной параметрической подгонке.

Еще один способ увидеть разницу между гистограммой и ЧП для нормальных данных показан на рисунке 15. На двойной логарифмической шкале, не только различные скорости сходимости легко увидеть, но и

различия в оптимальной ширине области. Продолжая до миллиона нормальных точек, оптимальная ширина области для гистограммы и ЧП является 0,035 и 0,136, соответственно. Они находятся в соотношении 4:1, помечено $h = 4h^*$. Стабильность (малая дисперсия) гистограммы с $h = 4h^*$ очевидна; Однако, в результате то же смещение форме лестницы. ЧП сохраняет стабильность этой гистограммы, в то время как линейная интерполяция резко снижает уклон. ISE из ЧП равен примерно $5{,}40 \times 10^{-6}$, который составляет 14 % от ISE лучшей гистограммы.

Для того, чтобы подчеркнуть различия с соответствующими результатами гистограммы, будут представлены некоторые правила ширины области для ЧП. Правило плагина, основанного на (2.29).

ЧП правило нормальной ссылки: $\hat{h} = 2{,}15\hat{\sigma}n^{-1/5}$,

где $\hat{\sigma}$ представляет собой оценку, может быть устойчивую, стандартного отклонения. Она надежная

Таблица 1 – Пример для $N(0{,}1)$ данных, так что AMISE * ≈ 1/400 и 1/4,000

Estimator	Equivalent Sample Sizes	
$N(\bar{x}, 1)$	57	571
$N(\bar{x}, s^2)$	100	1,000
Optimal FP	546	9,866
Optimal histogram	2,297	72,634

выбор подбирается на основании межквартильного диапазона являющегося $\hat{\sigma} = IQR/1{,}348$, где 1,348 является $\Phi^{-1}(0{,}75) - \Phi^{-1}(0{,}25)$. Факторы, изменяющие правила (4.8), основанные на выборке асимметрии и эксцесса были показаны на рисунке 3.4. Факторы, которые основаны на отношениях

$$\frac{h_y^*}{h_N} = \left[\frac{R\left(\phi''; 0, \sigma_y^2\right)}{R(g''(y))}\right]^{1/5} \tag{2.31}$$

Теперь $R(\phi'') = 3/(8\sqrt{\pi}o_y^5)$ и шероховатость $R(g'')$ логнормальны и t_ν плотности соответственно.

$$\frac{(9\sigma^4 + 20\sigma^2 + 12)e^{25\sigma^2/4}}{32\sqrt{\pi}\sigma^5} \quad \text{and} \quad \frac{12\Gamma\left(\nu + \frac{5}{2}\right)\Gamma\left(\frac{\nu+5}{2}\right)^2}{\sqrt{\pi}\nu^{5/2}\Gamma(\nu+5)\Gamma\left(\frac{\nu}{2}\right)^2}, \tag{2.32}$$

Необъективность и беспристрастность алгоритмов кросс-валидации лишь немного сложнее реализовать для частотного полигона. Для BCV, Скоттом и Терреллом (1987) была предложена следующая оценка $R(f'')$:

$$\hat{R}(f'') = \frac{1}{n^2h^5}\sum_k(\nu_{k+1} - 2\nu_k + \nu_{k-1})^2 - \frac{6}{nh^5}. \tag{2.33}$$

Затыкание этой оценки в выражении AMISE приводит к

$$\mathrm{BCV}(h) = \frac{271}{480nh} + \frac{49}{2880n^2h}\sum_k (\nu_{k+1} - 2\nu_k + \nu_{k-1})^2. \tag{2.34}$$

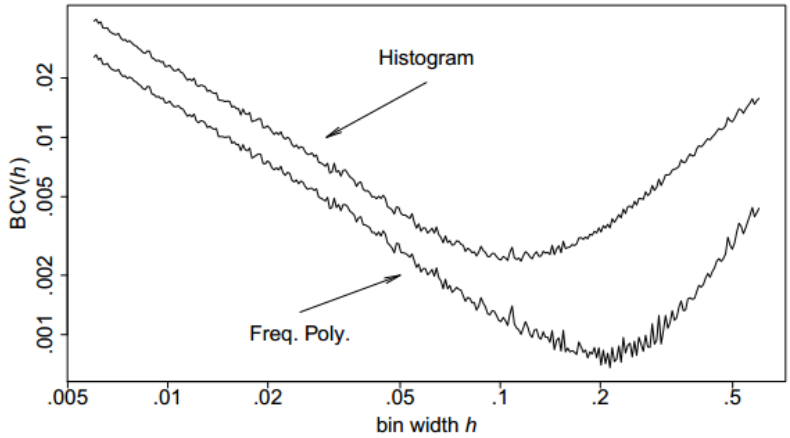

Рисунок 15 – BCV для гистограммы и частотного полигона для данных немецких доходов.

Рассмотрим теоретическое улучшение возможности при применении частотного полигона к адаптивным гистограммным сеткам. Однако, следует отметить, что соединение гистограммных средин в адаптивной сетке не приводит к оценке, интегрирующей до 1, за исключением асимптотики. С этой оговоркой следующие результаты являются следствием уравнения (2.29). Теорема: асимптотические свойства оптимального адаптивного частотного полигона, построенные путем соединения середин адаптивной гистограммы

$$\mathrm{AMSE}(x) = \frac{2f(x)}{3nh} + \frac{49}{2{,}880}h^4 f''(x)^2 \tag{2.35}$$

откуда следует, что

$$\begin{aligned} h^*(x) &= 2[15f(x)/49f''(x)^2]^{1/5}n^{-1/5} \\ \mathrm{AMSE}^*(x) &= (5/12)[49/15]^{1/5}[f''(x)^2 f(x)^4]^{1/5}n^{-4/5} \\ \mathrm{AAMISE}^* &= (5/12)[49/15]^{1/5}\left\{\int [f''(x)^2 f(x)^4]^{1/5}dx\right\} n^{-1/5}. \end{aligned} \tag{2.36}$$

Сравнивая полученные уравнения, отметим, что

$$\mathrm{AAMISE}^* \leq \mathrm{AMISE}^* \Leftrightarrow \int [f''(x)^2 f(x)^4]^{1/5}dx \leq \left[\int f''(x)^2 dx\right]^{1/5}, \tag{2.37}$$

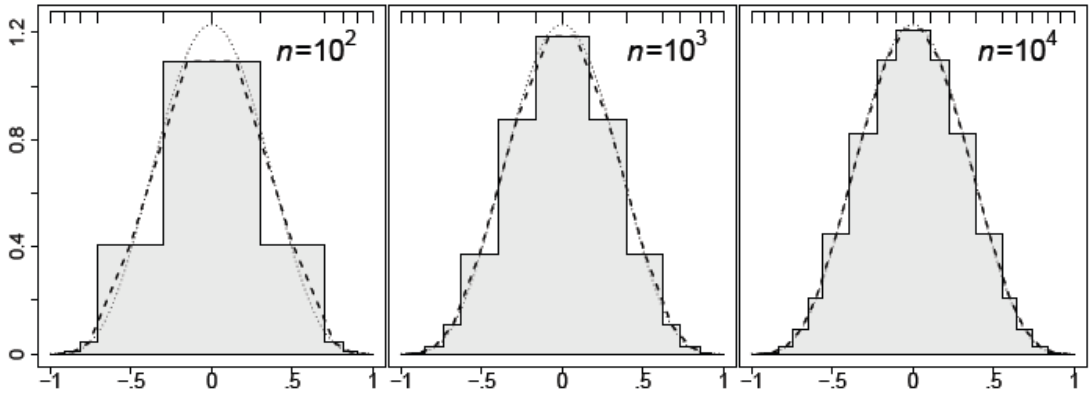

Рисунок 16 – Оптимальные адаптивные сетки частотного полигона для масштабируемой плотности. Отображены гистограммы, на которых ЧП (пунктирная линия) является производной. Минимальные шаги для адаптивной сетки показаны над каждой фигурой.

эквивалентно следующему неравенству (что справедливо в силу неравенства Йенсена):

$$\mathrm{E}\left[\frac{f''(x)^2}{f(x)}\right]^{1/5} \leq \left[\mathrm{E}\frac{f''(x)^2}{f(x)}\right]^{1/5}. \tag{2.38}$$

Таким образом, асимптотически, MISE адаптивного ЧП является лишь 91,5 % и 76,7 % от Mise из фиксированной ширины области ЧП для нормальных и Коши данных, соответственно.

MISE для ЧП адаптивной гистограммы можно вычислить точно. Так как в результате адаптивного ЧП не интегрируется в 1, суждение зарезервировано, как ее практическое значение; Тем не менее, есть большой интерес рассмотреть в структуре оптимальной сетки. Отметим, что асимптотически оптимальный адаптивный ЧП будет интегрироваться с 1, так как лежащая в основе адаптивная гистограмма точно интегрирует 1.

Общая картина в адаптивной сетке ЧП может быть выведена из теоремы. Сетка ЧП кажется вне фазы с оптимальной адаптивной гистограммной сеткой в критических точках. Области ЧП являются широкими, где вторая производная мала, как в точках перегиба так и, в меньшей степени, в хвостах. В промежутках области могут быть довольно узкими в зависимости от величины $f''(x)$. Рассмотрим оптимальную адаптивную сетку из масштабируемой близко к нормальной $Beta$ (5,5) плотности на рисунке 16. В хвостах, оптимальные области не намного шире. На самом деле, на картине относительно трудно увидеть для самого большого размера выборки за исключением того. Принимая во внимание не только сложность оптимальной адаптивной сетки, но и относительно умеренное снижение MISE, практические адаптивные алгоритмы не спешат появляться. Промежуточная стратегия будет выполнять преобразования данных, чтобы минимизировать

асимметрию или для обработки широко разделенных кластеров по отдельности.

2.2.3 Гистограммная арифметика

2.2.3.1 Операции над двумя гистограммными величинами

Пусть имеется система двух непрерывных случайных величин *(x_1,x_2)* с плотностью распределения *$p(x_1,x_2)$*. Известны аналитические формулы для определения плотности вероятности результатов арифметических действий над случайными величинами. Например, для нахождения плотности вероятности p_{x1+x2} суммы двух случайных величин x_1+x_2 используется соотношение

$$p_{x_1+x_2}(x) = \int_{-\infty}^{+\infty} p(x-v,v)dv = \int_{-\infty}^{+\infty} p(v,x-v)dv \qquad (2.39)$$

для нахождения плотности вероятности $p_{x1/x2}$ частного двух случайных величин x_1/x_2

$$p_{x_1/x_2}(x) = \int_0^{\infty} vp(x,v)dv - \int_{-\infty}^{0} vp(v,xv)dv \qquad (2.40)$$

плотность вероятности p_{x1x2} произведения двух случайных величин x_1x_2 определяется соотношением

$$p_{x_1x_2}(x) = \int_0^{\infty} \left(\frac{1}{v}\right) p\left(\frac{x}{v},v\right) dv - \int_{-\infty}^{0} (1/v)p(v,x/v)dv \qquad (2.41)$$

Однако эти формулы не всегда удобны для численных расчетов.

Основные принципы разработки гистограммных операций продемонстрируем на примере операции сложения [4]. Пусть $z = x_1 + x_2$ и носители $x_1 - [a_1, a_2]$, $x_2 - [b_1, b_2]$, $p(x_1, x_2) -$ плотность распределения вероятностей случайного вектора (x_1, x_2). Заметим, что прямоугольник $[a_1, a_2] \times [b_1, b_2]$ - носитель плотности распределения вероятностей $p(x_1, x_2)$ и плотность вероятности z отлична от нуля на интервале $[a_1 + b_1, a_2 + b_2]$. Обозначим $z_i, i = 0, \dots, n -$ точки деления этого интервала на n отрезков. Тогда вероятность попадания величины z в интервал $[z_i, z_{i+1}]$ определяется по формуле

$$P(z_i < z < z_{i+1}) = (\int\int_{\Omega_i} p(x_1, x_2)dx_1dx_2 \qquad (2.42)$$

где $\Omega_i = \{(x_1, x_2)|z_i \le x_1 + x_2 \le z_{i+1}\}$. И окончательно p_{zi} имеет вид

$$p_{zi} = (\int\int_{\Omega_i} p(x_1, x_2)dx_1dx_2/(z_{i+1} - z_i). \qquad (2.43)$$

Рассмотренный подход обобщается на случай большего числа переменных [9].

2.2.3.2 Арифметика неопределенных данных на основе гистограмм второго порядка

Для осуществления численных операций над «неопределенными» переменными, заданными своими функциями плотности в виде гистограмм второго порядка, в условиях неопределенности, определим арифметику для гистограмм второго порядка (ГВП).

Пусть X, Y – ГВП, определяются сетками $\{v_i, i=0,1,\ldots n\}$, $\{w_i, i=0,1,\ldots n\}$ и наборами гистограмм $\{Px_i\}$, $\{Py_i\}$. Пусть $Z=X*Y$, где $*\in\{+,-,*,/,\uparrow\}$. Построим Z как ГВП. Зададим сетку $\{z_i, i=0,1,\ldots n\}$, тогда гистограмма Pz_i на отрезке $[z_{i-1}, z_i]$ определяется по формуле

$$Pz_i = \int_{\Omega_i} X(\xi)Y(\eta)d\xi d\eta. \qquad (2.44)$$

где $\Omega_i = \{(\xi,\eta)|z_i \le \xi * \eta \le z_{i+1}\}$. Заметим, что на каждом прямоугольнике $[v_{i-1}, v_i]\times[w_{j-1},w_j]$ функция $X(\xi)Y(\eta)$ – есть постоянная гистограмма Px_iPy_j. Интеграл от гистограммы по некоторой области – есть значение гистограммы, умноженное на площадь области.

Проиллюстрируем, как работает гистограммная арифметика в случае сложения двух ГВП.

Пример. Пусть необходимо сложить две гистограммы второго порядка X и Y. Гистограммы X и Y порождены равномерными случайными величинами, заданными соответственно на отрезках $[0, t_1]$ и $[t_2, 2]$, где t_1 – равномерная случайная величина, заданная на отрезке $[1, 2]$, t_2 – равномерная случайная величина, заданная на отрезке $[0, 1]$. Результат сложения двух гистограмм представлен в виде гистограммы второго порядка Z, изображенной на рисунке 17. Носителем Z является отрезок $[0, 4]$, высота 1, значения плотности вероятности представлены оттенками серого.

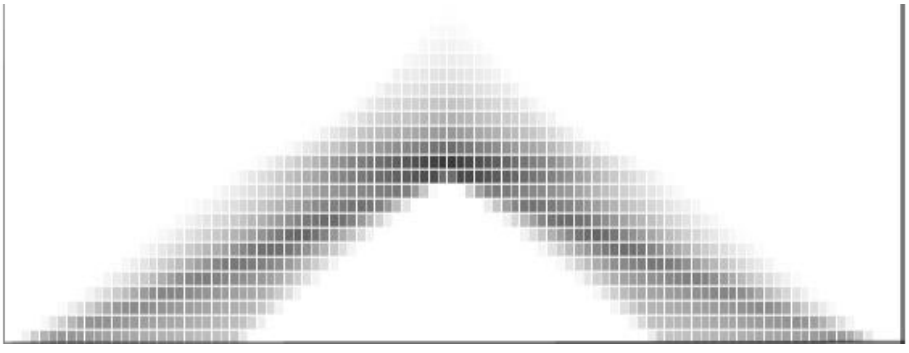

Рисунок 17 – Сумма двух гистограмм второго порядка Z = X+Y

Использование гистограмм второго порядка в условиях неопределенности вероятностных характеристик параметров системы имеет широкую область применения к решению разнообразных практических задач.

Например, к решению задач оценки эффективности инвестиционных проектов. В данном случае мы имеем дело с высоким уровнем рыночной неопределенности, где стандартная финансовая модель не учитывает вероятностную природу характеристик оцениваемых показателей и соответственно не может дать достаточных оснований для принятия эффективных инвестиционных решений, а используемые методы, например, метод дисконтированных денежных потоков, не учитывает вероятностный характер результатов инвестиционных проектов. Гистограммы второго порядка также могут успешно применяться в задачах оценки показателей надежности и оценки безотказной работы сложных технических систем, для изучения гидрологических и других систем. Решение практических задач с использованием методов интерполяции и экстраполяции также лежит в сфере применения ГВП [10].

2.2.4 Вероятностные расширения и их свойства

Пусть $f(x_1,...,x_n)$ — рациональная функция, тогда для вычисления гистограммы F заменим арифметические операции на гистограммные, а переменные x_1, x_2, ..., x_n— их гистограммными значениями. Полученную гистограмму F будем называть — естественным гистограммным расширением [12].

Теорема. Пусть $f(x_1,...,x_n)$ — рациональная функция, каждая переменная которой встречается только один раз и x_1,..., x_n — независимые случайные величины. Тогда естественное гистограммное расширение аппроксимирует вероятностное расширение с точностью $O(h^\alpha)$ Доказательство проведем по индукции. Для n = 2 утверждение справедливо [10]. Пусть справедливо для $n = k$ и гистограмма F_k аппроксимирует плотность вероятности функции $f(x_1, ..., x_k)$ с некоторой точностью $O(h^\alpha)$. Покажем, что это справедливо и для $n = k + 1$. Действительно, $F_k+ 1=F_k*x_k+1$, но

$F_k+1 - f(x_1, ..., x_k, x_{k+1}) = (F_k - f(x_1, ..., x_k)) *x_{k+1} \le Ch^\alpha *\text{supp}\{x_{k+1}\}$. Теорема доказана.

Пример. Для рациональной функции $f(x, y) = xy+ x + y + 1 = (x + 1)(y + 1)$ только второе представление попадает под условие Теоремы 1 и, следовательно, естественное гистограммное расширение будет аппроксимировать вероятностное с некоторой точностью $O(h^\alpha)$. Теорема 1 легко обобщается на следующий случай [14,17].

Замечание 1. Пусть для функции $f(x_1,...,x_n)$ возможна замена переменных, такая что $f(z_1,..., z_k)$ — рациональная функция от переменных z_1,..., z_k, удовлетворяющая условиям Теоремы 1 и z_i — функции от множества переменных x_i, $i \in Ind_i$, причем множества Ind_i попарно не пересекаются. Пусть для каждой z_i существуют вероятностные расширения. Тогда естественное расширение $f(z_1,..., z_k)$ будет аппроксимировать вероятностное с некоторой точностью [10].

Сравнение с Монте-Карло. Известно, что метод Монте-Карло имеет сходимость $O(1/\sqrt{N})$, где N — число повторов, гистограммные расширения имеют скорость сходимости $O(1/n^{\alpha})$. Пусть необходимо достигнуть точности ε, число операций метода Монте-Карло при этом $\approx \varepsilon^{-2}$ в сравнении с гистограммными расширениями $\approx \varepsilon^{-2/\alpha}$. Таким образом, гистограммная арифметика эффективней метода Монте-Карло примерно в $\varepsilon^{-2(1-1/\alpha)}$ раз.

2.2.5 Сглаживание эмпирических данных

Если данные сильно зашумлены, то имеет смысл произвести их сглаживание. Однако следует иметь ввиду, что сглаживание данных приводит к тому, что стандартное предположение относительно нормального распределения ошибки не будет выполняться. Поэтому сглаживание обычно применяется для получения информации о возможном выборе типа параметрической модели, а сам процесс подбора параметров параметрической модели проводят для исходных (несглаженных) данных.

Сглаживание производится по нескольким расположенным подряд данным, причем их число обычно подбирается экспериментально [34].

В методе скользящего среднего исходные данные y_i сглаживаются по следующему правилу:

$$ys_i = \frac{1}{2N+1}\sum_{k=-N}^{N} y(i+k), \tag{2.45}$$

где $2N + 1$ - число точек, выбираемых для сглаживания, т.е. слева и справа от текущей точки выбирается по N точек (ясно, число точек, участвующих в сглаживании, должно быть нечетным). Данные, расположенные в точках, близких к границам отрезка, не сглаживаются, т.к. не хватает точек справа или слева от текущей, в которой в данный момент производится Сглаживание [19]. Ниже на рисунке 18 приведены исходные зашумленные данные: *x=0:0.002:5 и y=x*sin(3*x)+2*randn(size(x)).*

Сглаженные с $N = 33$ точками и сглаженные с $N = 191$ точкой. Красная линия соответствует незашумленным данным

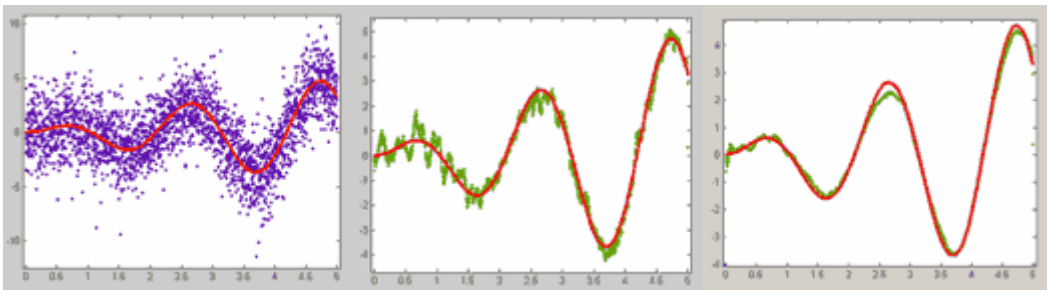

Рисунок 18 – Левый график отображает зашумленные и незашумленные данные; Средний график – сглаженные с $N = 33$ и незашумленные. Правый график – сглаженные с $N = 191$ и незашумленные.

При сглаживании при помощи взвешенной локальной регрессии для каждого сглаживаемого значения данных, заданного в точке , выбирается набор из фиксированного числа рядом расположенных точек, каждой из которых назначается вес по следующей формуле

$$w_i = \left(1 - \left|\frac{x_{k_c} - x_i}{d(x_{k_c})}\right|^3\right)^3, \tag{2.46}$$

где $d(x_k)$ - расстояние от x_k до наиболее удаленной точки из набора. Т.е. наибольший вес (равный единице) будет у y_k, а наименьший (равный нулю) у данных, расположенных на границах набора. Распределение весов для некоторой точки внутри интервала, где заданы данные, и на краях, показано на следующем графике:

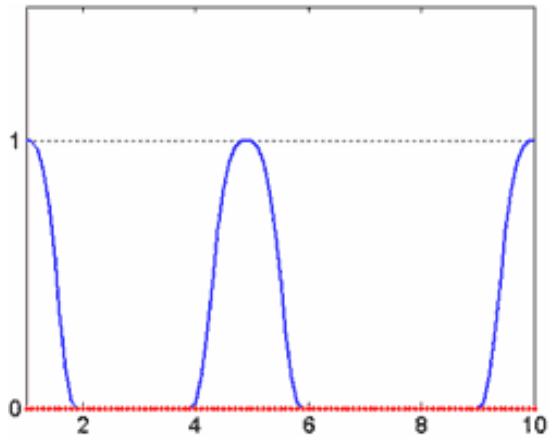

Рисунок 19 - Значения весов для взвешенной локальной регрессии, в которых заданы данные

2.3 Вывод по главе 2

Анализ методов обработки данных позволил разделить этим методы на 2 группы, обрабатывающие данные с элиторной неопределенностью, и данные с эпистемической неопределенностью.

На рисунке 20 представлена классификации методов для обработки данных в зависимости от типа неопределенности в данных.

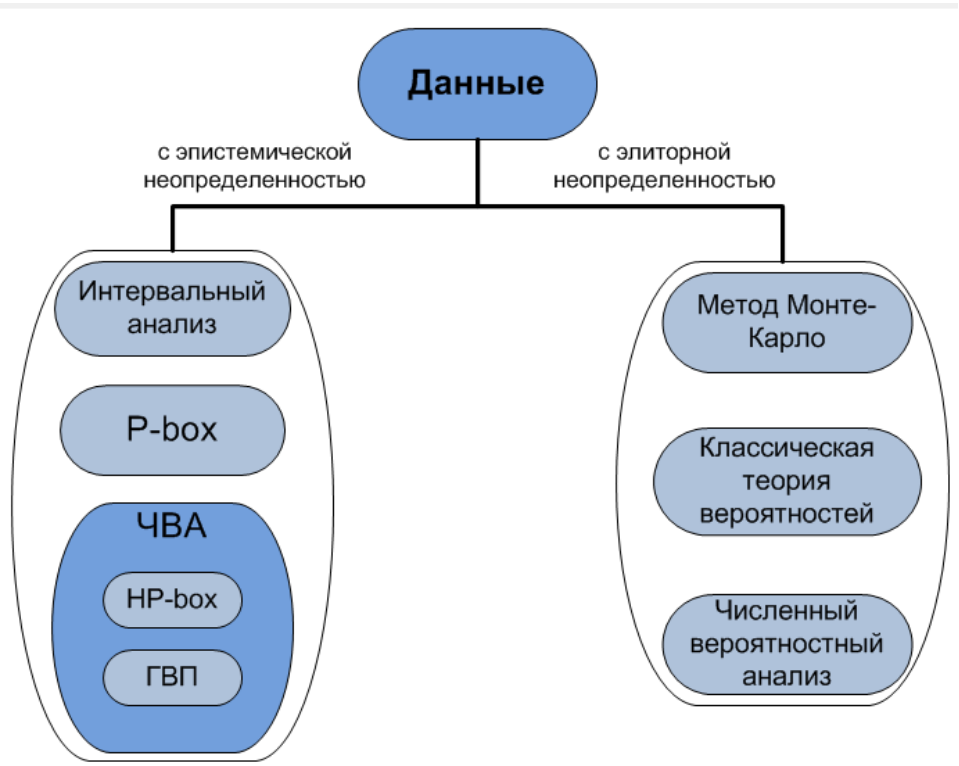

Рисунок 20 – Классификация методов для обработки данных в условиях неопределенностей

Однако, данные методы направлены на решение задач с эпистемической неопределенностью, обусловленной неопределенностью вероятностных оценок, а задачи с эпистемической неопределенностью, обусловленной недостатком знаний о системе, например, задачи в условиях малых объемов выборок.

3 Разработка модуля для численного моделирования эмпирических данных

3.1 Общая характеристика модуля

В процессе проектирования системы на первом шаге необходимо определить решаемые системой проблемы.

Повышение точности определения поведения системы позволит аналитикам при решении задач в условиях неопределенностей принимать эффективные решения.

Данный программный модуль предназначен для проведения численных экспериментов по восстановлению функции плотности вероятности на основе различных подходов.

Программный модуль представляет собой самостоятельный раздел, решающий задачу восстановления функции плотности вероятности случайной величины X методом ядерного восстановления функции плотности вероятности и методами оценки функции плотности вероятности случайной величины, представленной в виде полиграмм.

3.2 Структура и описание основных блоков

В модуле реализована оценка функции плотности вероятности методом ядерного восстановления функции плотности вероятности, методом полиномиального сглаживания случайной величины X, представленной в виде полиграммы, оценка функции плотности вероятности методом скользящего среднего.

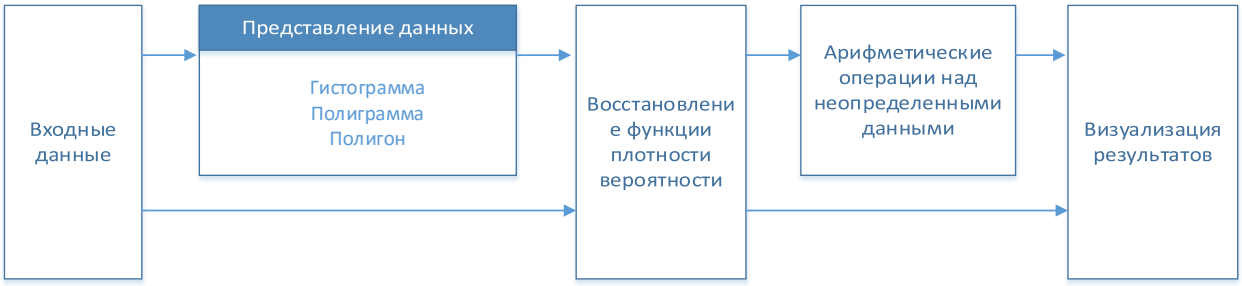

Рисунок 21 – Общая структура модуля

Объектом исследования является случайная величина, имеющая треугольное распределение.

Случайная величинам ξ имеет треугольное распределение (распределение Симпсона) на отрезке $[a,b](a<b)$, если функция плотности вероятности имеет вил:

$$f(x)=\begin{cases}\frac{2}{b-a}-\frac{2}{(b-a)^2}|a+b-2x|, & x\in[a,b],\\ 0, & x\notin[a,b]\end{cases}. \quad (3.1)$$

Характеристическая функция (комплекснозначная функция вещественной переменной) треугольного распределения

$$\varphi(t) = \left[\frac{2(e^{itb/2} - e^{ita/2})}{(b-a)it}\right]^2. \tag{3.2}$$

Свойства треугольного распределения [22]:

– моменты:

$$\mathbf{E}\xi^k = \frac{4}{(b-a)^2(k+1)(k+2)}\left[a^{k+2} + b^{k+2} - 2\left(\frac{a+b}{2}\right)^{k+2}\right]; \tag{3.3}$$

– дисперсия:

$$\mathbf{D}\xi = \frac{(b-a)^2}{24}; \tag{3.4}$$

– коэффициент асимметрии $\gamma_1 = 0$;

– коэффициент эксцесса $\gamma_2 = -3/5$.

Если ξ_1 и ξ_2 - независимые случайные величины, равномерно распределенные на отрезке $[a/2, b/2]$, то случайная величина $\xi = \xi_1 + \xi_2$ имеет треугольное распределение [22].

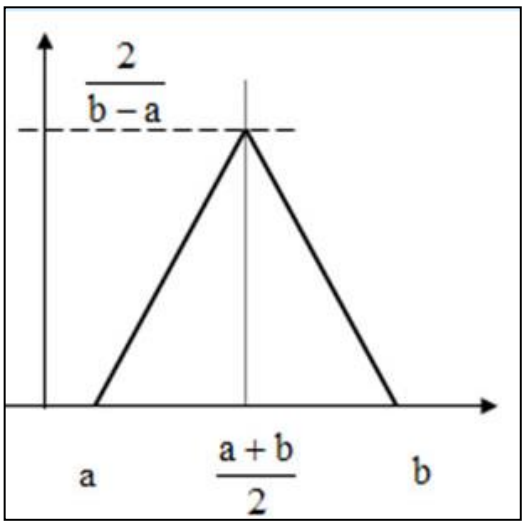

Рисунок 22 - Треугольное распределение

3.2.1 Постановка задачи

Имеется случайная величина *X*, которая характеризуется *f(x)*, статистической характеристикой - дисперсией. Предположим, что имеющаяся случайная величина имеет треугольное распределение.

1) Используя гауссову ядерную оценку плотности распределения непрерывной случайной величины восстановить функцию распределения случайной величины.

Необходимо сгенерировать случайную величину *X* объемом выборки *n = 5, 10, 20, 40*.

Для каждой выборки подобрать значение параметра дисперсии, при которой разница между полученной оценкой функции плотности распределения и треугольным распределением во второй норме будет минимальна.

Повторить эксперимент 10 раз для каждого объема выборки.

2) Построить идеальную выборку *n = 5, 10, 20, 40* с треугольным законом распределения, рассчитать для нее ядерную оценку плотности вероятности, рассчитать ошибку, построить графики.

3.2.2 Алгоритм решения задачи

С помощью программы, написанной на языке Pascal, генерируем случайную величину с треугольным законом распределения объемом выборки n с помощью генератора случайных чисел следующим образом:

x[i] = random + random (3.5)

где random - генератор случайных чисел от 0 до 1 с равномерным законом распределения.

Ядерная оценка плотности распределения непрерывной случайной величины определяется по следующей формуле:

$$f_n(t) = \frac{1}{n\sigma\sqrt{2\pi}} \sum_{i=1}^{n} \exp\left[\left(\frac{t-\xi_i}{\sigma\sqrt{2}}\right)^2\right] \quad (3.6)$$

где n - объем выборки; σ - параметр локальности (ширина окна ядра оценки).

На рисунке 22 представлена блок-схема алгоритма решения задачи.

Обозначения на блок-схеме:

– n - объем выборки;

– σ - дисперсия;

– разница между полученной ядерной оценкой распределения и треугольным распределением определяется по формуле:

–

$$\varepsilon_n = ||p - \varphi_n||_2 = \left(\int_0^2 (p(x) - \varphi(x))^2 dx\right)^2 = \left(\left[\sum_{l=1}^{N-1} g(lh) + \frac{1}{2}g(0) + \frac{1}{2}g(2)\right] 2\right)^{1/2} \quad (3.7)$$

где $\left(p(x) - \varphi(x)\right)^2 = g(x), \quad h = 2/N.$

На рисунке 22 представлена блок-схема решения задачи ядерного восстановления функции плотности вероятности случайной величины.

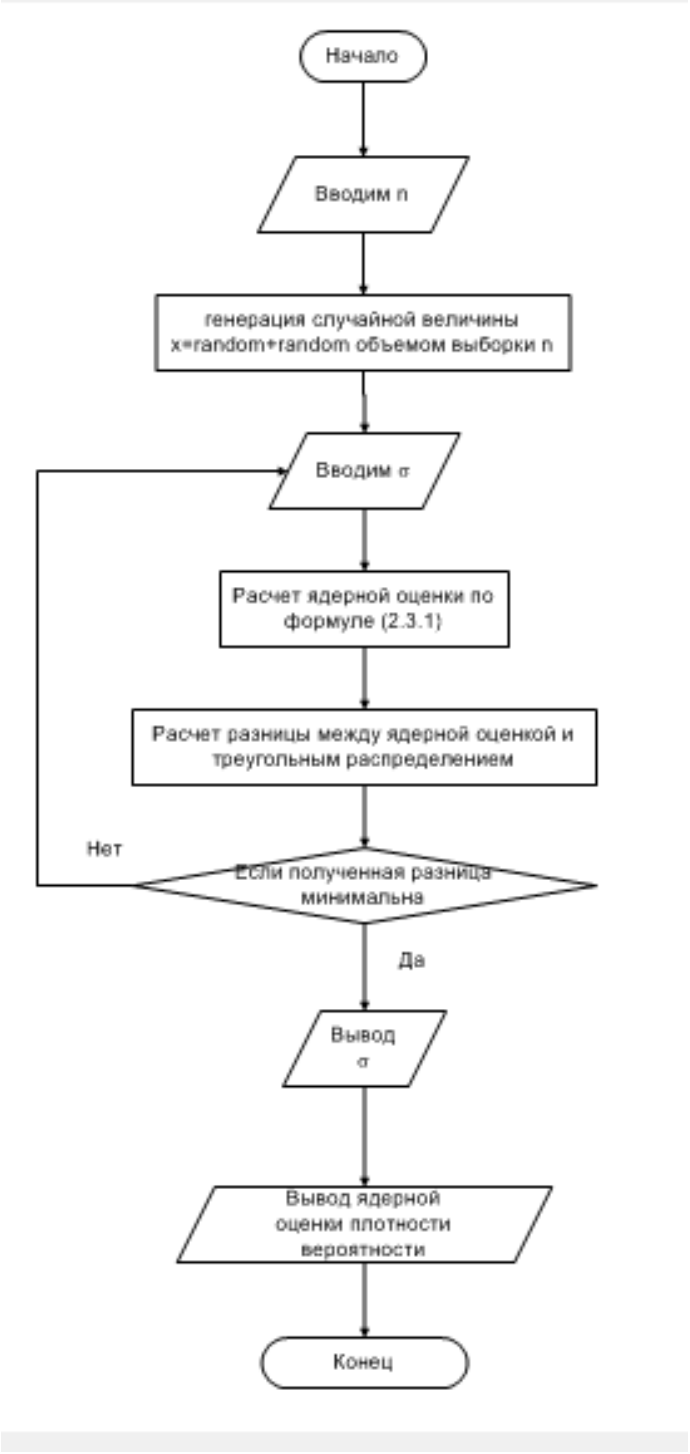

Рисунок 22 - Блок схема решения задачи

3.2.3 Результаты ядерного восстановления функции плотности вероятности

В таблице 2 представлены результаты численного эксперимента подбора параметра σ для выборок объемом n = 5, 10, 20, 40, по 10 экспериментов для каждого объема выборки n. Столбцы представляют номер эксперимента для

каждой выборки. Ячейки таблицы содержат пару чисел: параметр σ, который необходимо было подобрать таким образом, при котором разница между полученной оценкой функции плотности распределения и треугольным распределением во второй норме будет минимальна, p-φ представляет саму разницу между треугольным распределением и оценкой функции плотности распределения.

Таблица 2 - Результаты численного эксперимента

		Номер эксперимента									
Объем выборки		1	2	3	4	5	6	7	8	9	10
n=5	σ	0,4300	0,6150	0,3550	0,2700	0,4100	0,4350	0,2440	0,5850	0,3450	0,3750
	p-ϕ	0,0481	0,1550	0,0722	0,0140	0,0620	0,0512	0,0315	0,1222	0,1068	0,0069
n=10	σ	0,3200	0,2100	0,3700	0,2700	0,2900	0,3400	0,3400	0,2800	0,3900	0,2600
	p-ϕ	0,0258	0,0226	0,0708	0,0043	0,0086	0,0686	0,0155	0,0188	0,0454	0,0274
n=20	σ	0,2400	0,2300	0,2800	0,1800	0,3200	0,2500	0,2400	0,2400	0,3200	0,3000
	p-ϕ	0,0144	0,0206	0,0255	0,0098	0,0288	0,0111	0,0205	0,0184	0,0031	0,0259
n=40	σ	0,1500	0,2000	0,3400	0,1800	0,2200	0,1900	0,2900	0,2500	0,2300	0,2100
	p-ϕ	0,0047	0,0134	0,0258	0,0089	0,0396	0,0282	0,0086	0,0549	0,0244	0,0222

Из полученной таблицы видно, что при большем объеме выборки n параметр σ в среднем принимает меньшие значения. Так же можно отметить, что при меньших объемах выборки разница между треугольным распределением и полученной оценкой функции плотности распределения больше, чем при больших объемах выборки.

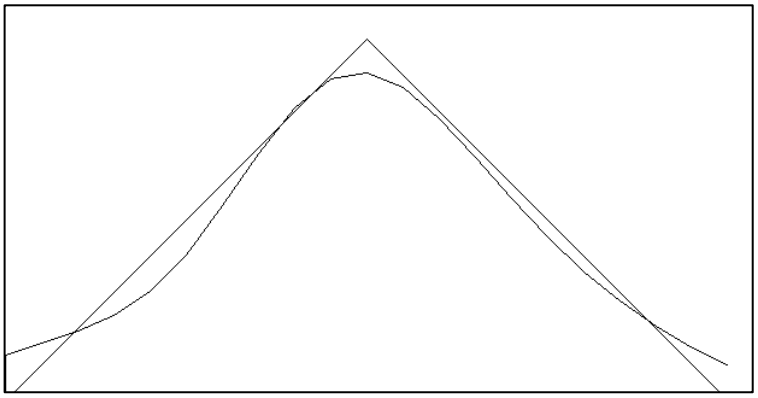

Рисунок 23 - Результат ядерного восстановления функции плотности вероятности

Так же был проведен численный эксперимент с идеальными выборками объемом n = 5, 10, 20, 40. На таблице 2 представлены значения идеальных выборок.

Таблица 3 – Значения идеальных выборок

Идеальная выборка с треугольным законом распределения							
n=5	n=10	n=20		n=40			
0,63	0,45	0,32	1,05	0,22	0,74	1,03	1,33
0,89	0,63	0,45	1,11	0,32	0,77	1,05	1,37
1,11	0,77	0,55	1,16	0,39	0,81	1,08	1,41
1,37	0,89	0,63	1,23	0,45	0,84	1,11	1,45
2,00	1,00	0,71	1,29	0,5	0,87	1,13	1,5
	1,11	0,77	1,37	0,55	0,89	1,16	1,55
	1,23	0,84	1,45	0,59	0,92	1,19	1,61
	1,37	0,89	1,55	0,63	0,95	1,23	1,68
	1,55	0,95	1,68	0,67	0,97	1,26	1,78
	2,00	1,00	2,00	0,71	1,00	1,29	2,00

Для каждой выборки был проведен расчет ядерной оценки функции плотности распределения случайной величины X, подобран параметр σ, такой, что разница между треугольным распределением и полученной оценкой функции плотности распределения принимала минимальное значение, построены графики полученных оценок.

На рисунке 24 представлен график ядерной оценки функции плотности распределения случайной величины X с объемом выборки n = 5.

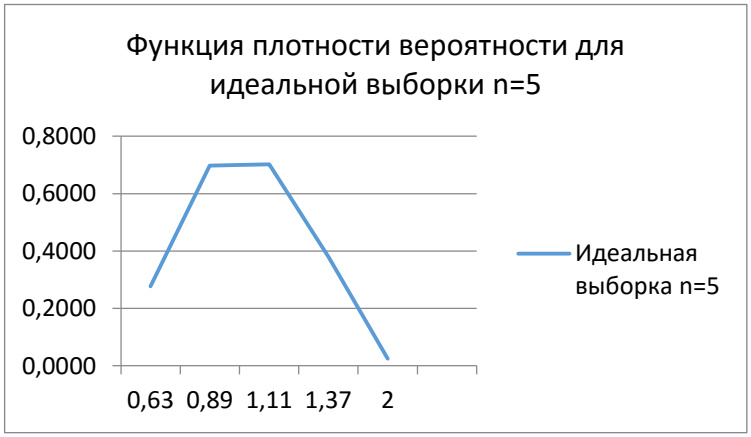

Рисунок 24 - Результат ядерного восстановления функции плотности вероятности

На рисунке 25 представлен график ядерной оценки функции плотности распределения случайной величины X с объемом выборки n = 10.

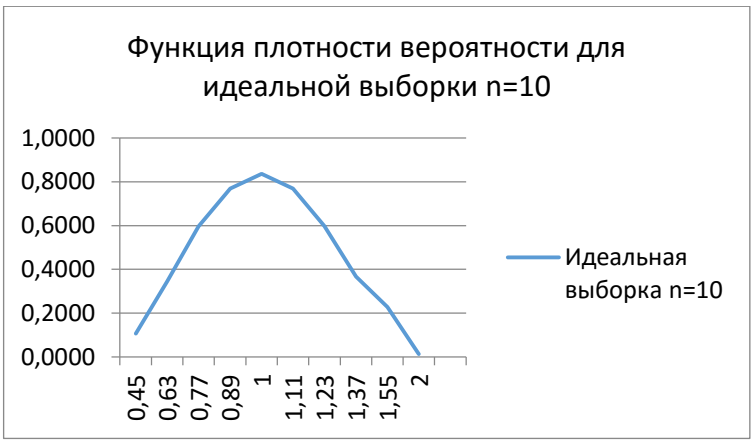

Рисунок 25 - Результат ядерного восстановления функции плотности вероятности

На рисунке 26 представлен график ядерной оценки функции плотности распределения случайной величины X с объемом выборки n = 20.

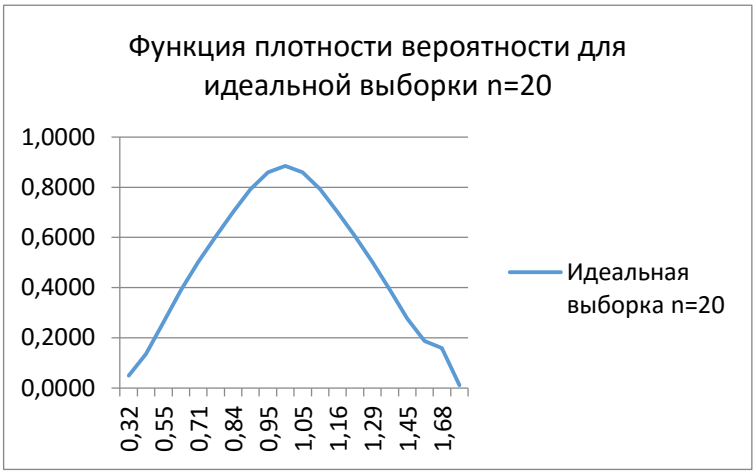

Рисунок 26 - Результат ядерного восстановления функции плотности вероятности

На рисунке 27 представлен график ядерной оценки функции плотности распределения случайной величины X с объемом выборки n = 40.

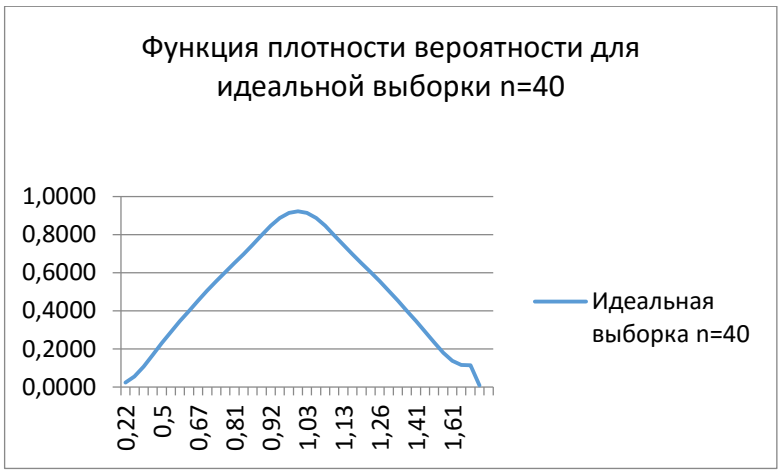

Рисунок 27 - Результат ядерного восстановления функции плотности вероятности

Из рисунков видно, что чем больше объем выборки n, тем больше ядерная оценка функции плотности распределения стремится к треугольному распределению.

Так же, численный эксперимент показал, что при большем объеме выборки n, ошибка ядерной оценки функции плотности распределения случайной величины уменьшается.

В таблице 4 представлены результаты данного эксперимента при параметре σ = 0,2.

Таблица 4 - Расчет ошибки ядерной оценки для идеальной выборки при постоянном параметре σ

	Объем выборки			
	n = 5	n = 10	n = 20	n = 40
σ	0,2000	0,2000	0,2000	0,2000
$\varepsilon = \|p - \varphi\|_2$	0.0510	0.0135	0.0061	0.0047

Таким образом, численный эксперимент показал, что

$$\varepsilon(5) > \varepsilon(10) > \varepsilon(20) > \varepsilon(40).$$

3.3 Оценка функции плотности с помощью полиномиального сглаживания

С помощью программы, написанной на языке Pascal, генерируем случайную величину с треугольным законом распределения объемом выборки n с помощью генератора случайных чисел следующим образом:

$$x[i] = random + random. \quad (3.8)$$

где random - генератор случайных чисел от 0 до 1 с равномерным законом распределения.

Полиномиальное сглаживание будет производиться с помощью парабол. Для этого были сформированы следующие базисные функции согласно рисунка 28:

$$a_1(1-x^2)=\varphi_1$$
$$a_2(1-x)x=\varphi_2$$
$$-a_3(1-x)x=\varphi_3 \quad (3.9)$$

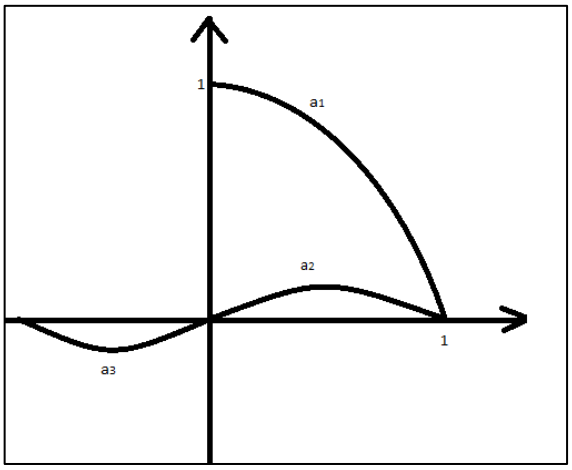

Рисунок 28 – Полиномы для сглаживания полиграммы

При этом $x \to \frac{x-x_0}{h}$.

Методом наименьших квадратов определяются коэффициенты базисных функций:

$$\Phi(a)=\sum_{j=1}^{N}(\sum_{i=1}^{3} a_i\varphi_i(\xi_j)-z_j)^2 \to min \quad (3.10)$$

$$\frac{d\Phi}{da_i}=\sum_{j=1}^{N}\left(\sum_{i=1}^{3} a_i\varphi_i(\xi_j)-z_j\right)\varphi_l(\xi_j)=0 \quad (3.11)$$

$$a_{il}=\sum_{j=1}^{N}\varphi_i(\xi_j)\varphi_l(\xi_j) \quad (3.12)$$

$$b_l=\sum_{j=1}^{N} z_j\varphi_l(\xi_j) \quad (3.13)$$

В результате примененного полиномиального сглаживания получили следующие результаты. На рисунке 29 представлен результат восстановления функции плотности вероятности случайной величины *X*, представленной в виде полиграммы, с помощью полиномиального сглаживания.

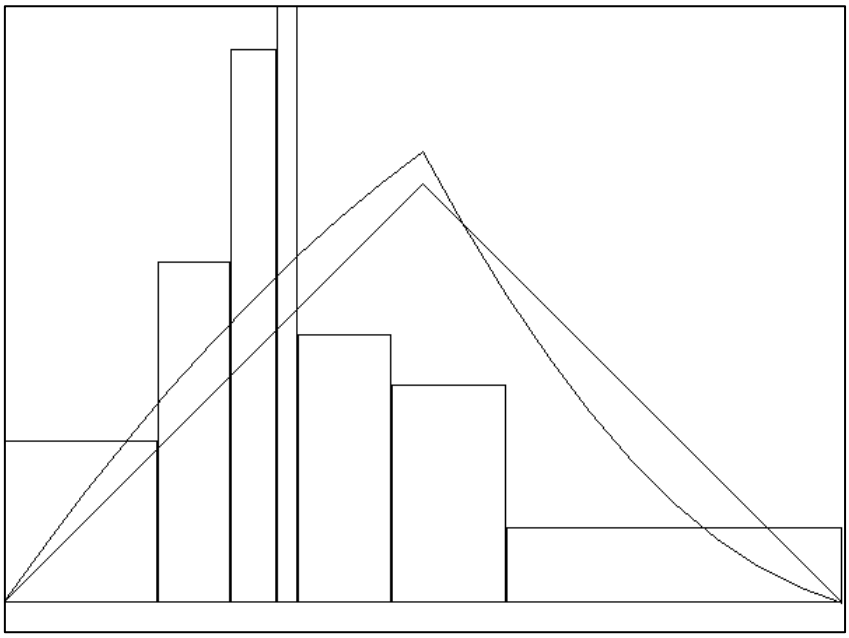

Рисунок 29 - Результат полиномиального сглаживания для восстановления функции плотности вероятности

Результаты оценки плотности вероятности случайной величины *X*, представленной в виде полиграммы, методом скользящего среднего представлены на рисунке 30.

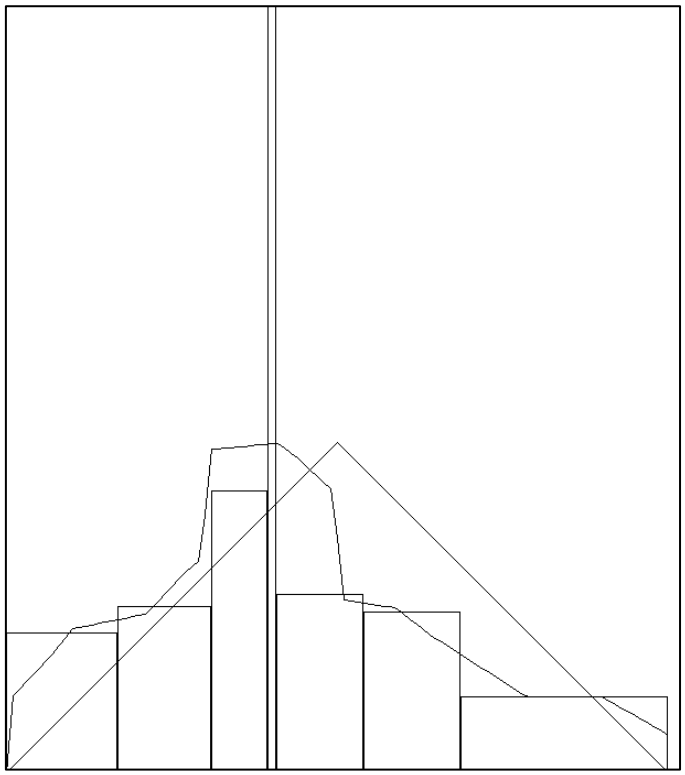

Рисунок 30 - Результат сглаживания методом скользящего среднего для восстановления функции плотности вероятности

3.4 Сравнение рассмотренных методов на примере одной задачи

Рассмотрим несколько примеров оценки плотности вероятности случайной величины *X*, построенной по треугольному закону распределения:

1) Пусть имеется выборка $\Xi = \{\xi_1,...,\xi_n\}$, $n = 7$ случайной величины *X*. Плотность случайной величины *X* аппроксимирована полиграммой. Далее, методом скользящего среднего с шириной окна равной *0,06* производим сглаживание полиграммы. Результат представлен на рисунке 32.

2) Пусть как и примере (1) имеется выборка $\Xi = \{\xi_1,...,\xi_n\}$, $n = 7$ случайной величины *X*. Приблизим плотность вероятности *X* с помощью полиграммы. Далее, произведем сглаживание полиграммы с помощью параболического базиса. Результат представлен на рисунке 33.

3) Пусть имеется выборка $\Xi = \{\xi_1,...,\xi_n\}$, $n = 7$ случайной величины *X,* та же, что использовалась в первом и втором экспериментах. Произведем гауссову ядерную оценку плотности вероятности для случайной величины *X,* используя простое прямоугольное ядро с параметром $\sigma=0,28$. На рисунке 31. представлена ядерная оценка плотности вероятности в пределах отрезка [0,2].

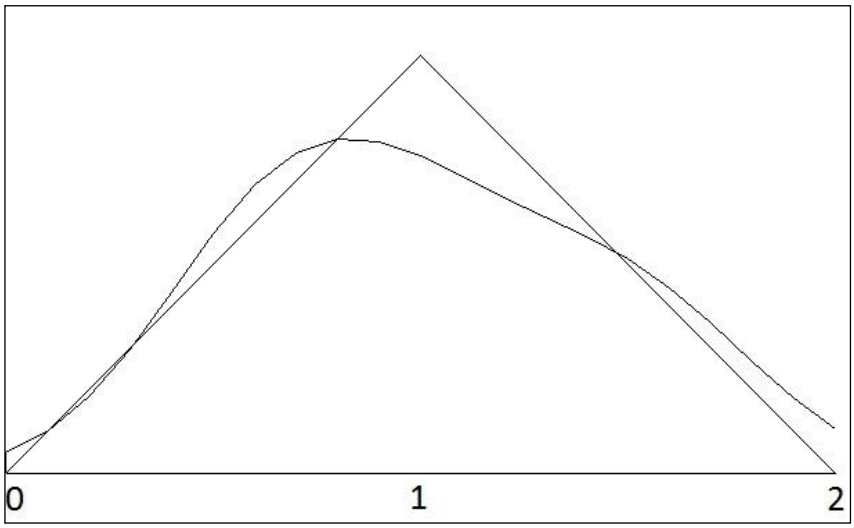

Рисунок 31 – Результат ядерного восстановления функции плотности вероятности случайной величины

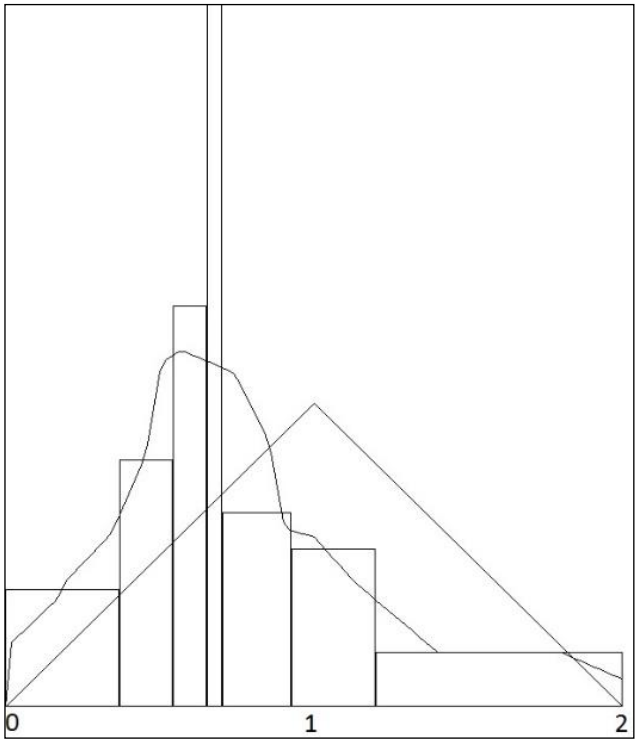

Рисунок 32 - Результат восстановления функции плотности вероятности случайной величины, представленной в виде полиграммы, с помощью сглаживания методом скользящего среднего

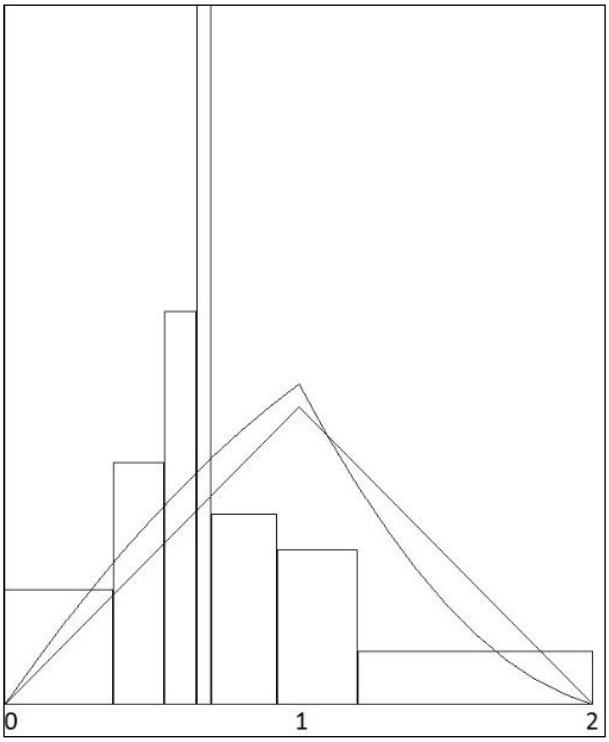

Рисунок 33 - Результат восстановления функции плотности вероятности случайной величины, представленной в виде полиграммы, с помощью полиномиального сглаживания

По результирующим рисункам видно, что наиболее близкая к треугольному распределению оценка плотности вероятности получилась в результате проведения второго эксперимента [21].

3.5 Вывод по главе 3

Проведены численные эксперименты восстановления функции плотности вероятности с использованием ядерных оценок. Данные эксперименты показали, что ядерное восстановление функции плотности вероятности с увеличением объема выборки данных дает более лучшие результаты. Был проведен эксперимент на идеальных выборках, когда случайная величина была отобрана таким образом, что каждое значение лежало на функции распределения. В условиях идеальной выборки ядерные оценки показали достаточно точные результаты при объеме выборки n=20.

Так же, был проведен численный эксперимент восстановления функции плотности вероятности случайной величины X в условиях малой выборки, объем выборки n=7. Восстановление функции плотности вероятности проводилось методом ядерного восстановления функции плотности вероятности, методом сглаживания случайной величины, представленной в виде полиграммы скользящим средним и методом полиномиального сглаживания. Результаты показали, что представление случайной величины с помощью полиграмм и последующая их обработка является перспективным направлением, и уже на первоначальных этапах исследования, с применением простых методов, дает достаточно точные результаты.

ЗАКЛЮЧЕНИЕ

Неопределенность в данных можно классифицировать по нескольким признакам: по типу неопределенностей можно выделить элиторную неопределенность, которая характеризуется изменчивостью процессов и состояний систем, эпистемическую неопределенность, характеризующуюся неопределенностью самих вероятностных оценок и недостаточностью знаний о системе; по видам неопределенных данных выделяют случайные, нечеткие, интервальные данные. Данные, содержащие случайную неопределенность, задаются некоторыми вероятностными распределениями их возможных значений; «нечеткие» данные задаются лингвистически сформулированными распределениями их возможных значений; данные, содержащие интервальную неопределенность, задаются интервалами их возможных значений без указания какого-либо распределения возможных значений числа внутри заданного интервала. Следует отметить, что для каждого вида неопределенных данных разработана своя арифметика.

Анализ методов обработки данных позволил разделить этим методы на 2 группы, обрабатывающие данные с элиторной неопределенностью, и данные с эпистемической неопределенностью.

Однако, данные методы направлены на решение задач с эпистемической неопределенностью, обусловленной неопределенностью вероятностных оценок, а задачи с эпистемической неопределенностью, обусловленной недостатком знаний о системе, например, задачи в условиях малых объемов выборок.

Проведены численные эксперименты восстановления функции плотности вероятности с использованием ядерных оценок. Данные эксперименты показали, что ядерное восстановление функции плотности вероятности с увеличением объема выборки данных дает более лучшие результаты. Был проведен эксперимент на идеальных выборках, когда случайная величина была отобрана таким образом, что каждое значение лежало на функции распределения. В условиях идеальной выборки ядерные оценки показали достаточно точные результаты при объеме выборки $n \geq 20$.

Проведен численный эксперимент восстановления функции плотности вероятности случайной величины X в условиях малой выборки, объем выборки $n=7$. Восстановление функции плотности вероятности проводилось методом ядерного восстановления функции плотности вероятности, методом сглаживания случайной величины, представленной в виде полиграммы скользящим средним и методом полиномиального сглаживания. Результаты показали, что представление случайной величины с помощью полиграмм и последующая их обработка является перспективным направлением, и уже на первоначальных этапах исследования дает достаточно точные результаты. Эксперимент показал перспективность исследования по применению численных методов обработки данных, представленных на основе полиграмм.

В результате, по теме магистерской диссертации опубликованы две статьи:

1) арифметики и численный вероятностный анализ неопределенных данных;
2) полиграммы для представления случайных данных.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1) Ferson S. Arithmetic with uncertain numbers / S. Ferson, J.G. Hajagos // Reliability Engineering and System Safety. - 2004. - 85 - С. 135–152.
2) Ferson S. What Monte-Carlo cannot do / S. Ferson // Human and Ecological Risk Assessment: An International Journal. - 1996.
3) Гаскаров Д. В., Шаповалов В. И. Малая выборка. – М.: Статистика, 1978. – 248 с
4) Герасимов В. А. Численные операции гистограммной арифметики и их применения / В. А. Герасимов, Б. С. Добронец, М. Ю. Шустров. - ВЦ СО РАН СССР, Красноярск, 1991. - С. 83-88.
5) Граничин О. Н. Введение в методы стохастической оптимизации и оценивания: Учеб. пособие / О. Н. Граничин. - СПб.: Издательство С.-Петербургского университета, 2003. - 131с. УДК 519.712, ББК 32.811.7
6) Добронец Б. С. Гистограммный подход к представлению и обработке данных космического и наземного мониторинга / Б. С. Добронец, О. А. Попова // Известия Южн. фед. ун-та. Технические науки. - 2013. - С. 14-23.
7) Добронец Б. С. Интервальная математика: Учеб. пособие / Б. С. Добронец. - Красноярск: Краснояр. гос. ун-т, 2004. - 216с. УДК 519
8) Добронец Б. С. Представление и обработка неопределенности на основе гистограммных функций распределения и P-Boxes / Б. С. Добронец, О. А. Попова
9) Добронец Б. С. Численный вероятностный анализ для исследования систем в условиях неопределенности / Б. С. Добронец, О. А. Попова // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. - 2012. - № 4 - С. 39-46.
10) Добронец Б. С. Численный вероятностный анализ неопределенных данных: монография / Б. С. Добронец, О. А. Попова. - Красноярск: Сиб. фед. ун-т., 2014. - 168 с.
11) Добронец Б. С. Численный вероятностный анализ для оценки инвестиционных проектов // Б. С. Добронец, О. А. Попова, Е. В. Головчанская // XI Межд. конф. ФАМЭБ - Красноярск, 2012.
12) Добронец Б. С. Численные операции над случайными величинами и их приложения // Б. С. Добронец, О. А. Попова // Журнал Сиб. фед. ун-та, Математика и физика. - 2011 - 4(2) - С. 229-239.
13) Добронец Б. С. Элементы численного вероятностного анализа / Б. С. Добронец, О. А. Попова.
14) Dobronets B. S. Software implementation of numerical operations on random variables / B. S. Dobronets, A. M. Krantsevich, N. M. Krantsevich // Журнал Сиб. фед. ун-та, Математика и физика. - 2013. - 6(2) - С. 168-173.
15) Дыбов А. М. Особенности оценки инвестиционных проектов с учетом факторов риска и неопределенности / А. М. Дыбов // Вестник Удмуртского университета - 2010 / Экономика и право. — Ижевск, УдГУ — 2010, Вып. 2, с. 7-14.

16) Ермаков С. М. Метод Монте-Карло в вычислительной математике / С. М. Ермаков. - СПб: 2009. - 192 с.
17) Ибрагимов В. А. Элементы нечеткой математики / В. А. Ибрагимов. - Баку: Азер. гос. нефт. академ., 2010. - 392 с.
18) Иванюк В. А. Моделирование сложных экономических систем на основе методов искусственного интеллекта / В. А. Иванюк // Успехи современного естествознания - 2011. - №1 - С. 151-152.
19) Квасов Б. И. Методы изогеометрической аппроксимации сплайнами : монография / Б. И. Квасов. - Москва : Физматлит [Физико-математическая литература], 2006. - 360 с. : ил. - Список лит.: с.348-356.
20) Корчикова Д. И. Арифметики и численный вероятностный анализ неопределенных данных / Д. И. Корчикова. – Новосибирск: Новосиб. гос. ун-т, 2015.
21) Корчикова Д. И. Полиграммы для представления случайных данных / Д. И. Корчикова. – Новосибирск: Новосиб. гос. ун-т, 2015.
22) Лапко А. В. Непараметрические системы обработки неоднородной информации [Электронный ресурс] / А. В. Лапко, В. А. Лапко ; Сиб. федер. ун-т, Ин-т космич. и информ. технологий. - Электрон. текстовые дан. (PDF, 16,18 Мб). - Новосибирск : Наука, 2007. - 174 с. - Библиогр.: с. 167-171. - ISBN 978-5-02-023180-1 : Б. ц.
23) Лукашов А. В. Метод Монте-Карло для финансовых аналитиков: краткий путеводитель / А. В. Лукашов // Управление корпоративными финансами. - 2007 - 01(19). - С. 22-39.
24) Лю Б. Теория и практика неопределенного программирования / Б. Лю; Пер. с англ. - М.: БИНОМ. Лаборатория знаний, 2005. - 416 с.: ил. - (Адаптивные и интеллектуальные системы). УДК 517.11+519.92, ББК 22.18
25) Петрушин В. Н. Интервальная арифметика: эмпирико-статистический подход к оценке результатов действий / В. Н. Петрушин, Е. В. Никульчев.
26) Попова О. А. Гистограммы второго порядка для численного моделирования в задачах с информационной неопределенностью / О. А. Попова // Известия Южн. фед. ун-та, Технические науки. - С. 6-14.
27) Соболь И. М. Численные методы Монте-Карло / И. М. Соболь. - М.: Наука, 1973. - 312 с.: с илл.
28) Тарасенко Ф. П. Непараметрическая статистика: монография / Ф. П. Тарасенко. - Томск: ТГУ, 1976. - 294 с.
29) Третьяков Н. П. Имитационное моделирование методом Монте-Карло и развитие методологии прогнозных оценок макроэкономических показателей / Н. П. Третьяков, Е. О. Щербакова // Интернет-журнал "Технологии техносферной безопасности" - 2009. - №6 - С. 1-16.
30) Тутубалин В. Н. Границы применимости (вероятностно-статистические методы и их возможности) / В. Н.Тутубалин. - М.: Знание, 1977. - 64 с.
31) Углев В. А. Выбор между методом Монте-Карло и гистограммной арифметикой при реализации моделей с элементами случайности / В. А. Углев // ИММОД. - 2013.

32) Черепанов Е. В. Математическое моделирование неоднородных совокупностей экономических данных. Монография / Московский государственный университет экономики, статистики и информатики (МЭСИ). - 2013. - С. 229.
33) Якобсен Х. Ш. Представление и расчет экономических неопределенностей: интервалы, нечеткие числа и вероятности / Х. Ш. Якобсен // Департамент производства техники, колледж в Копенгагене - инженерия, DK-2750 Ballerup, Дания. - 2000.
34) Ярлыкова Л. К., Медведев А. В. О непараметрических алгоритмах сглаживания при моделировании лавинообразных процессов /Актуальные проблемы авиации и космонавтики. 2013. Т. 1. № 9. С. 345-347.

ПРИЛОЖЕНИЕ А

Плакаты презентации

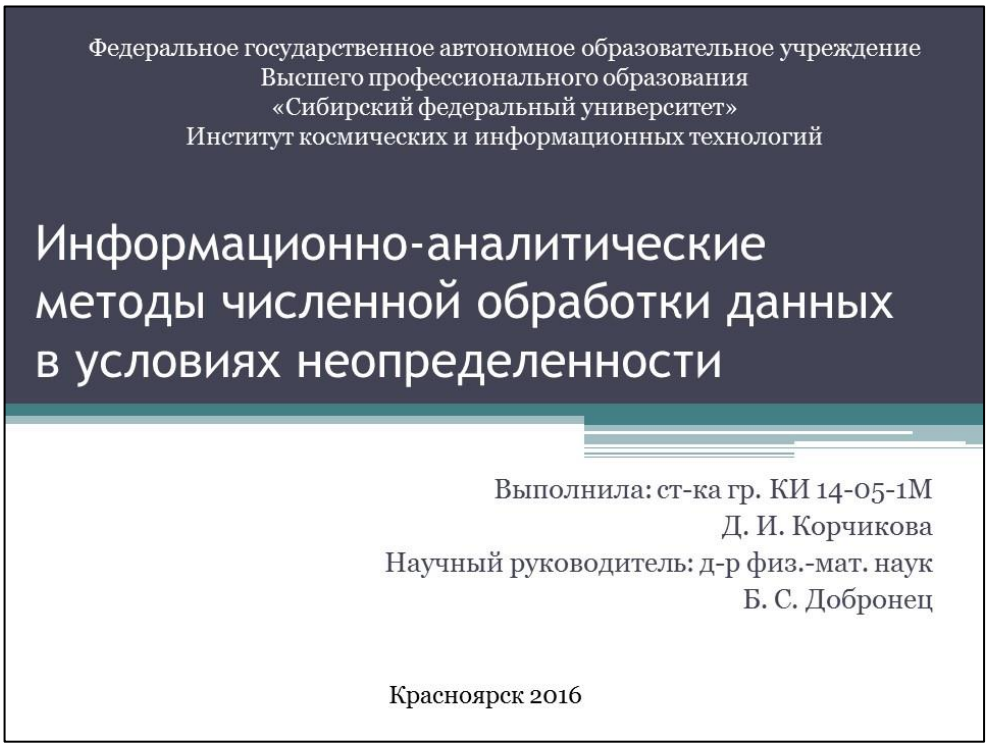

Рисунок А.1 – Титульный лист

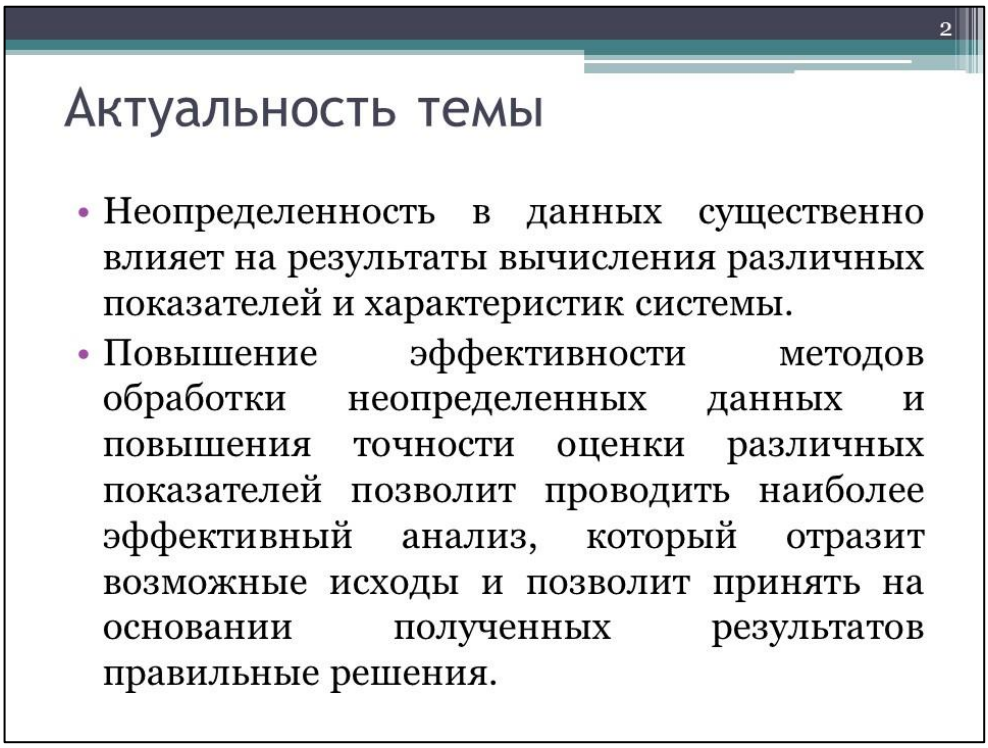

Рисунок А.2 – Актуальность темы

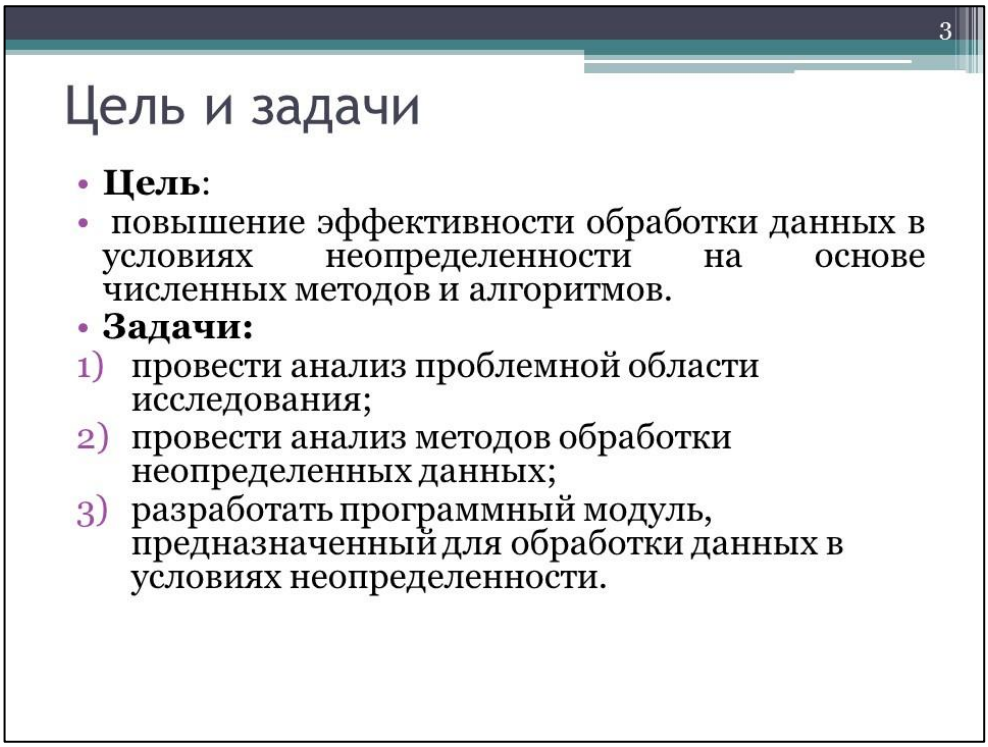

Рисунок А.3 – Цель и задачи

Рисунок А.4 – Виды неопределенностей

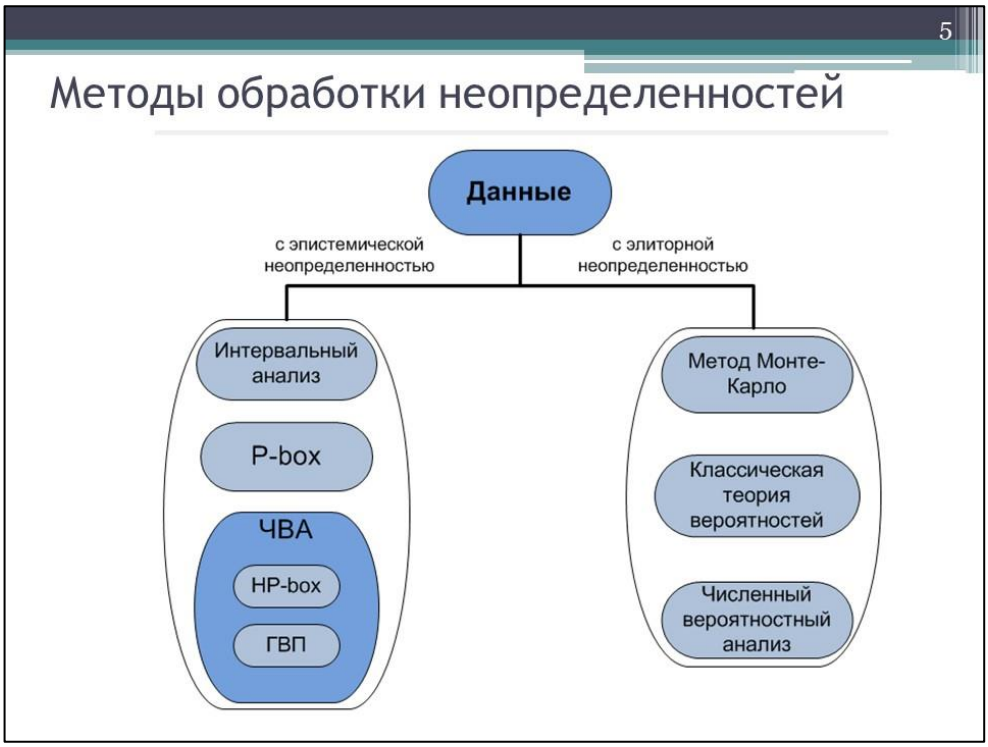

Рисунок А.5 – Методы обработки неопределенностей

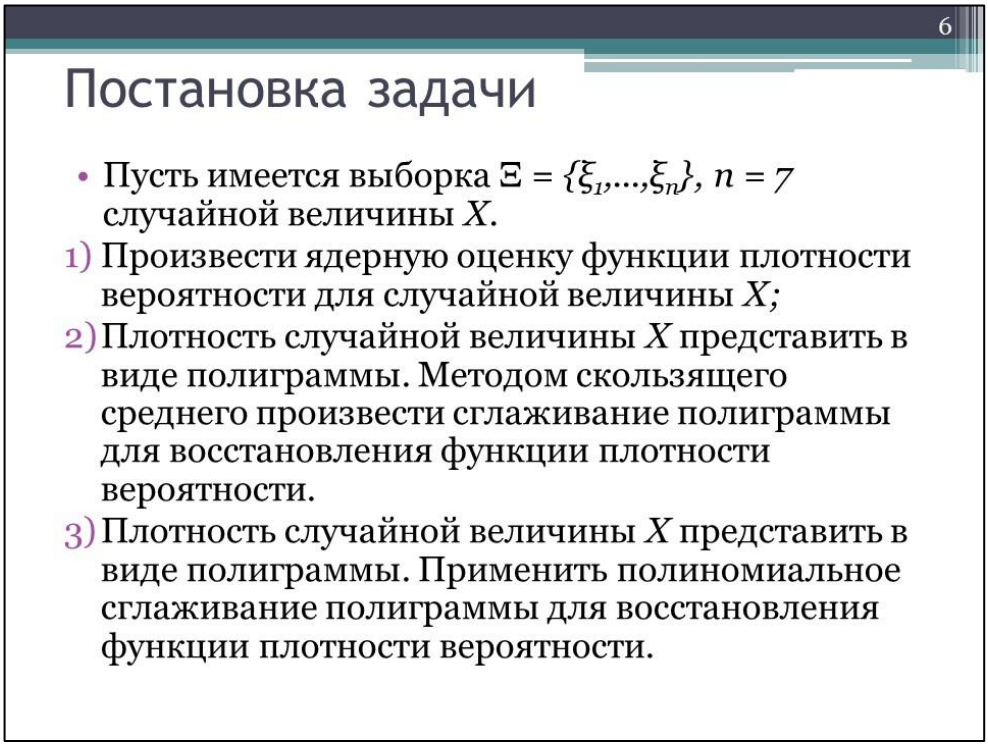

Рисунок А.6 – Постановка ззадачи

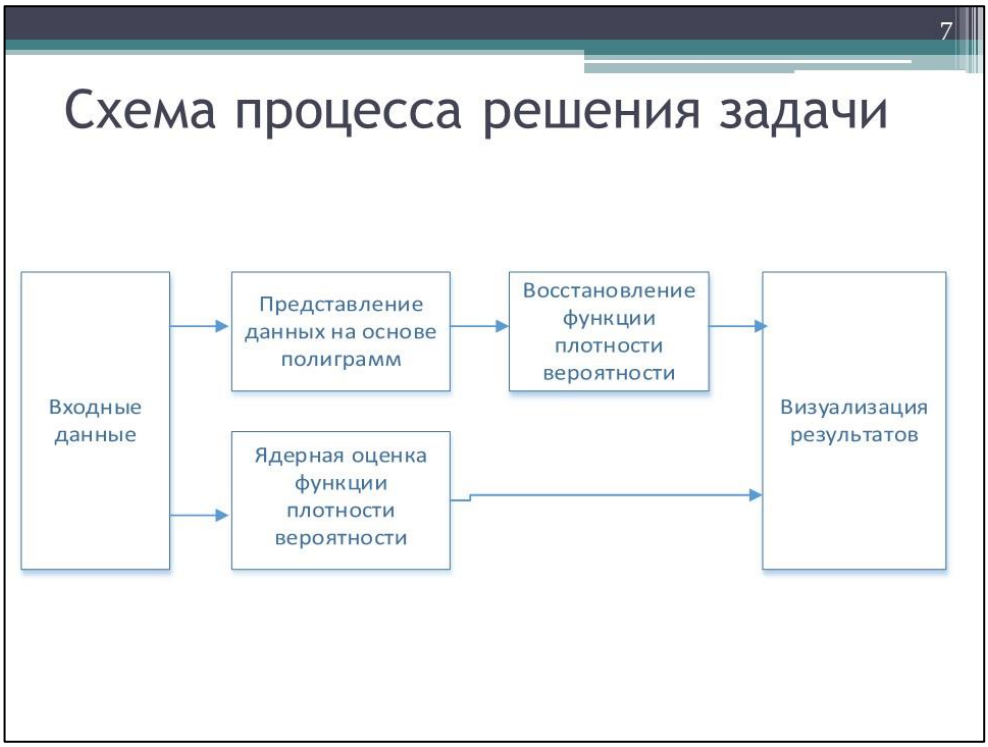

Рисунок А.7 – Схема решения задачи

Рисунок А.8 – Результат ядерного восстановления функции плотности вероятности

Рисунок А.9 – Восстановление функции плотности вероятности на основе полиграмм

Рисунок А.10 – Восстановление функции плотности вероятности на основе полиграмм

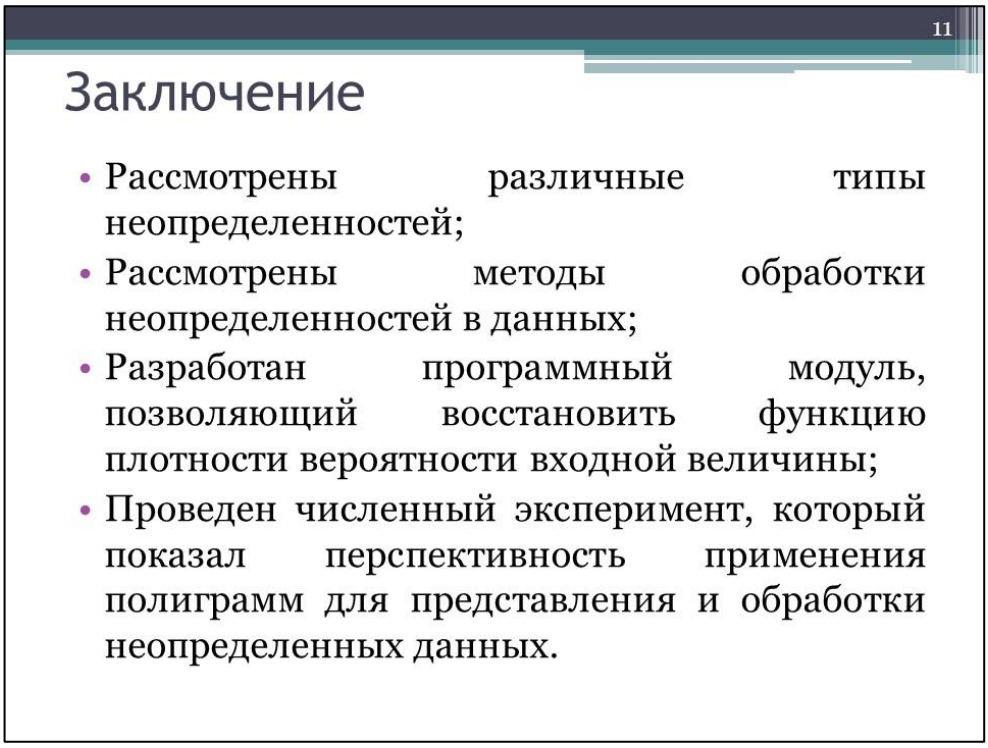

Рисунок А.11 - Заключение

Рисунок А.12 - Публикации

Рисунок А.13 - Публикации