Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Центр сопровождения образовательных инициативных проектов

УТВЕРЖДЕНО:

Руководитель сетевой ОПОП
В.В. Кашпур

Рабочая программа дисциплины

# Теория вероятности и математическая статистика для анализа данных

по направлению подготовки

**09.04.03 Прикладная информатика**

Направленность (профиль) подготовки :
**«Дата-аналитика для бизнеса»**

Форма обучения
**Очная**

Квалификация
**Магистр**

Год приема
**2023**

Томск – 2023

## 1. Цель и планируемые результаты освоения дисциплины

Целью освоения дисциплины является формирование следующих компетенций:

ОПК-1 Способен самостоятельно приобретать, развивать и применять математические, естественнонаучные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте.

ОПК-3 Способен анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями.

ОПК-7 Способен использовать методы научных исследований и математического моделирования в области проектирования и управления информационными системами;.

ПК-4 Способен разрабатывать и реализовывать маркетинговые программы с использованием инструментов комплекса маркетинга.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИОПК-1.1 Владеет фундаментальными математическими, естественнонаучными, социально-экономическими и профессиональными понятиями в контексте решения задач в области информационных технологий.

ИОПК-1.2 Определяет взаимосвязи, закономерности, обобщает, абстрагирует фундаментальные модели, законы, методики для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте.

ИОПК-1.3 Развивает и применяет математические, естественнонаучные, социально-экономические и профессиональные знания для решения задач

ИОПК-3.1 Осуществляет сбор, обработку и анализ научно-технической информации, необходимой для решения профессиональных задач.

ИОПК-3.2 Умеет работать с различными видами информации с помощью различных средств информационных и коммуникационных технологий.

ИОПК-7.1 Владеет методами научных исследований и математического моделирования для решения профессиональных задач в области проектирования и управления информационными системами.

ИОПК-7.3 Разрабатывает и применяет математические модели в области проектирования и управления информационными системами.

ИПК-4.3 Использует методы проведения маркетинговых исследований в области распределения (дистрибуции) и продаж.

## 2. Задачи освоения дисциплины

– Освоить аппарат теории вероятностей, математической и прикладной статистики для анализа данных.

– Научиться применять понятийный аппарат и основные методы вероятностного и статистического анализа для решения практических задач профессиональной деятельности.

## 3. Место дисциплины в структуре образовательной программы

Дисциплина относится к Блоку 1 «Дисциплина (модули)».

Дисциплина относится к обязательной части образовательной программы.

## 4. Семестр(ы) освоения и форма(ы) промежуточной аттестации по дисциплине

Семестр 1, зачет.

Семестр 2, экзамен.

**5. Входные требования для освоения дисциплины**

Для успешного освоения дисциплины требуются компетенции, сформированные в ходе освоения образовательных программ предшествующего уровня образования.

**6. Язык реализации**

Русский

**7. Объем дисциплины**

Общая трудоемкость дисциплины составляет 6 з.е., 216 часов, из которых:
-лекции: 20 ч.
-практические занятия: 40 ч.
Объем самостоятельной работы студента определен учебным планом.

**8. Содержание дисциплины, структурированное по темам**

Тема 1. Теория вероятностей. Случайные величины

Дискретные и непрерывные случайные величины. Способы задания случайных величин. Функции случайных величин. Основные числовые характеристики случайных величин. Основные законы распределения дискретных и непрерывных случайных величин. Другие числовые характеристики. Системы случайных величин.

Тема 2. Статистика. Введение.

Генеральная и выборочная совокупности. Способы представления выборок: табличные и графические.

Тема 3. Проверка статистических гипотез.

Проверка статистических гипотез. Алгоритм проверки. Критерий согласия. Пример работы алгоритма. P-value. Практика по работе с выборками. Генерация. Визуализация. Описательные статистики. Проверка вида распределения.

Тема 4. Закон больших чисел. Центральная предельная теорема

Нормальный закон распределения. ЗБЧ. ЦПТ.

Тема 5. Параметрические критерии сравнения групп

Параметрические критерии сравнения групп. Z-test, t-test, Fisher. Правило сложения дисперсий. ANOVA. Практика по параметрическим критериям.

Тема 6. Непараметрические критерии сравнения групп

U -критерии Манна-Уитни и критерий Вилкоксона. Критерии Краскала-Уолиса и Фридмана. Практика по непараметрическим критериям.

Тема 7. Корреляционный анализ

Корреляционный анализ количественных данных. Ранговая корреляция. Корреляционный анализ категоризованных данных. Таблицы сопряженности. Корреляционный анализ а Python.

Тема 8. Регрессионный анализ

Парная регрессия. Постановка задачи. Оценка параметров. Проверка качества модели парной регрессии. Множественная регрессия. Постановка задачи. Оценка параметров. Проверка качества модели множественной регрессии. Построение регрессионных моделей в Python.

**9. Текущий контроль по дисциплине**

Текущий контроль по дисциплине проводится путем контроля посещаемости вебинаров и практических занятий, тестов по лекционному материалу и фиксируется в форме контрольной точки не менее одного раза в семестр.

Оценочные материалы текущего контроля размещены на сайте ТГУ в разделе «Информация об образовательной программе» – https://www.tsu.ru/sveden/education/eduop/.

**10. Порядок проведения и критерии оценивания промежуточной аттестации**

Зачет в первом семестре проводится в форме теста, включающего в себя как вопросы по теории, так и решение небольших практических задач.

Тест состоит из 10 вопросов разной сложности, за каждый из которых можно набрать от 1 до 3 баллов. Максимум за тест 20 баллов.

| Баллы | Оценка |
|---|---|
| [13,20] | Зачтено |
| [0,13) | Незачтено |

Экзамен во втором семестре проводится в форме теста, включающего в себя как вопросы по теории, так и решение небольших практических задач.

Тест состоит из 15 вопросов разной сложности, за каждый из которых можно набрать от 1 до 3 баллов. Максимум за тест 30 баллов.

| Баллы | Оценка |
|---|---|
| [26,30] | Отлично |
| [21,26) | Хорошо |
| [16,21) | Удовлетворительно |
| [0,16) | Неудовлетворительно |

Примерные вопросы теста:

1) Какие из ниже приведенных законов являются законами распределения непрерывных случайных величин?
   а) равномерное
   б) геометрическое
   в) биномиальное
   г) нормальное
   д) Пуассона
   г) Стьюдента
   д) экспоненциальное

2) Какие из ниже приведенных способов не определены для дискретных случайных величин?
   а) функция распределения
   б) ряд распределения
   в) плотность распределения

3) Для случайной величины, распределенной по стандартному равномерному закону чему равна вероятность попадания в интервал от 0.4 до 0.7?
   (Ввести значение)

4) Доверительным интервалом называется
   а) любой интервал, содержащий истинное значение оцениваемого параметра
   б) интервал, содержащий истинное значение оцениваемого параметра с вероятностью 1.
   в) интервал, содержащий истинное значение оцениваемого параметра с вероятностью $1 - \alpha$.

5) Для визуализации зависимости одного количественного показателя от одного категориального фактора, имеющего два и более уровней, лучше всего подойдет график:

а) Histogram
б) Boxplot
в) Scatterplot
г) q-q plot

6) Пусть Z – статистика правостороннего критерия. Zr – критическая граница, соответствующая уровню значимости 0.05. Zs – выборочное значение статистики, полученное по элементам выборки. Что такое p-value?
а) $p - value = P\left(Z \geq Zr \mid H_0\right)$
б) $p - value = P\left(Z \geq Zs \mid H_0\right)$
в) $p - value = P\left(Z < Zr \mid H_0\right)$
г) $p - value = P\left(Z < Zs \mid H_0\right)$

7) По критерию Шапиро-Уилка были получены следующие результаты. Какой вывод можно сделать?

W = 0.97396, p-value = 0.04472
а) Выборка подчиняется нормальному закону распределения на уровне значимости 0.05
б) Выборка не подчиняется нормальному закону распределения на уровне значимости 0.05
в) Выборка подчиняется нормальному закону распределения на уровне значимости 0.03
г) Выборка не подчиняется нормальному закону распределения на уровне значимости 0.03
д) Выборка подчиняется нормальному закону распределения на уровне значимости 0.01
е) Выборка не подчиняется нормальному закону распределения на уровне значимости 0.01

8) Для двух порядковых переменных при расчете коэффициента Спирмена были получены следующие результаты.
r = - 0,17558892 p = 0,0316143305
Какой вывод можно сделать при уровне значимости $\alpha = 0{,}05$ ?
а) имеется прямая статистически значимая связь между переменными;
б) имеется обратная статистически значимая связь между переменными;
в) между переменными нет статистически значимой корреляционной связи.

9) В таблице представлены результаты регрессионного анализа для многофакторной линейной модели.

| | $R^2 = 0.7005 \quad R^2_{adj} = 0.691$<br>$F(4, 127) = 74.25 \quad p < 2.2e - 16 \quad S_e = 2037$ | | | |
|---|---|---|---|---|
| $n = 130$ | Estimate | Std. Error | t-value | $p - value$ |
| (Intercept) | 3934.02 | 631.68 | 6.228 | 6.37e-09 |
| X1 | 201.11 | 18.50 | 10.872 | < 2e-16 |
| X2 | 23.25 | 75.49 | 0.308 | 0.7586 |
| X3 | -38.31 | 446.60 | -0.086 | 0.9318 |
| X4 | 898.30 | 388.79 | 2.310 | 0.0225 |

Сколько значимых параметров в той модели на уровне значимости 0.05? (введите число)

10) Какой критерий можно использовать для сравнения по некоторому уровню трех независимых ненормальных совокупностей.
    а) t-test
    б) ANOVA
    в) Критерий Манна-Уитни
    г) Критерий Вилкоксона
    д) Критерий Краскала-Уолиса
    е) Критерий Фридмана

Оценочные материалы для проведения промежуточной аттестации размещены на сайте ТГУ в разделе «Информация об образовательной программе» – https://www.tsu.ru/sveden/education/eduop/.

**11. Учебно-методическое обеспечение**

а) Электронный учебный курс по дисциплине в LMS «Data-Diving».
б) Оценочные материалы текущего контроля и промежуточной аттестации по дисциплине (https://www.tsu.ru/sveden/education/eduop/).
в) Методические указания по организации самостоятельной работы студентов.

**12. Перечень учебной литературы и ресурсов сети Интернет**

а) основная литература:
– Практическая статистика для специалистов Data Science: Пер. с англ. /. П. Брюс, Э. Брюс. — СПб.: БХВ-Петербург, 2018. — 304 с.: ил. ISBN 978-5-9775-3974-6.

б) дополнительная литература:
– Кендалл М. Д. Статистические выводы и связи / М. Кендалл, А. Стьюарт; Пер. с англ. Л. И. Гальчука, А. Т. Терехина; Под ред. А. Н. Колмогорова. - М. : Наука. Физматлит, 1973. - 899, [1] с.: ил..
URL: http://sun.tsu.ru/limit/2016/000074332/000074332.djvu
– Маккинли Уэс. Python и анализ данных. - Москва : ДМК Пресс, 2015. - 482 с. - ISBN 978-5-97060-315-4
– Джеймс Г., Уиттон Д., Хасти Е., Тибширани Р., Введение в статистическое обучение с примерами на языке R. М.: ДМК Пресс, 2016 г., 450 с.
– Орлов А.И., Прикладная статистика. Учебник. / А.И.Орлов.- М.: Издательство «Экзамен», 2004. - 656 с.
– Кабанова Т. В. Применение пакета R для решения задач прикладной статистики : учебное пособие : [для студентов и аспирантов университетов] / Т. В. Кабанова ; М-во образования и науки РФ, Нац. исслед. Том. гос. ун-т. - Томск : Издательский Дом Томского государственного университета, 2019. - 123 с.: ил., табл..
URL: http://vital.lib.tsu.ru/vital/access/manager/Repository/vtls:000668036

в) ресурсы сети Интернет:
– открытые онлайн-курсы
– machine learning repository https://archive.ics.uci.edu/ml/index.php
– https://www.kaggle.com/
– https://docs.scipy.org/doc/scipy/reference/stats.html/
– https://docs.python.org/3/library/statistics.htm

**13. Перечень информационных технологий**

а) лицензионное и свободно распространяемое программное обеспечение:

– Microsoft Office Standart 2013 Russian: пакет программ. Включает приложения: MS Office Word, MS Office Excel, MS Office PowerPoint, MS Office On-eNote, MS Office Publisher, MS Outlook, MS Office Web Apps (Word Excel MS PowerPoint Outlook);

– публично доступные облачные технологии (Google Docs, Яндекс диск и т.п.);

– Python – https://pytorch.org/ ;

– Anaconda – https://www.anaconda.com/

– Jupyter notebook – https://jupyter.org/

– R – https://www.r-project.org/

– R Studio – https://www.rstudio.com/.

– JASP - https://jasp-stats.org/.

б) информационные справочные системы:

– Электронный каталог Научной библиотеки ТГУ – https://koha.lib.tsu.ru/

– Электронная библиотека (репозиторий) ТГУ – http://vital.lib.tsu.ru/vital/access/manager/Index

– ЭБС Лань – http://e.lanbook.com/

– ЭБС Консультант студента – http://www.studentlibrary.ru/

– Образовательная платформа Юрайт – https://urait.ru/

– ЭБС ZNANIUM.com – https://znanium.com/

– ЭБС IPRbooks – http://www.iprbookshop.ru/

**14. Материально-техническое обеспечение**

Занятия по учебной дисциплине проводятся с использованием дистанционных образовательных технологий. Каждый обучающийся обеспечен доступом к образовательной платформе https://edu.data-diving.ru/.

**15. Информация о разработчиках**

Кабанова Татьяна Валерьевна, кандидат физико-математических наук, доцент кафедры теории вероятностей и математической статистики ИПМКН ТГУ.