Санкт-Петербургский филиал федерального государственного автономного образовательного учреждения высшего образования "Национальный исследовательский университет "Высшая школа экономики"

Факультет Санкт-Петербургская школа экономики и менеджмента
Национального исследовательского университета
«Высшая школа экономики»

Департамент прикладной математики и бизнес-информатики

Рабочая программа дисциплины
Вычислительная статистика

для образовательной программы «Анализ больших данных в бизнесе, экономике и обществе»
направления подготовки 01.04.02«Прикладная математика и информатика»
уровень магистратура

Разработчик(и) программы
Шмилева Е.Ю., доцент,к.ф.-м.н., elena.shmileva@gmail.com

Согласована менеджером ОП Анализ больших данных в бизнесе, экономике и обществе

«30»августа 2016г.

Е.С. Авдониной _____

Утверждена Академическим руководителем образовательной программы

А.В. Сироткин _____

«30»августа 2016г.

Санкт-Петербург, 2016

1. Область применения и нормативные ссылки

Настоящая рабочая программа дисциплины устанавливает минимальные требования к знаниям и умениям студента, а также определяет содержание и виды учебных занятий и отчетности.

Программа предназначена для преподавателей, ведущих дисциплину «Вычислительная статистика», учебных ассистентов и студентов направления подготовки 01.04.02 «Прикладная математика и информатика», обучающихся по образовательной программе «Анализ больших данных в бизнесе, экономике и обществе».

Рабочая программа дисциплины разработана в соответствии с:

- Образовательным стандартом НИУ ВШЭ http://www.hse.ru/data/2016/11/02/1111123560/01.04.02%20Прикладная%20математика%20и%20информатика.pdf;
- Образовательной программой «Анализ больших данных в бизнесе, экономике и обществе», направление подготовки 01.04.02 «Прикладная математика и информатика», ;
- Объединенным учебным планом университета по образовательной программ «Анализ больших данных в бизнесе, экономике и обществе».

2. Цели освоения дисциплины

Целями освоения дисциплины «Вычислительная статистика» являются:

- изучение вычислительных и статистических методов обработки данных,
- освоение программного обеспечения для статистической обработки данных,
- подготовка к самостоятельной исследовательской деятельности в области статистики данных.

3. Компетенции обучающегося, формируемые в результате освоения дисциплины

В результате освоения дисциплины студент осваивает следующие компетенции:

Компетенция	Код по ОС ВШЭ	Уровень формирования компетенции	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции	Форма контроля уровня сформированности компетенции
Профессиональные компетенции					
А) Социально-личностные компетенции					
Способен использовать социальные и мультикультурные различия для решения проблем в профессиональной и социальной деятельности.	ПК-3	РБ, СД	Способен описывать проблемы на разном уровне детализации и объяснять постановку задачи людям с различным уровнем подготовки.	Семинарские занятия	Домашняя работа, экзамен .
Б) Инструментальные компетенции					
Способен описывать проблемы и ситуации профессиональной деятельности, используя язык и аппарат прикладной мате-	ПК-14	РБ, СД	Может переформулировать задачу поставленную на естественном языке, как задачу определения статистических различий.	Лекции, семинарские занятия	Контрольная работа, экзамен.

Компетенция	Код по ОС ВШЭ	Уровень формирования компетенции	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции	Форма контроля уровня сформированности компетенции
Профессиональные компетенции					
А) Социально-личностные компетенции					
матики при решении междисциплинарных проблем.					
Способен создавать, описывать и ответственно контролировать выполнение технологических требований и нормативных документов в профессиональной деятельности.	ПК-15	РБ, СД	Способен предложить последовательность методов обработки данных и статистических тестов для решения поставленной прикладной задачи.	Семинарские занятия, практические занятия	Домашняя работа, контрольная работа, экзамен
Способен использовать в профессиональной деятельности знания в области естественных наук, математики и информатики, понимание основных фактов, концепций, принципов теорий, связанных с прикладной математикой и информатикой.	ПК-16	РБ, СД	Может построить вероятностную модель заданного процесса.	Лекции, семинарские занятия	Домашняя работа, экзамен.
Способен строить и решать математические модели в соответствии с направлением подготовки и специализацией.	ПК-17	РБ, СД	Умеет использовать методы семплирования для решения прикладных задач.	Лекции, практические занятия	Домашняя работа, контрольная работа, экзамен.

4. Место дисциплины в структуре образовательной программы

Настоящая дисциплина относится к циклу дисциплин по выбору студента.
Изучение данной дисциплины базируется на следующих дисциплинах:

- «Теория вероятностей и математическая статистика»,
- «Линейная алгебра»,
- «Математический анализ».

Для освоения учебной дисциплины, студенты должны владеть следующими знаниями и компетенциями:

- использовать социальные и мультикультурные различия для решения проблем в профессиональной и социальной деятельности;
- описывать проблемы и ситуации профессиональной деятельности, используя язык и аппарат прикладной математики при решении междисциплинарных проблем;
- создавать, описывать и ответственно контролировать выполнение технологических требований и нормативных документов в профессиональной деятельности;
- использовать в профессиональной деятельности знания в области естественных наук, математики и информатики, понимание основных фактов, концепций, принципов теорий, связанных с прикладной математикой и информатикой;
- строить и решать математические модели в соответствии с направлением подготовки и специализацией.

Основные положения дисциплины должны быть использованы в дальнейшем при изучении следующих дисциплин:

- «Современные методы принятия решений»,
- «Распределенная обработка и анализ больших данных»,
- «Информационный поиск и обработка текстов на естественном языке».

5. Тематический план учебной дисциплины

ОБЪЕМ ДИСЦИПЛИНЫ – 4 зачетные единицы.

№	Название раздела	Всего часов	Аудиторные часы			Самостоятельная работа
			Лекции	Семинары	Практические занятия	
1	Основные вероятностные распределения и их математическое моделирование. Методы Монте-Карло.	27	3	3	3	18
2	Марковские цепи. MCMC. Алгоритм Метрополис-Гастингса. Многомерные распределения. Самплинг по Гиббсу	37	4	4	4	25
3	Улучшение и объединение оценок. Метод складного ножа и бутстрап.	23	2	2	1	18
4	Перестановочные тесты для проверки статистических гипотез.	10	1	1		8
5	Кросс-валидация. Калибровка модели, обучающие выборки.	11	1	1	1	8
6	Пространственная статистика. Геостатистика.	18	2	2	2	12
7	Финансовые модели. ARCH, GARCH процессы.	17	2	2	1	12
8	Фильтры Калмана. Принцип максимума априорной вероятности.	9	1	1		7
ИТОГО:		**152**	**16**	**16**	**12**	**108**

6. Формы контроля знаний студентов

Тип контроля	Форма контроля	2 модуль								3 модуль								Параметры
		1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	
Текущий	Контрольная работа														*			Письменная работа 90 минут
	Промежуточная контрольная работа						*											Письменная работа 90 минут
	Домашнее задание			*								*						Проверочное домашнее задание в письменной форме
Итоговый	Экзамен																*	Письменный экзамен 60 мин.

7. Критерии оценки знаний, навыков

Критерии оценки **контрольной работы и промежуточной контрольной работы**:

Полнота и правильность решения набора типичных задач по темам практических занятий.

Использование языка статистической обработки данных R для решения предложенных задач обязательно. Результатом контрольной работы преподаватель ожидает: теоретическое обоснование решения и письменный отчет о результатах компьютерных экспериментов.

При оценке решения каждой задачи учитываются следующие факторы:

- точность и правильность теоретического решения;
- понимание и инициативность в написании программ, решающих поставленные задачи.

За решение каждой задачи выставляются накопительные баллы в соответствии с точностью и полнотой решения и описания процесса решения задачи.

Общая сумма баллов за контрольную работу 10 баллов.

Критерии оценки **за домашнюю работу:**

Задание представляет собой написание программы для статистического исследования заданного или собранного набора данных.

Результаты работы предоставляются в виде кода на языке R и письменного отчета описывающего этапы исследования. По требованию преподавателя возможна устная защита исследования.

Домашняя работа с опозданием не принимается, за исключением случаев, когда они не выполнены в срок по уважительной причине (это оговаривается с преподавателем). В случае болезни студент обязан предупредить преподавателя заранее по э-майлу, после выздоровления предъявить медицинскую справку в учебный офис. Срок сдачи домашней работы в этом случае не должен превышать двух недель с момента выздоровления.

Максимальная оценка за домашнюю работу – **10 баллов.**

Критерии оценки **за экзамен:**

Экзамен будет содержать задачи, затрагивающие все темы курса, которые необходимо будет решить теоретически, и практические задания на компьютере.

За каждое правильно выполненное задание присваивается накопительные баллы, которые суммируются.

Максимальная оценка за экзамен – **10 баллов.**

8. Содержание дисциплины

Раздел 1 Основные вероятностные распределения и их математическое моделирование. Методы Монте-Карло.
Псевдослучайные числа. Линейный конгруэнтный генератор. Генерирование равномерного распределения. Генерирование дискретных и экспоненциального распределений. Метод обратного преобразования. Генерирование стандартного нормального распределения. Метод принятия/отвержения (выборка с отклонением). Вероятности, квантили, базовые статистики, гистограммы, Q-Q графики для сгенерированных выборок. Интегрирование методом Монте-Карло.

Основная литература

- J.E.Gentle, (2009) Computational Statistics, Springer, 720 pages
- Г.И.Ивченко, Ю.И.Медведев, Введение в математическую статистику, Издательство ЛКИ, Москва, 2014, 600 с.

Раздел 2 Марковские цепи. MCMC. Алгоритм Метрополис-Гастингса. Многомерные распределения. Самплинг по Гиббсу.
Случайные блуждания. Переходные вероятности марковских цепей. Монте-Карло для марковских цепей. Свойства и приемущества получаемых симуляций. Скорости сходимости алгоритмов.

Основная литература

- J.E.Gentle, (2009) Computational Statistics, Springer, 720 pages

Раздел 3 Улучшение и объединение оценок. Метод складного ножа и бутстрап.
Бутстрап для оценки среднего, дисперсии и других параметров выборки. Доверительные интервалы параметров с помощью бутстрепа. Свойства получаемых оценок. Коррекция смещения бутстрапом. Сравнение метода складного ножа и бутстрапа.

Основная литература

- J.E.Gentle, (2009) Computational Statistics, Springer, 720 pages
- Г.И.Ивченко, Ю.И.Медведев, Введение в математическую статистику, Издательство ЛКИ, Москва, 2014, 600 с.

Раздел 4 Перестановочный тест для проверки статистических гипотез.
Нахождение p-значения. Примеры. Точный тест Фишера.

Основная литература

- J.E.Gentle, (2009) Computational Statistics, Springer, 720 pages
- Г.И.Ивченко, Ю.И.Медведев, Введение в математическую статистику, Издательство ЛКИ, Москва, 2014, 600 с.

Раздел 5 Кросс-валидация.
Калибровка модели, обучающие выборки. Применение к линейной регрессии.

Основная литература

- J.E.Gentle, (2009) Computational Statistics, Springer, 720 pages
- Г.И.Ивченко, Ю.И.Медведев, Введение в математическую статистику, Издательство ЛКИ, Москва, 2014, 600 с.

Раздел 6. Пространственная статистика. Геостатистика.
Многомерное гауссовское распределение. Матрица ковариаций. Гауссовское случайное поле. Функция ковариаций. Стационарность полей. Параметрические семейства функций ковариаций. Вариограммный анализ. Оценивание функции ковариаций поля. Кригинг. Нестационарный кригинг. Методы обработки распределений с тяжелыми хвостами.

Основная литература

- C. Reimann, P. Filzmoser, R. Garrett, R. Dutter, (2008), Statistical Data Analysis Explained: Applied Environmental Statistics with R, Wiley
- E.Shmileva, E. Spodarev, S.Roth, Extrapolation of stationary random fields, in: V.Schmidt (ed). Stochastic Geometry, Spatial Statistics and Random Fields. Springer, Lecture Notes in Mathematics, Volume 2120, 2015, pp 321-368 Springer, 2015

Раздел 7. Финансовые модели.
ARCH и GARCH модели финансовых рядов. Элементы стохастического анализа.

Основная литература

- McNeil, R. Frey, P. Embrechts, Quantitative Risk Management: concepts, techniques and tools, Princeton series in Finance, 2015, 2nd edition
- P.Glasserman, (2003) Monte Carlo Methods in Financial Engineering, Springer.

Раздел 8. Фильтры Калмана. Принцип максимума априорной вероятности.
Фильтр Калмана в применении к финансовым рядам. Прогнозирование значений финансовых рядов.

9. Образовательные технологии

1. Работа в малых группах.
2. Проведение практических и семинарских занятий в компьютерных классах.

9.1 Методические рекомендации преподавателю

Особых рекомендаций для преподавателя нет.

9.2 Методические указания студентам по освоению дисциплины

Студентам рекомендуется регулярно посещать лекционные занятия и семинары. Основной прогресс в изучении области ожидается от самостоятельной работы над домашними заданиями, которыми снабжается каждое семинарское и практическое задание. Приветствуется коллективная работа над домашними работами. Рекомендуется параллельно с курсом лекций прочитывать соответствующие главы из книг списка литературы.

При выполнении проверочного домашнего задания приветствуется творческий подход и коллективная работа.

9.3 Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине

Студентам рекомендуется пользоваться сайтом вопросов и ответов qa.piterdata.ninja. Доступен на чтение всем, доступен по логину студентам программы, предоставляемому по запросу.

Для углубления понимания курса, можно ознакомиться со свободно доступными онлайн курсами:

- Introduction to Computational Thinking and Data Science: https://www.edx.org/course/introduction-computational-thinking-data-mitx-6-00-2x-4
- Computational Probability and Inference https://www.edx.org/course/computational-probability-inference-mitx-6-008-1x
- Probability: Distribution Models & Continuous Random Variables https://www.edx.org/course/probability-distribution-models-purduex-416-2x

10. Оценочные средства для текущего контроля и аттестации студента

10.1 Оценочные средства для оценки качества контроля освоения дисциплины в ходе текущего контроля

Примеры контрольной работы:
Промежуточная контрольная работа

- Задача (Длина самой длинной серии из 1). Пусть X1, X2,..., Xn - iid последовательность из 0 и 1. P(Xi = 1) = p. Постройте эмпирический закон распределения длины самой длинной серии из единиц для n = 20 и p = 0.9. Т.е. по горизонтали должно быть отложено i = 0..20, а по вертикали, частота, с которой самая длинная серия из единиц была длины i.
- Методом Монте-Карло посчитайте интеграл от exp^{-x^2} на интервале от (-5,5).
- Сгенерируйте 1000 равномерно распределенных на отрезке [0;1] с.в. Xi. Рассчитайте значения случайных величин X и $\sigma^{\wedge 2}$. Сравните их с E(Xi) и Var(Xi). Постройте гистограмму (эмпирическую функцию плотности) для полученных 1000 чисел.
- Проведя 1000 экспериментов на компьютере найдите 5%-ое пороговое значение статистик U_1 (MannWhitney, $n_1 = 5$ и $n_2 = 4$) и T^+ (Wilcoxon Signed Rank Test, $n = 9$) для двусторонней альтернативной гипотезы. Сравните его с асимптотическим.

Контрольная работа 2

- Рассмотрите X1, X2,..., Xn н.о.р.с.в. Какова оценка второго момента методом складного ножа? Будет ли эта оценка несмещенной?

- Оцените среднее и посчитайте 95% доверительный интервал бутсрапом или методом складного ножа для скорости движения, если контрольные камеры ДПС зафиксировали скорость движения лишь 6-и автомобилей: 100, 110, 128, 96, 109, 85.

- Нарисуйте 3 реализации броуновского движения, используя 5, 10, 100 и 1000 слагаемых.

- Используя Метрополис-Гастинг алгоритм сгенерируйте выборку из стандартного нормального распределения. В качестве плотности переходной вероятности вспомогательной Марковской цепи возьмите g(x|y)=1/2 exp{-|x-y|}. Выведите получившуюся последовательность на экран.

- Используя самплинг по Гиббсу сгенерируйте двумерное гауссовское распределение со средним компонент 0, дисперсией 1 и корреляцией компонент r = 0, 0.2, 0.5, 0.8. Опишите эффективность самплинга по Гиббсу для этой задачи.

- Подберите параметрическую модель вариограммы подходящую стационарным пространственным данным. Методом наименьших квадратов оцените параметры модели. Постройте кригинг экстраполяцию пространственных данных. Сделайте прогноз значений в нескольких точках пространства.

Примеры проверочного Домашнего Задания.

- Смоделируйте и проиллюстрируйте парадокс Дней Рождений (вероятность, что 2 человека в группе из 23 человек празднуют день рождения в один день, >1/2).
- Смоделируйте и проиллюстрируйте Санкт-Петербургский парадокс.
- Сгенерируйте 100 бросаний справедливой монетки. Посчитайте максимальную длину последовательности из 1 или из 0 (в реальном эксперименте где-то 6-8) и число переключений 01 и 10. Смоделируйте данный эксперимент 2000 раз и соберите статистику по изучаемым величинам. Сравните с реальными статистиками при подбрасывании монетки 100 раз.

10.2 Примеры заданий промежуточного /итогового контроля

Экзаменационные вопросы:

- Методом Монте-Карло посчитайте площадь окружности и приведите приближенное значение числа Пи. Как методом Монте-Карло посчитать приближенное значение *e*?
- Сгенерируйте 10 равномерно распределенных на отрезке [0;1] с.в. Рассчитайте сумму X = X1 +...+X10. Повторите эксперимент 1000 раз. Постройте эмпирическую функцию плотности полученных 1000 значений сумм.
- В одной партии игрок выигрывает 2^n рублей с вероятностью 2^{-n}, $n \in$ N. Постройте 3 реализации зависимости среднего выигрыша от числа партий.

- Предположим, что имеется выборка из гамма распределения размером n и параметрами a и 1. Опишите как бы с помощью бутстрепа вы бы построили 95% односторонний нижний доверительный интервал для стандартного отклонения.

- Результаты спортсменов на соревнованиях по прыжкам в длину: 183; 164; 227; 178; 189; 233; 161; 231 (в см). Бутстрап методом постройте оценку для средней длины прыжка и посчитайте двусторонний 95%-ый доверительный интервал для этой величины. Укажите сколько ресамплингов Вы проделали.

11. Порядок формирования оценок по дисциплине

Накопленная оценка по дисциплине рассчитывается с помощью взвешенной суммы оценок за отдельные формы текущего контроля знаний следующим образом:

$$О_{накопленная} = 1/3 \cdot О_{текущий1} + 1/3 \cdot О_{текущий\ 2} + 1/3 \cdot О_{текущий\ 3},$$ где

$О_{текущий1}$ – оценка за Контрольную Работу 1
$О_{текущий2}$ – оценка за Домашнюю Работу
$О_{текущий\ 3}$ – оценка за Контрольную Работу 2

Способ округления накопленной оценки текущего контроля: арифметический.

Результирующая оценка по дисциплине (которая идет в диплом) рассчитывается следующим образом:

$$O_{результ} = 0{,}7 \cdot O_{накопл} + 0{,}3 \cdot O_{экз}\text{, где}$$

$O_{накопл}$ – накопленная оценка по дисциплине
$O_{экз}$ – оценка за экзамен

Способ округления экзаменационной и результирующей оценок: арифметический.

12. Учебно-методическое и информационное обеспечение дисциплины

12.1 Основная литература

- Gentle J.E. Computational Statistics [Electronic Resource] / James E. Gentle.- Dordrecht, Heidelberg, London, New York :Springer, 2009.- 720 p. - Authorized access: http://link.springer.com/book/10.1007/978-0-387-98144-4#page-1 (Online Digital Library "Springer eBooks")/

12.2 Дополнительная литература

- A. McNeil, R. Frey, P. Embrechts. Quantitative Risk Management: concepts, techniques and tools [Electronic Resource] / Alexander J. McNeil, Rudiger Frey and Paul Embrechts.- Princeton University Press, 2005. – 554 p. - Authorized access: http://library.books24x7.com/searchresults.aspx (Online Digital Library "Books24x7").
- Shmileva E. Extrapolation of stationary random fields / Evgeny Spodarev, Elena Shmileva and Stefan Roth. 2013, 52 p. - URL: https://arxiv.org/pdf/1306.6205v1.pdf (Open access archive arXiv.org).

12.3 Справочники, словари, энциклопедии

Справочники, словари, энциклопедии не требуются.

12.4 Ресурсы информационно-телекоммуникационной сети «Интернет»

www.gks.ru – Федеральная служба государственной статистики
www.qrmtutorial.org –Электронные материалы и R-коды по теме финансовые модели
www.statistik.tuwien.ac.at/StatDA/R-scripts/ - Электронные материалы и R-коды по теме геостатистика

12.5 Программные средства

Для успешного освоения дисциплины, студент использует следующие **программные средства:**

- R - язык программирования для статистической обработки данных и работы с графикой

12.6 Информационные справочные системы

Справочные информационные системы не используются.

12.7 Дистанционная поддержка дисциплины

Дистанционная поддержка дисциплины не требуется.

13. Материально-техническое обеспечение дисциплины

Практические занятия и семинары выполняются в компьютерном классе.
Необходим проектор для лекций.