МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
«Новосибирский национальный исследовательский государственный университет»
(Новосибирский государственный университет, НГУ)

Физический факультет
Кафедра физики элементарных частиц ФФ

Рабочая программа дисциплины

МЕТОДЫ АНАЛИЗА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

направление подготовки: **03.04.01 Прикладные математика и физика**
Профиль: **Прикладные математика и физика. Информационные процессы и системы**

Форма обучения
Очная

Семестр	**Общий объем**	**Виды учебных занятий (в часах)**				**Промежуточная аттестация (в часах)**				
		Контактная работа обучающихся с преподавателем			Самостоятельная работа, не включая период сессии	Самостоятельная подготовка к промежуточной аттестации	Контактная работа обучающихся с преподавателем			
		Лекции	Практические занятия	Лабораторные занятия			Консультации	Зачет	Дифференцированный зачет	Экзамен
1	2	3	4	5	6	7	8	9	10	11
2	108	32	32		22	18	2			2
Всего 108 часов / 3 зачётные единицы, из них: - контактная работа 68 часов										
Компетенции ПК-1										

Руководитель программы
д.ф.-м.н. И. Б. Логашенко

Новосибирск, 2024

Содержание

1. Перечень планируемых результатов обучения по дисциплине, соотнесённых с планируемыми результатами освоения образовательной программы.

Целью освоения дисциплины «Методы анализа экспериментальных данных» является практическое ознакомление студентов с современными методами анализа результатов измерений, получаемых в физических экспериментах. Для достижения поставленной цели ставятся следующие задачи:

1) изучение практического применения методов теории вероятностей, математической статистики, методов Монте-Карло для анализа данных экспериментов, оценки погрешностей, классификации данных;
2) изучение специализированного программного обеспечения Geant4 и Root.

В современных экспериментах зачастую изучаются слабые эффект или редкие процессы. Это значительно усложняет анализ данных, т.к. исследуемый эффект может сильно искажаться наличием фона, наличием корреляций и т.п. Правильная оценка величины эффекта, оценка погрешностей и достоверности результата в таких случаях требует глубокого понимания математической статистики и применения специальных подходов.

Бурное развитие информационных технологий в последнее время привело к взрывному росту количества собираемой и анализируемой информации практически во всех направлениях жизнедеятельности человека. В связи с этим возникло целое направление исследований, интеллектуальный анализ данных, задачей которого является разработка методов и алгоритмов выявления скрытых закономерностей в больших массивах данных.

В ходе современных физических экспериментах часто возникают большие объемы информации. Характерный объем данных, накопленный в типичном эксперименте, обычно составляет 10-100 терабайт. На Большом Адронном Коллайдере каждый год планируется накапливать более 10 петабайт (10000 терабайт) в течение 10 лет. Такие объемы данных, а также их возросшая сложность, связанная с большим количеством регистрирующих систем, требуют использования автоматизированных методов анализа данных.

Обсуждаемые методы анализа данных применяются не только при анализе данных физического эксперимента, но и в других областях знаний, таких как экономика, биология и др.

Лекционная часть курса состоит из трех разделов. Первый раздел посвящен повторению и углублению знаний, полученных в курсах теории вероятностей и математической статистики. Подробно обсуждаются практические подходы к оценке параметров с помощью методов максимального правдоподобия и наименьших квадратов как в типичных случаях, так и в ситуациях, усложненных наличием фона, отсутствием достаточной экспериментальной статистики, наличием корреляций между измеряемыми величинами и т.п. Обсуждается применение байесовских методов при анализе данных физического эксперимента. Второй раздел курса посвящен применению методов интеллектуального анализа данных. Подробно обсуждаются такие методики, как нейронные сети, усиленные деревья принятия решения, линейный дискриминантный анализ и т.п., и их применение для решения задач классификации (отделения сигнала от фона), кластеризации, регрессии. В третьем разделе курса рассматриваются отдельные задачи, часто возникающие при анализе экспериментальных данных: задача реконструкции треков частиц, задача деконволюции, задача калибровки измерительной системы.

В практической части курса студенты получают возможность на практике применить полученные знания. Первые несколько занятий посвящены ознакомлению с программным обеспечением, используемым для моделирования эксперимента и для анализа данных. В дальнейшем студенты должны выполнить несколько обязательных заданий, отражающих

основные этапы анализа данных любого эксперимента: использование методов Монте-Карло, моделирование физических процессов, оценка параметров распределений, отделение сигнала от фона и т.п.

Дисциплина нацелена на формирование у обучающегося профессиональной компетенции:

Результаты освоения образовательной программы (компетенции)	Индикаторы	Результаты обучения по дисциплине
ПК-1 Способность осваивать и применять специализированные знания в области физико-математических и (или) естественных наук в своей профессиональной деятельности.	**ПК 1.1** Применяет специализированные знания естественных и (или) физико-математических наук при решении поставленных задач в специализированной области своей профессиональной деятельности. **ПК 1.2** Применяет классические и новые знания при решении поставленных задач в специализированной области своей профессиональной деятельности.	**Знать** теоретические положения математической статистики и теории вероятностей, лежащие в основе изучаемых методов анализа данных: методы оценки параметров распределений, методы максимального правдоподобия и наименьших квадратов, построение критериев согласия, основы теории проверки гипотез, основы теории принятия решений, байесовский подход к оценке вероятностей; знать основные алгоритмы многомерного анализа данных, в частности, методы построения функций правдоподобия, нейронных сетей, деревьев принятия решений. **Уметь** использовать методы Монте-Карло для моделирования эксперимента и оценки погрешностей; оценивать параметры распределений при наличии корреляций и фона; применять методы максимального правдоподобия и наименьших квадратов; применять методы многомерного анализа данных; применять комплексные алгоритмы при анализе больших массивов данных; решать типичные задачи, возникающие при анализе данных современного физического эксперимента. **Владеть** программным инструментарием для моделирования и анализа данных физического эксперимента: пакетами ROOT и GEANT4.

2. Место дисциплины в структуре образовательной программы.

Курс **«Методы анализа экспериментальных данных»** относится к дисциплинам по выбору.

Для освоения материала необходимо предшествующее успешное освоение курсов основ математической статистики и теории вероятностей, математического анализа, общей физики. Для успешного выполнения практических заданий необходимо знание языков программирования C++ или Python. В свою очередь, учебный курс «Методы анализа экспериментальных данных» предоставляет студентам теоретические знания и практические навыки, необходимые при прохождении преддипломной практики.

3. Трудоёмкость дисциплины в зачётных единицах с указанием количества академических часов, выделенных на контактную работу обучающегося с преподавателем (по видам учебных занятий) и на самостоятельную работу.

Трудоемкость дисциплины – 2 з.е. (72 час)
Форма промежуточной аттестации: 1 семестр – экзамен

№	Вид деятельности	Семестр
		1
1	Лекции, час	32
2	Практические занятия, час	32
3	Лабораторные занятия, час	-
4	Занятия в контактной форме, час, из них	68
5	из них аудиторных занятий, час	64
6	в электронной форме, час	-
7	консультаций, час.	2
8	промежуточная аттестация, час	2
9	Самостоятельная работа, час	40
10	Всего, час	72

4. Содержание дисциплины, структурированное по темам (разделам) с указанием отведённого на них количества академических часов и видов учебных занятий.

Программа и основное содержание лекций (32 часа)

Наименование темы и их содержание	Объем, час
1. Случайные величины. Дискретные и непрерывные распределения. Параметры распределений: среднее значение, дисперсия, моменты. Ковариационная матрица, коэффициент корреляции. Преобразование распределения при замене переменных. Основные распределения и их параметры: биномиальное, Пуассона, равномерное, нормальное, $\chi 2$. Центральная предельная теорема.	4
2.Метод Монте-Карло. Интегрирование методом Монте-Карло. Алгоритмы генерации случайных чисел: метод Неймана, метод трансформации, комбинированный. Алгоритм генерации нормально-распределенной величины. Алгоритм Метрополиса	2
3.Оценка параметров распределений по ограниченной выборке. Точечные и интервальные оценки. Свойства оценок: состоятельность, смещение, эффективность, робастность (устойчивость). Понятие информации Фишера и неравенство Рао-Крамера. Способы построения оценок, метод моментов. Способы построения несмещенной оценки, робастной оценки.	2
4.Метод максимального правдоподобия. Оценка погрешностей и построение доверительных интервалов в методе максимального правдоподобия. Примеры использования метода максимального правдоподобия для аппроксимации гистограммы, определения времени жизни, оценки дисперсии.	2
5.Метод наименьших квадратов. Оценка погрешностей в методе наименьших квадратов. Метод наименьших квадратов в линейном приближении. Пример использования метода наименьших квадратов для аппроксимации гистограмм	2
6.Критерий согласия и способы его построения. Критерий $\chi 2$. Оценка качества аппроксимации в методе максимального правдоподобия. Другие критерии согласия: проверка последовательностей, критерий Колмогорова-Смирнова	2
7.Байесовский подход к оценке вероятностей. Теорема Байеса. Формулировка теоремы Байеса для непрерывных распределений. Применение теоремы Байеса для оценки погрешностей. Связь теоремы Байеса и метода максимального правдоподобия. Примеры применения теоремы Байеса: определение эффективности, оценка верхнего предела при близости измеренного значения к границе интервала возможных значений, оценка уровня сигнала при наличии фона. Понятие Байесовских сетей	2
8.Нейронные сети. Однослойный и многослойный перцептрон. Обучение перцептрона, алгоритм обратного распространения ошибок. Глобальные методы оптимизации. Радиально-базисные сети. Задача кластеризации и сеть Кохонена. Применение нейронных сетей для классификации данных	4
9.Задача разделения сигнала и фона (задача проверки гипотез). Критерий разделения, мощность и значимость критерия. Методы сравнения критериев. Простые гипотезы, лемма Неймана-Пирсона и наилучший критерий разделения. Критерий разделения в случае сложных гипотез. Практические методы	2

построения критериев разделения: факторизация функции правдоподобия; линейный дискриминантный анализ Фишера; нейронные сети; усиленные деревья принятия решений; методы, основанные на подсчете числа событий. Метод главных компонент.	
10.Практические методы построения критериев разделения: факторизация функции правдоподобия; линейный дискриминантный анализ Фишера; нейронные сети; деревья принятия решений; методы, основанные на подсчете числа событий. Метод главных компонент	4
11.Задача обратной свертки (unfolding). Постановка задачи. Методы получения результатов без обратной свертки. Прямое решение задачи. Регуляризация. Регуляризация Тихонова. Метод максимальной энтропии.	2
12.Восстановление траекторий заряженных частиц. Задачи распознавания трека и определения параметров трека. Алгоритм гистограммирования и преобразование Хью (Hough). Рекурсивный метод наименьших квадратов (фильтр Калмана).	2
13.Задача калибровки измерительной системы. Простой алгоритм калибровки с помощью оценки среднего отклика. Использование многопараметрических методов оптимизации для калибровки системы по большому массиву данных.	2

Программа практических занятий (32 часа)

1. Знакомство с программным пакетом ROOT (ПО для обработки данных). (4 часа)
2. Выполнение задания №1: методы Монте-Карло и параметры распределений. (4 часа)
3. Знакомство с программным пакетом GEANT (ПО для моделирования взаимодействия частиц с веществом). (4 часа)
4. Выполнение задания №2: моделирование отклика простого детектора элементарных частиц. (4 часа)
5. Выполнение задания №3: анализ результатов, полученных в ходе выполнение задания №2, с использованием метода максимального правдоподобия. (4 часа)
6. Знакомство с пакетом TMVA (ПО для многомерного статистического анализа данных). (4 часа)
7. Выполнение задания №4: анализ результатов, полученных в ходе выполнение задания №2, с использованием нейронных сетей. (4 часа)
8. Выполнение задания №5: анализ результатов, полученных в ходе выполнение задания №2, с использованием различных алгоритмов многомерного анализа данных. (4 часа)

Самостоятельная работа студентов (40 часов)

Перечень занятий на СРС	Объем, час
Подготовка к практическим занятиям.	11
Изучение теоретического материала, не освещаемого на лекциях	11
Подготовка к экзамену	18

5. Перечень учебной литературы.

1. Лотов, Владимир ИвановичТеория вероятностей и математическая статистика : курс лекций / В.И. Лотов ; М-во образования и науки РФ, Новосиб. гос. ун-т, Фак. информ.

технологий2-е изд., испр. и допНовосибирск : Редакционно-издательский центр НГУ, 2011127 с. : ил. ; 20 см. (37 экз)

2. Д. Худсон. Статистика для физиков. М.: Мир, 1970.(21 экз)
3. Чистяков, Владимир ПавловичКурс теории вероятностей : [Учебник для втузов] / В.П. Чистяков3-е изд., испр.М. : Наука, 1987240 с. : ил. ; 21 см. (39 экз)
4. Боровков, Александр Алексеевич (1931-) Математическая статистика : учебное пособие : [для студентов Мех.-мат. фак. НГУ : в 2 ч.] / А.А. Боровков ; М-во высш. и сред. спец. образования РСФСР, Новосиб. гос. ун-т им. Ленин. Комсомола Новосибирск : Редакционно-издательский центр НГУ, 1983-1984. 20 см. (Ч1- 322 экз, Ч2-322 экз)

6. Перечень учебно-методических материалов по самостоятельной работе обучающихся.

Самостоятельная работа студентов поддерживается следующими учебными пособиями:

5. В. И. Лотов. Теория вероятностей и математическая статистика. Новосибирск: НГУ, 2006.
6. Д. Худсон. Статистика для физиков. М.: Мир, 1970.
7. Чистяков В.П. Курс теории вероятностей. М.: Наука, 1982.
8. Боровков Л.Л. Математическая статистика. Оценка параметров. Проверка гипотез. Новосибирск, 1984.

7. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины.

Для освоения дисциплины используются следующие ресурсы:
- электронная информационно-образовательная среда НГУ (ЭИОС);
- образовательные интернет-порталы;
- информационно-телекоммуникационная сеть Интернет.

Интернет-ресурсы:

1. Описание пакета ROOT. http://root.cern.ch/drupal/content/users-guide
2. Описание пакета TMVA. http://tmva.sourceforge.net/docu/TMVAUsersGuide.pdf
3. Описание пакета Geant4. http://geant4.cern.ch/support/userdocuments.shtml

7.1 Современные профессиональные базы данных

Не используются.

7.2. Информационные справочные системы

Не используются.

8. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине.

Для обеспечения реализации дисциплины используется стандартный комплект программного обеспечения (ПО), включающий регулярно обновляемое лицензионное ПО Windows и MS Office.

Использование специализированного программного обеспечения для изучения дисциплины не требуется.

9. Материально-техническая база, необходимая для осуществления образовательного процесса по дисциплине.

Для реализации дисциплины используются специальные помещения:

1. Учебные аудитории для проведения занятий лекционного типа, практических занятий, курсового проектирования (выполнения курсовых работ), групповых и индивидуальных консультаций, текущего контроля, промежуточной и итоговой аттестации.

2. Помещения для самостоятельной работы обучающихся.

Учебные аудитории укомплектованы специализированной мебелью и техническими средствами обучения, служащими для представления учебной информации большой аудитории.

Помещения для самостоятельной работы обучающихся оснащены компьютерной техникой с возможностью подключения к сети "Интернет" и обеспечением доступа в электронную информационно-образовательную среду НГУ.

Материально-техническое обеспечение образовательного процесса по дисциплине для обучающихся из числа лиц с ограниченными возможностями здоровья осуществляется согласно «Порядку организации и осуществления образовательной деятельности по образовательным программам для инвалидов и лиц с ограниченными возможностями здоровья в Новосибирском государственном университете».

10. Оценочные средства для проведения текущего контроля и промежуточной аттестации по дисциплине.

10.1 Порядок проведения текущего контроля и промежуточной аттестации по дисциплине

Текущий контроль

Текущий контроль осуществляется в ходе семестра путем проверки решений студентами заданий для самостоятельного решения.

Промежуточная аттестация

Освоение компетенций оценивается согласно шкале оценки уровня сформированности компетенции. Положительная оценка по дисциплине выставляется в том случае, если заявленная компетенция ПК-1 сформирована не ниже порогового уровня в части, относящейся к формированию способности использовать специализированные знания в области методов анализа экспериментальных данных в профессиональной деятельности.

Окончательная оценка работы студента в течение семестра происходит на экзамене. Экзамен проводится в конце семестра в экзаменационную сессию по билетам в устной форме. Вопросы билета подбираются таким образом, чтобы проверить уровень сформированности компетенции ПК-1.

Вывод об уровне сформированности компетенций принимается преподавателем. Положительная оценка ставится, когда все компетенции освоены не ниже порогового уровня. Оценки «отлично», «хорошо», «удовлетворительно» означают успешное прохождение промежуточной аттестации.

Соответствие индикаторов и результатов освоения дисциплины

Таблица 10.1

Код компетенции	Индикатор	Результат обучения по дисциплине	Оценочные средства
ПК-1	**ПК 1.1** Применяет специализированные знания естественных и (или) физико-математических наук при решении поставленных задач в специализированной области своей профессиональной деятельности.	**Знать** теоретические положения математической статистики и теории вероятностей, лежащие в основе изучаемых методов анализа данных: методы оценки параметров распределений, методы максимального правдоподобия и наименьших квадратов, построение критериев согласия, основы теории проверки гипотез, основы теории принятия решений, байесовский подход к оценке вероятностей; знать основные алгоритмы многомерного анализа данных, в частности, методы построения функций правдоподобия, нейронных сетей, деревьев принятия решений.	Решение заданий, экзамен
	ПК 1.2 Применяет классические и новые знания при решении поставленных задач в специализированной области своей профессиональной деятельности.	**Уметь** использовать методы Монте-Карло для моделирования эксперимента и оценки погрешностей; оценивать параметры распределений при наличии корреляций и фона; применять методы максимального правдоподобия и наименьших квадратов; применять методы многомерного анализа данных; применять комплексные алгоритмы при анализе больших массивов данных; решать типичные задачи, возникающие при анализе данных современного	Решение заданий, экзамен

		физического эксперимента. **Владеть** программным инструментарием для моделирования и анализа данных физического эксперимента: пакетами ROOT и GEANT4.	

Таблица 10.2

Критерии оценивания результатов обучения	Шкала оценивания
Решение заданий: – задание решено правильно, – работа оформлена аккуратно, четкие рисунки и чертежи, – осмысленность, логичность и аргументированность изложения материала, – точность и корректность применения терминов и понятий. «Сдать задачу» означает объяснение хода её решения и, при необходимости, ответы на дополнительные вопросы преподавателя, имеющие принципиальное значение для данной дисциплины. Свободно и аргументированно отвечает на дополнительные вопросы. В ответах на вопросы преподавателя обучающийся мог допустить непринципиальные неточности. **Экзамен:** – самостоятельность, осмысленность, структурированность, логичность и аргументированность изложения материала, отсутствие затруднений в объяснении процессов и явлений, а также при формулировке собственных суждений, – точность и корректность применения терминов и понятий, – наличие исчерпывающих ответов на дополнительные вопросы. При изложении ответа на вопрос(ы) преподавателя обучающийся мог допустить непринципиальные неточности.	*Отлично*
Решение заданий: – задание решено правильно, – работа оформлена аккуратно, четкие рисунки и чертежи, – осмысленность, логичность и аргументированность изложения материала, наличие затруднений в формулировке собственных суждений, – точность и корректность применения терминов и понятий, при наличии незначительных ошибок. «Сдать задачу» означает объяснение хода её решения и, при необходимости, ответы на дополнительные вопросы преподавателя, имеющие принципиальное значение для данной дисциплины. Отвечает на дополнительные вопросы. В ответах на вопросы преподавателя обучающийся мог допустить непринципиальные неточности. **Экзамен:**	*Хорошо*

– самостоятельность, осмысленность, структурированность, логичность и аргументированность изложения материала, наличие затруднений в объяснении отдельных процессов и явления, а также при формулировке собственных суждений, – точность и корректность применения терминов и понятий при наличии незначительных ошибок, – наличие полных ответов на дополнительные вопросы с возможным присутствием ошибок.	
Решение заданий: – задание решено правильно, - работа оформлена неаккуратно – неосознанность и неосновательность выбранных методов анализа, – нет осмысленности в изложении материала, наличие ошибок в логике и аргументации, – корректность применения терминов и понятий, при наличии незначительных ошибок. «Сдать задачу» означает объяснение хода её решения и, при необходимости, ответы на дополнительные вопросы преподавателя, имеющие принципиальное значение для данной дисциплины. При ответах на вопросы допускает ошибки **Экзамен:** – теоретический и фактический материал в слабой степени подкреплен ссылками на научную литературу и источники, – частичное понимание и неполное изложение причинно-следственных связей, – самостоятельность и осмысленность в изложении материала, наличие ошибок в логике и аргументации, в объяснении процессов и явлений, а также затруднений при формулировке собственных суждений, – корректность применения терминов и понятий, при наличии незначительных ошибок, – наличие неполных и/или содержащих существенные ошибки ответов на дополнительные вопросы.	*Удовлетворительно*
Решение заданий: – задание решено неправильно, – компилятивное, неосмысленное, нелогичное и неаргументированное изложение материала, – грубые ошибки в применении терминов и понятий, «Сдать задачу» означает объяснение хода её решения и, при необходимости, ответы на дополнительные вопросы преподавателя, имеющие принципиальное значение для данной дисциплины. На дополнительные вопросы не отвечает. **Экзамен:** – фрагментарное и недостаточное представление теоретического и фактического материала, не подкрепленное ссылками на научную литературу и источники, – непонимание причинно-следственных связей, – отсутствие осмысленности, структурированности, логичности и аргументированности в изложении материала, – грубые ошибки в применении терминов и понятий, – отсутствие ответов на дополнительные вопросы.	*Неудовлетво-рительно*

10.3 Типовые контрольные задания и материалы, необходимые для оценки результатов обучения

Список заданий для самостоятельного решения

Toy Monte-Carlo

Постановка задачи

Часто требуется оценить статистическую и систематическую ошибки выбранной процедуры анализа данных. В простых случаях, например, когда все распределения являются нормальными и используются хорошо известные статистические приемы, существуют аналитические формулы для подобных оценок. Однако в более сложных случаях необходимо использовать моделирование методом Монте-Карло.

Одна из наиболее распространенных процедур такого моделирования – toy Monte-Carlo, «простое» моделирование. В рамках этой процедуры исходные распределения для сырых данных считаются известными. По заданным распределениям генерируется массив сырых данных, которые затем проходят выбранную процедуру анализа и ее результаты сравниваются с параметрами, заложенными в моделирование. Как правило, генерируется множество независимых массивов данных, и из распределения полученных результатов извлекается информация о статистической ошибке и смещении процедуры.

Более сложными процедурами являются полное или параметризованное моделирование методом Монте-Карло, в котором сырые данных генерируются с использованием математических моделей, учитывающих лежащие в основе физические процессы. Пакет Geant4 предназначен для моделирования именно такого типа. Как правило, такие процедуры требуют гораздо больше вычислительных ресурсов, и они используются для оценки влияния параметров детектора на результаты анализа. Для оценки статистических свойств анализа достаточно toy Monte-Carlo.

В задании требуется провести небольшое исследование с помощью простого моделирования методом Монте-Карло.

1. Необходимо написать генератор случайных чисел, распределенных согласно плотности, заданной преподавателем. Следует использовать метод обратной трансформации или комбинированный метод
2. Продемонстрировать свойства двух различных оценок положения распределения: выборочного среднего и выборочной медианы. Пусть в одном эксперименте генерируется n_{data} (типичное значение 100) величин, распределенных согласно заданному распределению. По полученной выборке определяются значения выборочного среднего и выборочной медианы. Для того, чтобы изучить свойства этих двух оценок, повторим процедуру для n_{exp} (типичное значение 1000) экспериментов. Из распределения полученных значений оценок для различных экспериментов необходимо извлечь следующую информацию.
    1) Оценить мат. ожидание заданного распределения, определить точность этой оценки и сравнить полученное значение с расчетным.
    2) Оценить асимптотическую эффективность выборочной медианы в сравнении с выборочным средним. Для этого нужно провести моделирование для разных значений n_{data} и определить, чему асимптотически равно отношение дисперсий двух оценок.
    3) Продемонстрировать робастность выборочной медианы. Для этого модифицируйте генератор случайных чисел, добавив в него небольшую вероятность «выбросов». Покажите, как смещение и дисперсия двух оценок зависят от доли выбросов.
3. Продемонстрируйте предсказание центральной предельной теоремы. Для этого сгенерируйте распределение нормированной суммы 2, 5 и 100 случайных величин, подчиняющихся заданному распределению. Сходится ли полученное распределение к нормальному?

Дополнительные вопросы

1. При каком уровне выбросов стоит использовать выборочную медиану вместо выборочного среднего?

2. Сколько нужно сложить случайных величин, чтобы 95% доверительный интервал, определенный в предположении нормального распределения, был правильным с заданной точностью?
3. Замените свое распределение распределением Коши. Работает ли в этом случае центральная предельная теорема?
4. Вычислите доверительный интервал для параметров `Mean` (выборочное среднее) и `RMS` (выборочная дисперсия) для распределения h3 на Рисунке 22, зная свойства исходного распределения и постановку эксперимента. Попали ли значения этих параметров, полученные в вашем численном эксперименте, в полученные интервалы?

Моделирование детектора

Постановка задачи

Требуется смоделировать отклик детектора элементарных частиц. Детектор представляет собой простой многослойный калориметр, либо куб, либо цилиндр, разделенный на N слоев ($N = 10$). В качестве активного вещества используются широко распространенные в калориметрии материалы: кристаллы NaI, CsI, BGO, LSO, сжиженные благородные газы Ar, Kr, Xe. Каждый студент использует индивидуальный активный материал, предложенный преподавателем.
Регистрируемые частицы влетают в калориметр вдоль оси детектора. Возможны 3 типа частиц: электроны e, мюоны μ и гамма-кванты γ. Начальная энергия всех частиц одинакова, тип частицы должен выбираться случайным образом для каждого события.
Программа моделирования должна быть написана таким образом, что следующие параметры могут быть легко изменены без значительного переписывания кода: размеры детектора, количество слоев, активное вещество, энергия влетающих частиц.
В качестве результата необходимо провести моделирование 10000 событий. Для каждого события необходимо сохранить: тип частицы, полное энерговыделение в калориметре, энерговыделение в каждом слое. Результаты моделирования в форме дерева ROOT необходимо сохранить, т.к. эти данные будут использованы в последующих заданиях.

Основные и дополнительные вопросы

1. Объясните на качественном уровне отличия между полученными распределениями для разных типов частиц.
2. Попробуйте изменить список возможных взаимодействий для какого либо типа частиц в `ExN03PhysicsList` и посмотрите, как изменятся полученные распределения. Объясните результаты на качественном уровне.

Метод максимального правдоподобия

Постановка задачи

Пусть детектор регистрирует частицы двух типов и для каждой частицы измеряется только полное энерговыделение. Необходимо с помощью метода максимального правдоподобия оценить число событий каждого типа, анализируя измеренное распределение полного энерговыделения.
Из массива экспериментальных данных, сгенерированного в рамках задания 0, необходимо сформировать гистограмму полного энерговыделения в калориметре для событий двух типов (например, e и μ). Данная гистограмма представляет собой модель экспериментальных данных. В экспериментальных данных перемешаны энерговыделения двух типов событий. Для того, чтобы оценить число событий каждого типа, необходимо подогнать данную гистограмму функцией

$$f(x) = n_e f_e(x) + n_\mu f_\mu(x)$$

где x – энерговыделение, n_k – число событий k-того типа, $f_k(x)$ – плотность распределения энерговыделения для событий k-того типа. Подгонку необходимо реализовать методом максимального правдоподобия. По результатам подгонки необходимо получить оценку числа событий каждого типа и точность этой оценки.

Основные и дополнительные вопросы

1. Сравните полученные оценки с их истинным значением (поскольку данные получены с помощью моделирования, точно известно, сколько событий какого типа). Согласуются ли полученные оценки с точным значением?
2. Как можно оценить смещение полученной оценки?
3. Какие выводы можно сделать на основе формы доверительного эллипса?
4. Пусть в результате эксперимента нужно получить не число событий каждого типа n_e и n_μ, а их сумму $(n_e + n_\mu)$ (или разность $(n_e - n_\mu)$, отношение n_e/n_μ и т.п.). Как вычислить значение суммы (отношения, разности и т.п.) и оценить ее ошибку, получив оценку n_e и n_μ с помощью метода максимального правдоподобия? Указание: необходимо учесть корреляцию между n_e и n_μ.
5. Получите оценку суммы (разности, отношения и т.п.) из вопроса 4 напрямую с помощью метода максимального правдоподобия. Совпала ли полученная оценка с расчетным значением, полученным в вопросе 4? Указание: для этого необходимо переписать функцию правдоподобия так, чтобы необходимая величина (сумма, разность, отношение и т.п.) являлась свободным параметром при минимизации.

Нейронные сети

Постановка задачи

Пусть стоит задача идентификации типа частицы, зарегистрированной детектором. В отличие от предыдущего задания, в котором было необходимо определить полное число событий каждого типа, теперь требуется идентифицировать каждое событие.

Постройте нейронную сеть, которая будет решать поставленную задачу. Используйте массив экспериментальных данных, сгенерированный в рамках задания 0, для обучения сети. В качестве входов сети используйте энерговыделения в каждом слое, в качестве выхода – тип частицы. В качестве сигнала используйте события с гамма-квантами, в качестве фона – события с электронами.

Результатами выполнения задания являются структура сети, распределения ответа сети для событий сигнала и фона, и кривая $\alpha - \beta$ (ROC).

Основные и дополнительные вопросы

1. Нарисуйте кривые $\alpha - \beta$ для различных конфигураций сети.
2. Используйте подмножество слоев в качестве входов сети. Нарисуйте соответствующие кривые $\alpha - \beta$ для различных наборов входов и сделайте вывод о том, какие слои предоставляют больше информации о типе частицы.
3. Покажите влияние разрешения детектора на результат работы сети. Для этого модифицируйте энерговыделения в слоях, добавив к ним шум, и постройте семейство кривых $\alpha - \beta$. Сеть должна быть обучена на исходных (не модифицированных) данных.
4. Покажите влияние калибровки детектора на результат работы сети. Для этого модифицируйте энерговыделения в слоях, умножив их на масштабный коэффициент. Масштабный коэффициент должен быть выбран случайно для каждого слоя, но один и тот же коэффициент должен быть использован для всех событий. Такой подход позволяет смоделировать именно ошибку калибровки, когда существует неизвестный масштабный фактор, который не меняется от события к событию. Необходимо провести несколько раундов такого моделирования, и для каждого раунда построить кривую $\alpha - \beta$. Сеть должна быть обучена один раз на исходных (не

модифицированных) данных. Ширина семейства кривых покажет влияние ошибки калибровки.

Многомерная классификация данных

Постановка задачи

В этом задании требуется решить ту же задачу, что и в предыдущем задании - необходимо идентифицировать тип частицы, зарегистрированной детектором. Однако теперь эту задачу надо решить не с помощью нейронной сети, а с помощью других алгоритмов многомерной классификации.

Используйте массив экспериментальных данных, сгенерированный в рамках задания 0, для обучения алгоритмов. В качестве входных данных используйте энерговыделения в каждом слое, в качестве результата работы алгоритма – тип частицы. В качестве сигнала используйте события с гамма-квантами, в качестве фона – события с электронами.

Необходимо сравнить следующие алгоритмы:

1) отношение правдоподобий в предположении независимости входных данных;
2) подсчет числа событий в тренировочном множестве в окрестности текущего события;
3) усиленные деревья принятия решений.

Требуется реализовать эти алгоритмы, построить отклик классификатора для событий сигнала и фона и кривую $\alpha - \beta$ (ROC) для каждого алгоритма, и сделать вывод, какой алгоритм лучше всего работает в данном случае.

Основные и дополнительные вопросы

1. Оцените, улучшает ли предварительная подготовка данных с помощью метода главных компонент характеристики выбранного алгоритма классификации.
2. С какими весами следует сложить энерговыделения в слоях калориметра, чтобы наилучшим образом разделить электроны и гамма-кванты?
3. Покажите влияние разрешения детектора на результат работы сети. Для этого модифицируйте энерговыделения в слоях, добавив к ним шум, и постройте семейство кривых $\alpha - \beta$. Выбранные алгоритмы должны быть обучены на исходных (не модифицированных) данных. Характеристики какого алгоритма ухудшаются наибольшим образом при добавлении шума? Восстанавливаются ли исходные характеристики выбранных алгоритмов, если обучить их на зашумленных данных?

Билеты к экзамену

Билет №1	1. Случайные величины. Дискретные и непрерывные распределения. Параметры распределений: среднее значение, дисперсия, моменты. Ковариационная матрица, коэффициент корреляции. Преобразование распределения при замене переменных. 2. Задача разделения сигнала и фона. Критерий разделения, мощность и значимость критерия. Лемма Неймана-Пирсона и наилучший критерий разделения. Практические методы построения критериев разделения: факторизация функции правдоподобия. Метод главных компонент.
Билет №2	1. Метод Монте-Карло. Интегрирование методом М.-К. Алгоритмы генерации случайных чисел: метод Неймана, метод трансформации, комбинированный. Алгоритм генерации нормально-распределенной величины. 2. Нейронные сети. Однослойный и многослойный перцептрон. Обучение перцептрона, алгоритм обратного распространения ошибок. Глобальные методы оптимизации.
Билет №3	1. Основные распределения и их параметры: биномиальное, Пуассона, равномерное, нормальное, $\chi 2$. Центральная предельная теорема. 2. Задача разделения сигнала и фона. Критерий разделения, мощность и значимость критерия. Лемма Неймана-Пирсона и наилучший критерий разделения. Практические методы построения критериев разделения: факторизация функции правдоподобия; методы, основанные на подсчете числа событий.
Билет №4	1. Оценка параметров распределений по ограниченной выборке. Свойства оценок: состоятельность, смещение, эффективность, робастность (устойчивость). Способы построения оценок, метод моментов. Способы построения несмещенной оценки. 2. Задача разделения сигнала и фона. Критерий разделения, мощность и значимость критерия. Лемма Неймана-Пирсона и наилучший критерий разделения. Практические методы построения критериев разделения: факторизация функции правдоподобия; линейный дискриминантный анализ Фишера.
Билет №5	1. Метод максимального правдоподобия. Оценка ошибок в методе максимального правдоподобия. Примеры использования метода максимального правдоподобия для подгонки гистограммы, определения времени жизни. 2. Задача обратной свертки (unfolding). Постановка задачи. Методы получения результатов без обратной свертки. Прямое решение задачи. Регуляризация. Регуляризация Тихонова.
Билет №6	1. Метод максимального правдоподобия. Оценка ошибок в методе максимального правдоподобия. Примеры использования метода максимального правдоподобия для подгонки гистограммы с фиксированным числом событий, оценки дисперсии. 2. Способы построения критерия согласия. Критерий $\chi 2$. Оценка качества подгонки в методе максимального правдоподобия.

Билет №7	1. Метод наименьших квадратов. Оценка ошибок в методе наименьших квадратов. Метод наименьших квадратов в линейном приближении. Пример использования метода наименьших квадратов для подгонки гистограммы 2. Задача обратной свертки (unfolding). Постановка задачи. Методы получения результатов без обратной свертки. Прямое решение задачи. Регуляризация. Метод максимальной энтропии.
Билет №8	1. Теорема Байеса. Формулировка теоремы Байеса для непрерывных распределений. Применение т.Б. для оценки погрешностей. Связь т.Б. и метода максимального правдоподобия. Пример применения т.Б. для оценки эффективности, для оценки верхнего предела при близости измеренного значения к границе интервала возможных значений. 2. Задача разделения сигнала и фона. Критерий разделения, мощность и значимость критерия. Лемма Неймана-Пирсона и наилучший критерий разделения. Практические методы построения критериев разделения: нейронные сети; деревья принятия решений.
Билет №9	1. Теорема Байеса. Формулировка теоремы Байеса для непрерывных распределений. Применение т.Б. для оценки погрешностей. Связь т.Б. и метода максимального правдоподобия. Пример применения т.Б. для оценки эффективности, для оценки уровня сигнала при наличии фона. 2. Способы построения критерия согласия. Критерий серий. Критерий Колмогорова.
Билет №10	1. Интервальные оценки, доверительные интервалы. Построение доверительных интервалов методом Неймана. Доверительные интервалы в случае нормального распределения. 2. Нейронные сети. Однослойный и многослойный перцептрон. Радиально-базисные сети. Сети с самоорганизацией, сеть Кохонена.
Билет №11	1. Интервальные оценки, доверительные интервалы. Построение доверительных интервалов методом Неймана. Построение доверительных интервалов в методах максимального правдоподобия и наименьших квадратов. 2. Задача разделения сигнала и фона. Критерий разделения, мощность и значимость критерия. Лемма Неймана-Пирсона и наилучший критерий разделения. Задача проверки сложных гипотез. Метод отношения правдоподобий.

Оценочные материалы по промежуточной аттестации, предназначенные для проверки соответствия уровня подготовки по дисциплине требованиям ФГОС ВО, хранятся на кафедре-разработчике РПД в печатном и электронном виде.

Лист актуализации рабочей программы
по дисциплине «Методы анализа экспериментальных данных»

№	Характеристика внесенных изменений (с указанием пунктов документа)	Дата и № протокола Учёного совета ФФ НГУ	Подпись ответственного

Аннотация

к рабочей программе дисциплины
«Методы анализа экспериментальных данных»
направление подготовки: **03.04.01 Прикладные математика и физика**
Профиль: **Прикладные математика и физика. Информационные процессы и системы**

Программа дисциплины **«Методы анализа экспериментальных данных»** составлена в соответствии с требованиями ФГОС ВО к уровню магистратуры по направлению подготовки **03.04.01 Прикладные математика и физика,** а также задачами, стоящими перед Новосибирским государственным университетом по реализации Программы развития НГУ. Дисциплина реализуется на физическом факультете Федерального государственного автономного образовательного учреждения высшего профессионального образования Новосибирский национальный исследовательский государственный университет (НГУ) кафедрой физико-технической информатики в качестве дисциплины по выбору. Дисциплина изучается студентами **первого** курса **магистратуры** физического факультета в весеннем семестре.

Цели курса – ознакомление студентов с современными методами анализа результатов измерений, получаемых в физических экспериментах. Первая часть курса посвящена повторению и углублению знаний, полученных в курсах теории вероятностей и математической статистики. Вторая часть курса посвящена применению методов интеллектуального и многопараметрического анализа данных. В третьей части курса рассматриваются отдельные задачи, часто возникающие при анализе экспериментальных данных. В рамках практических занятий студенты получают возможность использовать полученные знания для решения индивидуально подобранных задач.

Дисциплина нацелена на формирование у обучающегося профессиональной компетенции:

Результаты освоения образовательной программы (компетенции)	Индикаторы	Результаты обучения по дисциплине
ПК-1 Способность осваивать и применять специализированные знания в области физико-математических и (или) естественных наук в своей профессиональной деятельности.	**ПК 1.1** Применяет специализированные знания естественных и (или) физико-математических наук при решении поставленных задач в специализированной области своей профессиональной деятельности. **ПК 1.2** Применяет классические и новые знания при решении поставленных задач в специализированной области своей профессиональной деятельности.	**Знать** теоретические положения математической статистики и теории вероятностей, лежащие в основе изучаемых методов анализа данных: методы оценки параметров распределений, методы максимального правдоподобия и наименьших квадратов, построение критериев согласия, основы теории проверки гипотез, основы теории принятия решений, байесовский подход к оценке вероятностей; знать основные алгоритмы многомерного анализа данных, в частности, методы построения функций правдоподобия, нейронных сетей, деревьев принятия решений. **Уметь** использовать методы Монте-Карло для моделирования эксперимента и оценки погрешностей; оценивать параметры распределений при

Результаты освоения образовательной программы (компетенции)	Индикаторы	Результаты обучения по дисциплине
		наличии корреляций и фона; применять методы максимального правдоподобия и наименьших квадратов; применять методы многомерного анализа данных; применять комплексные алгоритмы при анализе больших массивов данных; решать типичные задачи, возникающие при анализе данных современного физического эксперимента. **Владеть** программным инструментарием для моделирования и анализа данных физического эксперимента: пакетами ROOT и GEANT4.

Преподавание дисциплины предусматривает следующие формы организации учебного процесса: лекции, практические занятия, самостоятельная работа студента, консультации, экзамен.

Программой дисциплины предусмотрены следующие виды контроля:
Текущий контроль успеваемости: решение заданий из задания для самостоятельного решения
Промежуточная аттестация: экзамен.

Общая трудоемкость рабочей программы дисциплины составляет **3** зачетные единицы /**108** академических часов.