Министерство науки и высшего образования Российской Федерации

Томский государственный университет
систем управления и радиоэлектроники

Н. Э. Лугина

Теория вероятностей и математическая статистика

Методические указания к лабораторным работам
и организации самостоятельной работы для студентов направления
«Программная инженерия»
(уровень бакалавриата)

Томск
2022

УДК 519.2
ББК 22.17
Л83

Рецензент:
Сидоров А. А., заведующий кафедрой автоматизации обработки информации Томского государственного университета
систем управления и радиоэлектроники, канд. экон. наук, доцент

Лугина, Наталья Эдуардовна
Л83 Теория вероятностей и математическая статистика: методические указания к лабораторным работам и организации самостоятельной работы для студентов направления «Программная инженерия» (уровень бакалавриата) / Н. Э. Лугина. – Томск : Томск. гос. ун-т систем упр. и радиоэлектроники, 2022. – 49 с.

Курс «Теория вероятностей и математическая статистика» формирует у студентов понятия, знания и компетенции, позволяющие строить и анализировать модели реального мира с помощью вероятностно-статистических методов. Лабораторные работы, закрепляющие практические навыки, ориентированы на использование математического пакета MathCad и электронных таблиц Excel.

Для студентов высших учебных заведений, обучающихся по направлению подготовки «Программная инженерия».

Одобрено на заседании кафедры АОИ, протокол № 1 от 18.01.2022

УДК 51922
ББК 22.17

ОГЛАВЛЕНИЕ

1 ВВЕДЕНИЕ 4
2 МЕТОДИЧЕСКИЕ УКАЗАНИЯ К ЛАБОРАТОРНЫМ РАБОТАМ 5
2.1 Лабораторная работа «Описательная статистика» 5
2.2 Лабораторная работа «Проверка статистических гипотез» 7
2.3 Лабораторная работа «Метод Монте-Карло» 8
2.4 Лабораторная работа «Дисперсионный анализ» 11
2.5 Лабораторная работа «Корреляционный анализ» 13
2.6 Лабораторная работа «Регрессионный анализ» 19
2.7 Лабораторная работа «Временные ряды» 24
2.8 Лабораторная работа «Цепи Маркова» 26
3 МЕТОДИЧЕСКИЕ УКАЗАНИЯ К САМОСТОЯТЕЛЬНОЙ РАБОТЕ 29
3.1 Теоретическая подготовка 29
3.2 Подготовка к лабораторным работам 31
3.3 Подготовка к промежуточной аттестации 31
4 РЕКОМЕНДУЕМЫЕ ИСТОЧНИКИ 39
ПРИЛОЖЕНИЕ А 40
ПРИЛОЖЕНИЕ Б 41
ПРИЛОЖЕНИЕ В 43
ПРИЛОЖЕНИЕ Г 44
ПРИЛОЖЕНИЕ Д 45
ПРИЛОЖЕНИЕ Е 46
ПРИЛОЖЕНИЕ Ж 47
ПРИЛОЖЕНИЕ З 48
ПРИЛОЖЕНИЕ И 49

1 ВВЕДЕНИЕ

Курс «Теория вероятностей и математическая статистика» ориентирован на формирование у студентов представлений о понятиях, которые в большинстве своем являются формализацией понятий, взятых из реальной жизни. Основное внимание уделяется математическим методам построения вероятностных моделей, изучению основ статистического описания данных, постановок и методов решения фундаментальных задач математической статистики и статистическим выводам в рамках данных моделей. Для формирования соответствующих компетенций курс содержит лабораторные работы, самостоятельную работу по изучению теоретического материала и подготовку к лабораторным работам, самостоятельную подготовку к промежуточной аттестации.

Цели лабораторных работ:

– отработка навыков применения изученных моделей и методов при решении практических (ситуационных) задач;

– закрепление умения использования расчетными формулами, теоремами, таблицами при решении статистических задач;

– формирование способности и готовности применять статистические методы при обработке результатов измерений и алгоритмов;

– получение опыта использования современных математических программных пакетов при представлении, анализе и интерпретации статистических данных;

– получение навыка представления результатов исследования и их анализа в виде отчета.

Задание студенту (команде студентов) формулируется в терминах некоторой предметной области. Первый этап работы состоит в формализации задачи, выборе метода решения, установлении последовательности шагов решения. Второй этап лабораторной работы состоит в выборе программного инструментария для решения задачи и реализации задачи на выбранном или рекомендованном преподавателем программном средстве. Третий этап работы заключается в анализе полученных результатов, оформлении отчета и защите лабораторной работы.

Отчет по лабораторной работе должен содержать следующие элементы:

– титульный лист;

– вариант задания;

– расчетные формулы и результаты расчетов;

– графический и справочный материал;

– выводы по работе.

При оценке лабораторной работы учитывается содержание отчета, правильность применения расчетных формул, а также качество выполнения и оформления, своевременность сдачи и умение студента обосновывать и защищать сделанные выводы.

Целями самостоятельной работы является систематизация, закрепление и расширение теоретических знаний, приобретение навыков практической и исследовательской деятельности. Реализация этих целей достигается в процессе теоретической подготовки, подготовки к лабораторным работам, подготовки к промежуточной аттестации.

2 МЕТОДИЧЕСКИЕ УКАЗАНИЯ К ЛАБОРАТОРНЫМ РАБОТАМ

2.1 Лабораторная работа «Описательная статистика»

Цель работы

Знакомство с методами описательной статистики, получение навыков первоначальной обработки данных средствами Excel.

Форма проведения

Выполнение индивидуального задания средствами Excel. Варианты заданий выдаются преподавателем каждому студенту в виде файла.

Форма отчетности

Защита отчета. Отчет оформляется в виде файла Word.

Задание на лабораторную работу

Для выборочных данных своего варианта выполнить следующую обработку, пояснив полученные результаты:

1) по выборке найти среднее арифметическое, выборочную дисперсию, выборочное среднее квадратичное отклонение, проиллюстрировать выполнение правила «*3σ*»;

2) по вариационному ряду найти моду, медиану, размах выборки, оценить среднее квадратичное отклонение с помощью размаха;

3) найти верхнюю и нижнюю выборочные квартили, пояснить их смысл;

4) составить сгруппированный статистический ряд, оценить математическое ожидание и дисперсию по сгруппированному ряду, сравнить эти значения с найденными по выборке;

5) построить гистограмму; найти модальный интервал и сравнить его середину с найденным по вариационному ряду значением моды;

6) составить сгруппированный статистический ряд с накопленными частотами, построить эмпирическую функцию распределения; найти медианный интервал, сравнить середину интервала с найденным по вариационному ряду значением медианы.

Теоретические основы

Перед выполнением лабораторной работы повторите теоретический материал о выборочном методе математической статистики ([1], глава 1).

Замечания к пунктам 1–6 задания на лабораторную работу.

К пункту 1. Среднее арифметическое обозначается $\bar{x}$ и вычисляется по формуле $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot n_i$; выборочная дисперсия $\hat{D} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \cdot n_i$; выборочное среднее квадратическое отклонение $\sigma = \sqrt{D}$. Правило «трех сигма» выполняется для большинства унимодальных законов распределения, то есть для выборок из таких генеральных совокупностей: более 99 % выборочных значений лежат в интервале $\left(\bar{x} - 3\sigma;\ \bar{x} + 3\sigma\right)$. Появление выборочных значений за пределами указанного интервала может свидетельствовать об аномальности этих значений ([1], глава 3, п. 3.4), либо о том, что генеральная совокупность не является нормальной.

К пунктам 2, 3. Оценка моды $\hat{M}$ – варианта с наибольшей частотой. Медиана делит выборку на две части: половина вариант меньше медианы, половина – больше. Можно найти три числа q_1, q_2, q_3, которые аналогичным образом делят выборку на четыре равные части. Эти числа называются квартилями. Число q_2 совпадает с медианой, q_1 называется нижней, а

q_3 – верхней квартилью. Размах $R = x_{(n)} - x_{(1)}$ равен разности наибольшей и наименьшей вариант. Этой характеристикой пользуются при работе с малыми выборками. Грубая оценка выборочного среднего квадратического отклонения σ может быть выполнена по формуле: $\sigma = \frac{R}{6}$.

К пункту 4. Сведения о синтаксисе функции ЧАСТОТА. Имя и параметры: ЧАСТОТА(*массив_данных*; *массив_карманов*).

Массив_данных – это выборочные данные, для которых рассчитываются частоты попадания в «карманы». Если массив данных не содержит значений, то функция ЧАСТОТА возвращает нули.

Массив_карманов – это массив правых концов интервалов, в которых *группируются значения массива данных.*

Функция ЧАСТОТА возвращает распределение частот в виде вертикального массива, причем количество элементов в возвращаемом массиве на единицу больше числа элементов в массиве карманов. Дополнительный элемент в возвращаемом массиве содержит количество значений, больших, чем правая граница последнего интервала. Для вызова функции сначала выделяют область, куда попадут результаты вычисления, задают значения аргументов, затем выходят нажатием сочетания клавиш *Ctrl+Shift+Enter*.

К пункту 5. Для сгруппированной выборки находят модальный интервал – интервал с наибольшей частотой. В качестве моды можно взять середину модального интервала.

К пункту 6. Для нахождения медианы по сгруппированному ряду накопленных частот выделяют промежуток, на котором накопленная частота становится больше или равной 0,5 (а левее этого промежутка – меньше 0,5). За медиану принимают середину найденного интервала.

Порядок выполнения работы

1) Прочитать задание на лабораторную работу, повторить понятия математической статистики из предыдущего семестра.
2) Записать основные расчетные формулы; составить предварительный отчет.
3) Для пунктов задания 1), 2), 3) найти выборочные характеристики, используя встроенные статистические функции: СРЗНАЧ; МАКС; МИН; МЕДИАНА; КВАРТИЛЬ и прочие.
4) Для пункта задания 5) использовать встроенную функцию ЧАСТОТА; построить гистограмму (в меню: Мастер диаграмм).
5) Сравнить результаты оценок числовых характеристик по вариационному ряду и по сгруппированной выборке, сделать выводы.
6) Оформить отчет. Защитить работу перед преподавателем.

Варианты заданий

Варианты заданий формируются преподавателем для каждого студента индивидуально на основе данных Федеральной службы государственной статистики (https://rosstat.gov.ru) и выдаются студентам в виде файла или ссылки на соответствующую страницу сайта. Так же студентам предлагаются для анализа данные фитнес браслета за любой месяц года.

Контрольные вопросы

1) Что означают термины «генеральная совокупность» и «выборка»?
2) В чем суть выборочного метода?
3) Дайте определение оценки параметра распределения.
4) Какая оценка называется несмещенной?
5) Какая оценка называется состоятельной?
6) Приведите пример несмещенной и состоятельной оценки.
7) Приведите пример смещенной и состоятельной оценки.
8) Сравните понятия математического ожидания и среднего арифметического.

9) Поясните правило «*3σ*».
10) Что такое аномальные наблюдения?
11) Что такое «вариационный ряд»?
12) Какие оценки удобно находить по вариационному ряду?
13) Какие параметры указываются при обращении к функции КВАРТИЛЬ?
14) Поясните метод построения сгруппированного статистического ряда.
15) Как называется оценка плотности распределения, построенная по сгруппированному статистическому ряду?
16) Как называется оценка функции распределения, построенная по сгруппированному статистическому ряду?
17) Поясните, как найти медиану по сгруппированному статистическому ряду.
18) Поясните, чем вызвано расхождение числовых значений оценок, вычисленных по вариационному и по сгруппированному рядам.

2.2 Лабораторная работа «Проверка статистических гипотез»

Цель работы

Изучение алгоритмов проверки параметрических гипотез, проверки гипотез об однородности данных.

Форма проведения

Работа в парах. Решение ситуационных задач. Решение задачи о проверке гипотез о виде распределения. Решение задач о проверке гипотез об однородности данных. Выполнение заданий средствами пакета MathCad.

Форма отчетности

Защита отчета. Отчет оформляется в виде файла MathCad.

Теоретические основы

При подготовке к лабораторной работе повторите тему «Проверка статистических гипотез» по конспекту лекций или по литературе ([2], глава 19, параграфы 1 – 6). Справочный материал для проверки параметрических гипотез приведен в Приложении А, более подробно эта информация изложена в [2], глава 19, параграфы 8 – 13, 18, 19.

Задача на проверку гипотезы об однородности двух выборок рекомендуется решать, используя критерий знаков, а также критерий Вилкоксона ([2], глава 19, параграф 27).

Порядок выполнения работы

Лабораторная работа состоит из четырех задач:

1. Проверка статистических гипотез (ситуационная задача).
2. Критерий согласия Пирсона.
3. Критерий знаков.
4. Критерий Вилкоксона.

1) Прочитать условия задач, выяснить, какую гипотезу требуется проверить.
2) Сформулировать основную и альтернативную гипотезы
3) Уровень значимости задан по условию задачи.
4) С помощью Приложения А подобрать статистику и построить кривую распределения, используя встроенные функции пакета MathCad dnorm(x,0,1), dt(x,d), $dF(x,d_1,d_2)$.
5) Найти границы критической области, используя встроенные функции пакета MathCad qnorm(α+0.5, 0, 1), qt(1–α, k), $qF(1–\alpha, k_1, k_2)$. (Для задач 3, 4 использовать таблицы соответствующих критериев в печатном виде).
6) Рассчитать наблюдаемое значение статистики и принять статистическое решение. Сформулировать ответ на вопрос задачи в терминах условий задачи.

7) Оформить отчет и защитить его перед преподавателем.

Варианты заданий

Примерные варианты заданий (задача 1, задача 2) приводятся в Приложении Б. Для задачи 3, 4 преподавателем генерируются две выборки заданного объема из генеральной совокупности с известными параметрами.

Контрольные вопросы

1) Поясните термин «гипотеза».
2) Приведите примеры гипотез.
3) Как обозначаются основная и альтернативная гипотезы? Дайте определение ошибки первого рода.
4) Дайте определение ошибки второго рода.
5) Какую роль при проверке параметрических гипотез играет уровень значимости?
6) Какую роль при проверке параметрических гипотез играет альтернативная гипотеза?
7) Чем определяется размер критической области?
8) Чем определяется форма критической области?
9) Какую задачу решают критерии согласия? Назовите известные Вам критерии согласия.
10) Какую задачу решают критерии однородности? Назовите известные Вам критерии однородности.
11) Опишите критерий согласия Пирсона (постановка задачи, условия применения, принятие решения).
12) Опишите критерий знаков (постановка задачи, условия применения, принятие решения).
13) Опишите критерий Вилкоксона (постановка задачи, условия применения, принятие решения).

2.3 Лабораторная работа «Метод Монте-Карло»

Цель работы

Изучение метода Монте-Карло на примере вычисления определенного интеграла.

Форма проведения

Выполнение индивидуального задания средствами пакета MathCad.

Форма отчетности

Защита отчета. Отчет оформляется в виде файла MathCad.

Теоретические основы

При подготовке к лабораторной работе следует повторить темы «Равномерное распределение непрерывной случайной величины» и «Нормальное распределение непрерывной случайной величины» ([3], глава 2, параграф 30), а также формулировку закона больших чисел (ЗБЧ) и центральной предельной теоремы (ЦПТ) ([3], глава 4, параграфы 41, 42). С общей идеей метода Монте-Карло и алгоритмами моделирования случайных величин можно познакомиться по конспекту лекций или по литературе ([4], глава 21, параграфы 1-3; [2], глава 2, параграфы 5, 6).

Порядок выполнения работы

Лабораторная работа состоит из двух частей:

Часть 1 – Генерация случайных чисел указанным методом и исследование качества полученной последовательности (2 часа).

Часть 2 – Вычисление интеграла простейшим методом с заданной точностью и заданной доверительной вероятностью (2 часа).

Порядок выполнения *Части 1*:

1) Задать начальные значения на основании теории об алгоритме генерации, сгенерировать n=100 случайных чисел (нечетные варианты – методом Неймана, четные – методом Лемера).

2) Изобразить полученные числа в виде точечного графика на плоскости и оценить визуально качество последовательности.

3) Выполнить проверку на равномерность распределения путем сравнения статистических параметров, характерных для равномерного распределения, и частотный тест. Для построения гистограммы и удобства сравнения с «эталонным» генератором взять количество интервалов равным 10.

4) Оценить качество последовательности с помощью критерия Пирсона, используя статистику

$$K = \sum_{i=1}^{k} \frac{\left(n_i^* - np_i\right)^2}{np_i} \sim \chi^2_{(k-r-1)}.$$

Уровень значимости принять равным α=0,05.

5) Задать другие начальные значения и повторить п.1) – 4) три раза.

6) В отчёте в расчётной части представить результаты одного опыта, демонстрирующего достижение цели лабораторной работы, результаты всех опытов занести в таблицу:

Таблица 2.1 – Качество генерации

№ опыта	Начальные значения	Визуальная оценка	Кнабл	Ктабл	Вывод
1 2 3					

Порядок выполнения *Части 2*:

1) Привести интеграл к виду, требуемому для вычисления с помощью метода Монте-Карло.

$$I = \int_a^b h(x) \cdot f(x)\, dx,$$

где $f(x)$ – плотность распределения случайной величины $X \sim R(a,b)$.

2) Оценить количество опытов, необходимое для вычисления интеграла с заданной точностью и заданной доверительной вероятностью. Для оценки выборочного с.к.о. построить график функции $y = h(x)$. Для определения квантили, соответствующей заданной доверительной вероятности, использовать встроенную функцию $qnorm\left(\frac{\beta}{2} + 0.5, 0, 1\right)$ пакета MathCad.

3) Используя стандартный датчик *rnd* пакета MathCad, сгенерировать n случайных чисел $\gamma_i \in (0;1)$ и преобразовать их в $x_i = a + (b-a) \cdot \gamma_i$

4) Вычислить оценку $\hat{I}$ для интеграла I:

$$\hat{I} = \frac{1}{n} \sum_{i=1}^{n} h(x_i)$$

5) Вычислить оценку дисперсии

$$\hat{D} = \frac{1}{n}\sum_{i=1}^{n}\left(h^2(x_i)\right) - \hat{I}^2$$

6) Используя оценку дисперсии и зная количество опытов, применить определение доверительного интервала и центральную предельную теорему:

$$P\{\left|\hat{I} - I\right| < \varepsilon\} = 2\Phi\left(\frac{\varepsilon\sqrt{\mathrm{n}}}{\sigma}\right) = \beta,$$

затем найти точность расчета ε_0 из уравнения

$$2\Phi\left(\frac{\varepsilon_0\sqrt{\mathrm{n}}}{\sqrt{\hat{D}}}\right) = \beta$$

и сравнить с заданной точностью ε.

7) Провести вычисления несколько раз, добиваясь требуемой точности – увеличивая значение *n*.

8) Оформить отчет, записать результаты вычислений в таблицу.

Таблица 2.2 – Вычисление интеграла

№ опыта	Количество опытов	Оценка интеграла	Оценка дисперсии	Оценка точности ε_0	Точность ε
1 2 3					

9) Убедиться, что заданная точность вычислений достигнута. Ответ записать в виде доверительного интервала $I = \left(\hat{I} - \varepsilon; \hat{I} + \varepsilon\right)$.

Варианты заданий

Приводятся варианты заданий (Приложение В).

Контрольные вопросы

1) Какие задачи решаются методом Монте-Карло?
2) Перечислите способы получения случайных чисел.
3) Укажите достоинства и недостатки генерации случайных чисел с помощью таблиц.
4) Укажите достоинства и недостатки генерации случайных чисел с помощью физического датчика.
5) Укажите достоинства и недостатки генерации псевдослучайных чисел с помощью алгоритма.
6) Какие основные проблемы возникают при оценке качества генерации случайных чисел?
7) В чем заключается метод середин квадратов?
8) В чем заключается метод вычетов?
9) Определите период и запишите последовательность различных значений псевдослучайных чисел, полученных методом вычетов с начальными значениями $m_0 = 1$, M=11, K=14.
10) На каких теоремах основан метод Монте-Карло?
11) Расшифруйте сокращение «ЗБЧ» и сформулируйте соответствующую теорему.
12) Расшифруйте сокращение «ЦПТ» и сформулируйте соответствующую теорему.
13) Как точность вычислений зависит от числовых характеристик моделируемой случайной величины?
14) За счет чего можно повысить точность вычислений?
15) Сравните по точности простейший и геометрический методы Монте-Карло.

16) Что такое трудоемкость метода Монте-Карло?

17) Какую случайную величину будем генерировать для вычисления интеграла $I = \int_0^{\frac{\pi}{2}} \cos x \, dx$ простейшим методом Монте-Карло? Приведите этот интеграл к виду, необходимому для применения метода Монте-Карло.

18) Оцените количество опытов, необходимых для вычисления интеграла I с точностью ε=0,04 и доверительной вероятностью γ=0,95.

19) Изложите идею геометрического метода для этого интеграла. Поясните на рисунке.

20) Вычисление интеграла $\int_0^{\frac{\pi}{2}} \cos x \, dx$ производится методом Монте-Карло на основании 1000 испытаний. Какую наибольшую погрешность вычислений можно гарантировать с надежностью 97,22%?

2.4 Лабораторная работа «Дисперсионный анализ»

Цель работы

Изучение методов однофакторного дисперсионного анализа.

Форма проведения

Выполнение индивидуального задания средствами Excel.

Форма отчетности

Защита отчета. Отчет оформляется в виде файла Word.

Теоретические основы

Постановка задачи однофакторного дисперсионного анализа как частного случая статистической гипотезы приведена в [5], глава 11, параграф 11.1. Более подробное изложение можно найти в [2], глава 20.

В **классической** схеме однофакторного дисперсионного анализа на некоторый количественный признак X (случайную величину X) действует фактор F, имеющий r уровней. Дисперсионный анализ позволяет ответить на вопрос: влияет ли фактор F на измеряемый признак X? На каждом уровне проводится ряд наблюдений исследуемого признака, которые рассматриваются как независимые выборочные значения из генеральных совокупностей $X_1, X_2, ..., X_r$, распределенных по нормальному закону с одинаковыми, хотя и неизвестными дисперсиями, и математическими ожиданиями $m_1, m_2, ..., m_r$. При этом предполагается, что ошибки наблюдений распределены нормально с математическим ожиданием равным нулю и одинаковыми дисперсиями. Задача дисперсионного анализа формулируется как задача о равенстве всех математических ожиданий

$$H_o: m_1 = m_2 = ... = m_r,$$

$$H_1: m_i \neq m_j \text{ для некоторых } i, j.$$

Основная идея дисперсионного анализа состоит в переходе от задачи сравнения средних на всех уровнях к эквивалентной ей задаче сравнения дисперсий: «факторной» дисперсии, оценивающей разброс, вносимый в результате воздействия фактора F, и «остаточной» дисперсии, оценивающей разброс, возникший в результате случайных причин. Для этого сумма квадратов отклонений наблюдаемых значений от их общего среднего по всей таблице $\bar{x}$ раскладывается на две части (см. [3,5]) в виде основного дисперсионного тождества:

$$\sum_{i=1}^{r}\sum_{j=1}^{k_i}\left(x_{ij}-\bar{x}\right)^2=\sum_{i=1}^{r}k_i\left(\bar{x}_i-\bar{x}\right)^2+\sum_{i=1}^{r}\sum_{j=1}^{k_i}\left(x_{ij}-\bar{x}_i\right)^2$$

В этом тождестве левая часть обозначается $Q^2_{общ}$ и служит для оценки общего разброса в наблюдаемых данных, первое слагаемое в правой части Q^2_F – для оценки «факторного» разброса, второе слагаемое правой части Q^2_ε – для оценки «случайного» разброса в данных.

С помощью этих сумм квадратов находятся оценки факторной D_F и остаточной дисперсий D_ε, для чего соответствующая сумма квадратов делится на число степеней свободы: $\hat{D}_F=\frac{Q^2_F}{r-1}$, $\hat{D}_\varepsilon=\frac{Q^2_\varepsilon}{n-r}$, где n – общее количество наблюдений. Задача о равенстве средних на всех уровнях эквивалентна задаче о равенстве дисперсий:

$$H_0:\ D_F=D_\varepsilon\ ,$$

$$H_1: D_F>D_\varepsilon\ .$$

Здесь основная гипотеза утверждает, что факторная дисперсия отличается от остаточной незначимо, то есть разброс, вносимый фактором, практически не отличается от разброса, обусловленного случайными причинами, следовательно, и средние значения на разных уровнях отличаются незначимо; поэтому в исходной постановке справедлива гипотеза H_0. Если же верной является альтернативная гипотеза, то влияние фактора значимо отличается от случайного, и в исходной постановке справедлива гипотеза H_1.

Для сравнения дисперсий рассматривается статистика $K=\frac{\hat{D}_F}{\hat{D}_\varepsilon}$, имеющая распределение Фишера $F_{(r-1,n-r)}$. Критическая область является правосторонней, поэтому если наблюдаемое значение статистики *Кнабл* превышает или равно табличному *Ктабл*, то основная гипотеза отвергается в пользу альтернативной, то есть влияние фактора на результирующий признак значимо.

Классическая схема однофакторного дисперсионного анализа требует нормальности распределения исследуемого признака и ошибок наблюдений. Если эти условия не выполняются, то для решения задачи о равенстве средних более предпочтительными являются непараметрические методы, которые позволяют проверить гипотезу о равенстве средних с минимальными требованиями к выборочным данным: предполагается, что ошибки наблюдений независимы и имеют непрерывное распределение.

При проверке гипотезы H_0 с помощью **критерия Крускала-Уоллиса** выборочные значения x_{ij} заменяются их рангами x'_{ij}. Напомним, что ранг – это число, соответствующее порядковому номеру наблюдаемого значения в данной выборке, если наблюдаемые значения расположить по возрастанию. Для каждого уровня вычисляется средний ранг $R_i=\frac{1}{k_i}\sum_{j=1}^{k_i}x'_{ij}$, $i=\overline{1,\ r,}$ и сравнивается с общим средним рангом, который, в предположении справедливости гипотезы H_0, равен $R=\frac{n+1}{2}$. Сравнение проводится по статистике Крускала-

Уоллиса $K = \frac{12}{n(n+1)} \sum_{i=1}^{r} k_i \left(R_i - R\right)^2$, критические точки которой зависят от количества наблюдений n, количества уровней r и количества наблюдений k_i на каждом уровне. Критическая область в данной задаче является правосторонней, поэтому если $K_{набл} \geq K_{табл}$, то гипотеза H_0 отвергается. При больших значениях n распределение статистики Крускала-Уоллиса приближается к распределению χ^2_{r-1}.

Порядок выполнения работы

1) Получить у преподавателя таблицу наблюдений (строки – уровни фактора).
2) Изучить теоретическую часть [2, 5]. Ответить на вопросы.
3) Решить задачу по классической схеме. Уровень значимости принять равным α=0,05.
4) Решить эту же задачу методом Крускала-Уоллиса. Уровень значимости принять равным α=0,05 (таким же, как и при решении задачи по классической схеме).
5) Проверить результат с помощью встроенного пакета «Статистика» Excel. Проанализировать результаты.
6) Написать отчет и защитить его перед преподавателем. Для отчета придумать содержательную постановку задачи дисперсионного анализа, указав, что является измеряемой величиной, а что – фактором, и какие значения принимают уровни фактора.

Варианты заданий

Приводятся варианты заданий (Приложение Г).

Контрольные вопросы

1) Какой вид имеет таблица наблюдений дисперсионного анализа?
2) Какие требования предъявляются к экспериментальным данным в классической схеме дисперсионного анализа?
3) Сформулируйте основную и альтернативную гипотезы дисперсионного анализа (о средних).
4) Сформулируйте гипотезу, эквивалентную основной гипотезе дисперсионного анализа (о дисперсиях).
5) В чем основная идея дисперсионного анализа?
6) Как оценивается межгрупповой разброс данных?
7) Как оценивается внутригрупповой разброс данных?
8) Запишите основное дисперсионное тождество.
9) Какая статистика используется при проверке гипотезы о дисперсиях?
10) В каких случаях применяется непараметрический анализ (какие требования предъявляются к экспериментальным данным)?
11) В чем состоит метод Крускала-Уоллиса?
12) Каким распределением аппроксимируется статистика Крускала-Уоллиса при большом объеме выборки?

2.5 Лабораторная работа «Корреляционный анализ»

Цель работы

Знакомство с числовыми коэффициентами, предназначенными для выявления связи между двумя переменными; оценка силы корреляционной связи.

Форма проведения

Выполнение индивидуального задания средствами MathCad.

Форма отчетности

Защита отчета. Отчет оформляется в виде файла MathCad.

Теоретические основы

Перед выполнением лабораторной работы рекомендуется повторить определение и свойства коэффициента корреляции Пирсона ([3], глава 3, параграф 36; [2], глава 14, параграф 17). Полезно сравнить свойства коэффициента корреляции Пирсона и рангового коэффициента корреляции Спирмена ([4], глава 17, параграф 10; [2], глава 19, параграф 25). Ознакомьтесь с формой представления сгруппированных двумерных наблюдений (корреляционные таблицы), с выборочным корреляционным отношением и его свойствами ([2], глава 18, параграфы 5, 11 – 13).

Расчетные формулы для несгруппированных данных

Часть 1. Для несгруппированных данных фактор – отклик расчеты выполняются в MathCad с использованием встроенных функций: Rank (x), Rank(y), mean(x), mean(y), corr(x,y), Spear(x,y).

Коэффициент корреляции Пирсона:

Точечной оценкой коэффициента корреляции r_{xy} является выборочный коэффициент корреляции r_{xy} , который можно рассчитать по формуле

$$r_{xy} = \frac{\sum\left(x_i - \bar{x}\right)\cdot\left(y_i - \bar{y}\right)}{\sqrt{\left(x_i - \bar{x}\right)^2}\cdot\sqrt{\left(y_i - \bar{y}\right)^2}}.$$

Эта точечная оценка для двумерной генеральной совокупности

- несмещенная;
- асимптотически эффективная.

Или, если выполнить тождественные преобразования числителя:

$$r_{xy} = \frac{\frac{1}{n}\sum_i x_i \cdot y_i - \bar{x}\cdot\bar{y}}{s_x \cdot s_y}$$

где $\left(x_i, y_i\right)$, $i = 1, .2, n$ – независимая выборка объема n из двумерной генеральной совокупности;

$\bar{x}, \bar{y}$ – средние арифметические значения (выборочные средние) переменных X и Y;

s_x, s_y – выборочные средние квадратические отклонения переменных X и Y.

Расчет выборочного коэффициента корреляции r_{xy} в MathCad можно выполнить по формуле

$$r_{xy} = \frac{\sum_{i=1}^{n}\left(x_i \cdot y_i\right) - n\cdot mean\left(x\right)\cdot mean\left(y\right)}{\sqrt{\left[\sum_{i=1}^{n} x_i^2 - n\cdot\left(mean\left(x\right)\right)^2\right]\cdot\left[\sum_{i=1}^{n} y_i^2 - n\cdot\left(mean\left(y\right)\right)^2\right]}}$$

Проверьте результат расчета, используя встроенную функцию corr(x,y). Результаты расчетов выборочного коэффициента корреляции $\tilde{r}_{xy}$ по различным формулам должны совпасть.

По вычисленному значению выборочного коэффициента корреляции $\tilde{r}_{xy}$ можно предположить

- слабую связь;
- умеренную связь;
- заметную связь;

- достаточно тесную связь;
- тесную связь;
- весьма тесную связь

(чем ближе $|r_{xy}| \to 1$, тем теснее связь). Затем, посмотрев на знак выборочного коэффициента корреляции r_{xy}, можно предположить наличие линейной положительной корреляционной зависимости между X и Y (если $r_{xy} > 0$) или, линейную отрицательную корреляционную зависимость между X и Y (если $r_{xy} < 0$). Далее формулируем альтернативную гипотезу: $H_1 : r_{xy} > 0 \ (< 0)$ – коэффициент корреляции значим, переменные X и Y связаны (здесь указываем) линейной положительной (линейной отрицательной) зависимостью. Уровень значимости принять равным α=0,01.

Проверка значимости коэффициента корреляции Пирсона проводится с помощью статистики

$$K_{набл} = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-\tilde{r}_{xy}^2}},$$

$$K_{табл} \sim T_{(\alpha;n-2)}.$$

Используйте встроенную функцию для квантили распределения Стьюдента qt(1–alfa,n–2). Используйте односторонние критические области.

Коэффициент корреляции Спирмена:

$$\tilde{r}_s = 1 - \frac{6 \cdot \sum \left(x_i' - y_i'\right)^2}{n^3 - n},$$

где $\left(x_i' - y_i'\right)$ – разность значений рангов несгруппированных данных фактора и отклика.

Выполните проверку расчетов при помощи встроенной функции Spear(x,y).

Проверка значимости коэффициента корреляции Спирмена проводится с помощью той же статистики, что и для коэффициента Пирсона:

$$K_{набл} = \frac{r_s\sqrt{n-2}}{\sqrt{1-\tilde{r}_s^2}},$$

$$K_{табл} \sim T_{(\alpha;n-2)}.$$

Итак, при проверке значимости выборочного коэффициента корреляции Пирсона принимается статистическое решение о существовании между переменными X и Y **линейной** положительной (отрицательной) **корреляционной связи.** При проверке значимости рангового коэффициента корреляции Спирмена принимается статистическое решение о существовании между переменными X и Y **монотонной** прямой (обратной) **корреляционной связи.**

Расчетные формулы для сгруппированных данных

Часть 2. При чтении корреляционной таблицы проводится предварительный анализ характера зависимости: по степени заполненности клеток таблицы условными частотами можно судить о тесноте связи. Если клетки таблицы заполнены только вокруг диагонали таблицы, то имеется относительно тесная связь между переменными. Если условные частоты содержатся почти во всех клетках таблицы, то это свидетельствует о большом рассеянии значений переменных и, следовательно, зависимость между ними проявляется слабо.

Например, по данной корреляционной таблице можно предположить существование линейной положительной корреляционной зависимости между *X* и *Y*: с ростом *X*, *Y* имеет тенденцию в среднем возрастать.

	Y					
X	10	15	20	25	30	35
14	1	5				
24		5	3			
34			9	40	2	
44			4	11	6	
54				4	7	3

При выполнении Части 2 в MathCad удобно перейти к матричной записи данных. Так, для данной корреляционной таблицы, значения переменных X и Y, а также условные частоты, представляют собой матрицы-столбцы:

$$x := \begin{pmatrix} 14 \\ 24 \\ 34 \\ 44 \\ 54 \end{pmatrix}; \quad nx := \begin{pmatrix} 6 \\ 8 \\ 51 \\ 21 \\ 14 \end{pmatrix}; \quad y := \begin{pmatrix} 10 \\ 15 \\ 20 \\ 25 \\ 30 \\ 35 \end{pmatrix}; \quad y := \begin{pmatrix} 1 \\ 10 \\ 16 \\ 55 \\ 15 \\ 3 \end{pmatrix}.$$

Матрица частот совместного распределения – матрица размера (m x k) в данном случае имеет размер (5x6)

$$nxy := \begin{pmatrix} 1 & 5 & 0 & 0 & 0 & 0 \\ 0 & 5 & 3 & 0 & 0 & 0 \\ 0 & 0 & 9 & 40 & 2 & 0 \\ 0 & 0 & 4 & 11 & 6 & 0 \\ 0 & 0 & 0 & 4 & 7 & 3 \end{pmatrix}.$$

Для корреляционной таблицы имеют место следующие соотношения:

$$\sum_{i=1}^{m}\sum_{j=1}^{k} nxy_{ij} = n, \quad \sum_{i=1}^{m} nx_i = n, \quad \sum_{j=1}^{k} ny_j = n.$$

Средние вычисляют по формулам:

$$\bar{x} = \frac{\sum_{i=1}^{m} x_i \cdot nx_i}{n}, \quad \bar{y} = \frac{\sum_{j=1}^{k} y_j \cdot ny_j}{n}.$$

Групповые средние вычисляют по формулам:

$$\bar{x}_j = \frac{\sum_{i=1}^{m} x_i \cdot nxy_{ij}}{ny_j}, \quad \bar{y}_i = \frac{\sum_{j=1}^{k} y_j \cdot nxy_{ij}}{nx_i}.$$

где nxy_{ij} – частоты совместного наблюдения значений *X* и *Y*,

nx_i – безусловное распределение частот переменной *x* (суммируем по строкам);

ny_j – безусловное распределение частот переменной *y* (суммируем по столбцам).

Чтобы изобразить графически поле корреляции точками на координатной плоскости *xOy*, запишем массив данных (это могут быть и два отдельных массива) по корреляционной таблице – множество точек $\left(x_i, y_j\right)$. На этом же рисунке постройте эмпирическую линию регрессии *Y* по *X* – вычислите групповые средние и изобразите их графически в виде ломаной, соединяющей точки $\left(x_i, \bar{y}_i\right)$. Сделайте выводы. Обязательно проверяйте правильность расчета ysr_i, ставя знак «=». Например, «ysr_i=». Для данной корреляционной таблицы должно быть выведено на экран 5 значений ysr_i. Следите за тем, чтобы не было переопределений переменных в MathCad. Наличие «черных» квадратиков в файле MathCad при выполнении лабораторной работы считается ошибкой.

	x_i	y_j
1	14	10
2	14	15
3	24	15
4	24	20
5	34	20
6	34	25
7	34	30
8	44	20
9	44	25
10	44	30
11	54	25
12	54	30
13	54	35

Коэффициент корреляции Пирсона:

Расчет коэффициента Пирсона выполните по формуле

$$\tilde{r}_{xy} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{k} n_{ij} x_i y_j - n \cdot \bar{x} \cdot \bar{y}}{\sqrt{\left[\sum_{i=1}^{m}\left[n\left(X = x_i\right) \cdot x_i^2\right] - n \cdot \bar{x}^2\right] \cdot \left[\sum_{j=1}^{k}\left[n\left(Y = y_j\right) \cdot y_j^2\right] - n \cdot \bar{y}^2\right]}}$$

Для примера корреляционной таблицы в среде MathCad эта формула примет вид

$$\tilde{r}_{xy} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{k} nxy_{i,j} \cdot x_i \cdot y_j - n \cdot xsr \cdot ysr}{\sqrt{\left[\sum_{i=1}^{m}\left[nx_i \cdot x_i^2\right] - n \cdot xsr^2\right] \cdot \left[\sum_{j=1}^{k}\left[ny_j \cdot y_j^2\right] - n \cdot ysr^2\right]}}$$

Расчет коэффициента Пирсона также может быть выполнен по формуле:

$$\tilde{r}_{xy} := \frac{n \cdot \sum_{i=1}^{m}\sum_{j=1}^{k}\left(nxy_{i,j} \cdot x_i \cdot y_j\right) - \left[\sum_{i=1}^{m} x_i \cdot nx_i\right] \cdot \left[\sum_{j=1}^{k} y_j \cdot ny_j\right]}{\sqrt{n \cdot \sum_{i=1}^{m}\left[nx_i \cdot x_i^2\right] - \left[\sum_{i=1}^{m} x_i \cdot nx_i\right]^2} \cdot \sqrt{n \cdot \sum_{j=1}^{k}\left[ny_j \cdot y_j^2\right] - \left[\sum_{j=1}^{k} y_j \cdot ny_j\right]^2}}.$$

Формулы в MathCad пишутся в явном символьном виде. Результаты расчетов должны совпасть.

Статистика для проверки значимости коэффициента корреляции Пирсона:

$$\frac{r_{xy}\sqrt{n-2}}{\sqrt{1-\tilde{r}_{xy}^2}} \sim T_{(\alpha;n-2)}$$

Используйте встроенную функцию для квантили распределения Стьюдента: qt(1–alfa,n–2). Используйте односторонние критические области.

Эмпирическое корреляционное отношение *Y* по *X*.

Расчет корреляционного отношения *Y* по *X* выполняется по формуле

$$\eta_{yx} = \sqrt{\frac{s^2_{межгрупп}}{s^2_{общая}}},$$

где межгрупповая дисперсия переменной *y*

$$s^2_{межгрупп} = \frac{1}{n}\sum_{i=1}^{m}\left(\bar{y}_{x_i} - \bar{y}\right)^2 n_{x_i},$$

общая дисперсия переменной

$$s^2_{общая} = \frac{1}{n}\sum_{j=1}^{k}\left(y_j - \bar{y}\right)^2 n_{y_j}.$$

Межгрупповая дисперсия выражает ту часть вариации *Y*, которая обусловлена изменчивостью *X*. Используя принятые ранее обозначения для решения задачи в среде MathCad, формула примет вид:

$$\eta_{yx} = \sqrt{\frac{\sum_{i=1}^{m}\left(ysr_i - ysr\right)^2 \cdot nx_i}{\sum_{j=1}^{k}\left(y_j - ysr\right)^2 \cdot ny_j}}.$$

Статистика для проверки гипотезы о проверке значимости корреляционного отношения:

$$\frac{\eta^2_{yx}\cdot(n-m)}{\left(1-\eta^2_{yx}\right)\cdot(m-1)} \sim F\left(m-1, n-m\right)$$

Используйте встроенную функцию для квантили распределения Фишера-Снедекора qF(1–alfa,m–1,n–m).

Статистика для проверки гипотезы о линейности связи:

$$\frac{\left(\eta^2_{yx} - \tilde{r}^2_{yx}\right)\cdot(n-m)}{\left(1-\eta^2_{yx}\right)\cdot(m-2)} \sim F\left(m-2, n-m\right)$$

Используйте встроенную функцию для квантили распределения Фишера-Снедекора qF(1–alfa,m–2,n–m).

Порядок выполнения работы

Лабораторная работа состоит из двух частей:

Часть 1 – Измерение корреляции несгруппированных данных.

Часть 2 – Исследование корреляции сгруппированных данных.

Порядок выполнения *Части 1*:

1) Получить набор несгруппированных данных (фактор – отклик).
2) Рассчитать коэффициент корреляции Спирмена. Выполнить проверку расчетов при помощи встроенной функции Spear(x,y). Проверить значимость коэффициента корреляции Спирмена.
3) Измерить связь между переменными с помощью коэффициента корреляции Пирсона. Выполнить проверку расчетов при помощи встроенной функции corr(x, y). Проверить значимость коэффициента корреляции Пирсона.
4) Сравнить значения коэффициентов.

Порядок выполнения *Части 2*:

1) Получить корреляционную таблицу.
2) Построить поле корреляции на плоскости *xOy*. На этом же рисунке построить эмпирическую линию регрессии Y по X. Сделать выводы.
3) Найти оценку коэффициента корреляции Пирсона. Проверить значимость коэффициента корреляции Пирсона.
4) Найти выборочное корреляционное отношение. Проверить значимость корреляционного отношения.
5) Проверить гипотезу о линейности корреляционной связи.
6) Проанализировать результаты.
7) Написать отчет и защитить его перед преподавателем.

Варианты заданий

Приводятся варианты заданий для несгруппированных данных (Приложение Д) и для сгруппированных данных (Приложение Е).

Контрольные вопросы

1) Какая зависимость величины *Y* от *X* называется функциональной?
2) Какая зависимость величины *Y* от *X* называется стохастической или вероятностной?
3) Назовите основные задачи корреляционного анализа.
4) Какие значения может принимать коэффициент корреляции Спирмена?
5) При каких условиях применяется коэффициент корреляции Спирмена?
6) Перечислите свойства коэффициента корреляции Спирмена.
7) При каких условиях применяется коэффициент корреляции Пирсона?
8) Как зависит сила линейной связи между переменными от величины коэффициента корреляции?
9) Перечислите свойства коэффициента корреляции Пирсона.
10) Что такое поле корреляции?
11) Какие выводы могут быть сделаны по полю корреляции и линии групповых средних?
12) Как проверить значимость коэффициента корреляции?
13) Что такое корреляционное отношение?
14) Какими свойствами оно обладает?
15) Как проверяется линейность связи с помощью сравнения коэффициента корреляции и корреляционного отношения?

2.6 Лабораторная работа «Регрессионный анализ»

Цель работы

Получение навыка построения и оценки качества регрессионных моделей.

Форма проведения

Выполнение индивидуального задания средствами MathCad.

Форма отчетности

Защита отчета. Отчет оформляется в виде файла MathCad.

Теоретические основы

Теоретический материал приведен в [1], глава 4; [2], глава 18, параграфы 1 – 4; [5], глава 13). При подготовке к лабораторной работе необходимо уяснить разницу между функциональной и стохастической зависимостью ([2], глава 18, параграф 1) и познакомиться с основным методом построения регрессионных моделей – методом наименьших квадратов.

Предположим, что нам необходимо описать в виде некоторой функции взаимосвязь двух переменных *X* и *Y* (*X* – фактор, независимая переменная; *Y* – отклик, зависимая переменная): *Y=f(X).* По результатам наблюдений мы можем оценить эту зависимость приближенно (в силу воздействия неучтенных факторов, случайных причин, ошибок измерения): $y = f(x) + \varepsilon$, где ε – случайная переменная, называемая возмущением. Предполагается, что среднее значение возмущения равно нулю: $M[\varepsilon] = 0$. При этом для каждого значения *X=x* имеем случайную переменную *Y* со средним значением (математическим ожиданием) $f(x) = M[Y \mid X = x]$. Функция $f(x)$ называется ***функцией регрессии*** случайной переменной *Y* на *X*, а график этой функции – ***линией регрессии***. Уравнение регрессии позволяет определить, каким в среднем будет значение отклика *Y* при том или ином значении фактора *X*.

Исследуемые явления, как правило, определяются большим числом одновременно и совокупно действующих факторов. В связи с этим возникает задача исследования одной зависимой переменной Y от нескольких объясняющих переменных $x_1, x_2, \ldots, x_n$, где *Y* – отклик, $y = \varphi(x_1, x_2, \ldots, x_n)$, $x_1, x_2, \ldots, x_n$ – факторы. Тогда модель множественной линейной регрессии можно представить в виде двух частей – слагаемых:

$$y = \varphi(x_1, x_2, \ldots, x_n) + \varepsilon,$$

где одна часть модели закономерно зависит от факторов, то есть является функцией факторов, а другая часть – случайна по отношению к факторам. Слагаемое ε выражает собой либо внутренне присущую отклику изменчивость, либо влияние неучтенных факторов, либо того и другого вместе; ε – ошибка эксперимента.

Для оценки неизвестных параметров множественной линейной регрессии будем применять метод наименьших квадратов, то есть выясним при каких параметрах достигается минимум выражения

$$\sum (y_{набл} - y_{модель})^2 \to \min.$$

Введём обозначения:

$Y = (y_1 \quad y_2 \quad \ldots \quad y_n)^T$ – матрица-столбец, или вектор, значений зависимой переменной размера $(n \times 1)$ (*n* строк, 1 столбец);

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$ – матрица значений объясняющей переменной или матрица плана размера $(n \times (k+1))$ (*n* строк, *k*+1 столбцов). В матрицу *X* дополнительно введен столбец, все элементы которого равны единице, то есть условно полагается, что в модели регрессии свободный член b_0 умножается на фиктивную переменную $x_{i0} \equiv 1,\ i = 1, 2, \ldots, n$.

$b = (b_0 \quad b_1 \quad \cdots \quad b_k)^T$ – матрица-столбец, или вектор, параметров размера $((k+1) \times 1)$ (*(k+1)* строк, 1 столбец);

$\varepsilon = \begin{pmatrix} \varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_n \end{pmatrix}^T$ – матрица-столбец, или вектор, возмущений (случайных ошибок, остатков) размера $(n \times 1)$ (n строк, 1 столбец).

Тогда в матричной форме модель регрессии примет вид:

$$y = X \cdot b + \varepsilon$$

Для оценки вектора неизвестных параметров b применим метод наименьших квадратов. Условие минимизации остаточной суммы квадратов запишется в виде:

$$Q(b) = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon^T \cdot \varepsilon .$$

На основании необходимого условия экстремума функции многих переменных вектор частных производных должен обратиться в нуль. Отсюда получим ***систему нормальных уравнений*** для определения параметров множественной линейной регрессии

$$\nabla_b Q(b) = 0 \Rightarrow X^T X b = X^T y .$$

Так как матрица $X^T X$ является невырожденной, то у неё существует ей обратная, следовательно, оценка вектора b неизвестных параметров примет вид:

$$\tilde{b} = \left(X^T \cdot X \right)^{-1} \cdot X^T \cdot y .$$

Эта формула позволяет найти оценки коэффициентов множественной линейной регрессии в матричной форме.

Перейдем теперь к ***оценке значимости*** коэффициентов регрессии $\tilde{b}_i$ и построению ***доверительных интервалов*** для параметров регрессионной модели $\tilde{b}_i$.

Проверим значимость каждого коэффициента регрессии: основная гипотеза $H_0 : b_i = 0$ – коэффициент регрессии незначим; альтернативная гипотеза $H_1 : b_i \neq 0$ – коэффициент регрессии значим.

Статистика

$$\frac{b_i - \tilde{b}_i}{S_{b_i}} \sim T_{(n-(k+1))}$$

имеет распределение Стьюдента с (n –(k+1)) степенями свободы. Поэтому b_i **значимо отличаются от нуля** на уровне значимости α, если

$$|t| = \frac{|\tilde{b}_i|}{S_{\tilde{b}_i}} > T_{\left(1-\frac{\alpha}{2};n-(k+1)\right)},$$

где $\tilde{b}_i$ – оценки значений коэффициентов регрессии;

$S_{\tilde{b}_i}$ – среднее квадратическое отклонение (стандартная ошибка) коэффициента регрессии b_i

$$S_{\tilde{b}_i} = S_\varepsilon \cdot \sqrt{C_{ii}}, \; i = \overline{1, k+1},$$

Оценка остаточной дисперсии S_ε^2 определяется по формуле:

$$S_\varepsilon^2 = \frac{Q(b)}{n-(k+1)} = \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n-(k+1)}$$

или

$$S_\varepsilon^2 = \frac{1}{n-(k+1)}\sum_{i=1}^{n}\left(y_i - \left(\tilde{b}_0 + \tilde{b}_1 x_{i1} + \tilde{b}_2 x_{i2} + \cdots + \tilde{b}_k\right)\right)^2;$$

C_{ii} – главные диагональные элементы матрицы C

$$C = \left(X^T \cdot X\right)^{-1};$$

$t_{табл}$ – квантиль распределения Стьюдента, $t_{табл} := qt\left(1-\frac{\alpha}{2}, n-(k+1)\right)$.

Если коэффициент регрессии b_i **значимо отличается от нуля** на уровне значимости α, то в этом случае и доверительный интервал **не накроет значение 0**. Убедитесь в совпадении результатов расчётов.

Построим $\gamma = (1-\alpha)\%$-ный доверительный интервал для параметров b_i. Доверительный интервал для параметров b_i имеет вид:

$$b_i = \tilde{b}_i \pm t_{табл} \cdot S_{\tilde{b}_i}.$$

Проверить адекватность модели – значит установить, соответствует ли математическая модель, выражающая зависимость между переменными, экспериментальным данным; **и** достаточно ли включенных в уравнение регрессии объясняющих переменных (одной или нескольких) для описания зависимой переменной.

Проверка значимости уравнения регрессии проводится на основании дисперсионного анализа как вспомогательного средства для изучения качества регрессионной модели.

Сформулируем основную гипотезу $H_0 : D_Y = D_\varepsilon$ – дисперсия зависимой переменной, обусловленная соответственно регрессией или объясняющими переменными, равна дисперсии, обусловленной воздействием неучтенных случайных факторов.

Сформулируем альтернативную гипотезу $H_1 : D_Y > D_\varepsilon$ – вклад, вносимый в дисперсию совместным одновременным влиянием объясняющих переменных $x_1, x_2, \ldots, x_n$ выше суммарного эффекта от воздействия всех неучтенных случайных факторов.

Вычислим оценку остаточной дисперсии как сумму квадратов отклонений наблюдаемых значений *y* от модельных значений

$$S_\varepsilon^2 = \frac{1}{n-(k+1)}\sum_{i=1}^{n}\left(y_i - \left(\tilde{b}_0 + \tilde{b}_1 x_{i1} + \tilde{b}_2 x_{i2} + \cdots + \tilde{b}_k\right)\right)^2$$

и вычислим оценку общей дисперсии как сумму квадратов отклонений значений зависимой переменной *y* от средней

$$S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2.$$

Тогда отношение несмещённой оценки дисперсий зависимой переменной, обусловленной соответственно регрессией или совместным одновременным влиянием объясняющих переменных к несмещённой оценке дисперсий, обусловленной воздействия всех неучтенных случайных факторов и ошибок имеет распределение Фишера-Снедекора

$$\frac{S_Y^2}{S_\varepsilon^2} \sim F_{(n-1;\, n-(k+1))},$$

определенное на уровне значимости α при $(n-1)$ и $(n-(k+1))$ степенях свободы. Отношение, то есть уравнение регрессии значимо, если $F_{набл} > F_{табл}$. Проверяя гипотезу, *мы только утверждаем непротиворечивость линейного вида функции регрессии имеющимся результатам наблюдений, но вовсе не утверждаем, что этот вид зависимости является единственно возможным*

Порядок выполнения работы

1) Получить у преподавателя набор экспериментальных данных.
2) Изучить теоретическую часть [1, 2, 5]. Ответить на вопросы.
3) Найти коэффициенты регрессионной модели матричным способом. Если по условию задачи дана полиномиальная (квадратичная) регрессия $y = c + bx + ax^2$, то

$$\begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} \equiv \begin{pmatrix} c \\ b \\ a \end{pmatrix}$$

4) Пояснить, что показывает уравнение регрессии. Если в варианте задания есть текст задачи («вес семян», «урожайность пшеницы» и т.д.), то по тексту задачи пояснить смысл b_0, b_1, b_2. Если текста задачи нет, то пояснить смысл коэффициентов регрессии в общем виде.
5) Проверить значимость коэффициентов регрессии. Уровень значимости принять равным α=0,05.
6) Построить доверительные интервалы для коэффициентов регрессионной модели.
7) Проверить адекватность модели. Сделать вывод об использовании модели для принятия решений и прогноза.
8) Построить на одном рисунке диаграмму рассеяния и график модельной кривой. Для вариантов заданий со множественной линейной регрессией по оси абсцисс отложить переменную, коэффициент регрессии при которой значим (x_1 или x_2).
9) Написать отчет и защитить его перед преподавателем.

Варианты заданий

Приводятся варианты заданий для построения множественной линейной регрессии (Приложение Ж).

Контрольные вопросы

1) Какая зависимость величины *Y* от *X* называется функциональной?
2) Какая зависимость величины *Y* от *X* называется стохастической или вероятностной?
3) Что называется регрессией?
4) Какие задачи решает регрессионный анализ?
5) Как называется график функции регрессии?
6) Как называются независимые переменные в регрессионном анализе? Зависимые?
7) Что описывает вектор невязок?
8) Как в матричной форме записать уравнение модели?
9) Для чего вводится фиктивная переменная x_0?
10) Изложите идею метода наименьших квадратов.
11) Запишите систему нормальных уравнений в матричном виде и используя понятие обратной матрицы выведите формулу для оценок коэффициентов регрессионной модели.
12) Какие требования предъявляются к экспериментальным данным?
13) Являются ли параметры регрессии b_0, b_1, b_2 случайными величинами? Если да, то какому распределению они подчиняются?
14) Что такое диаграмма рассеяния? Какие выводы она позволяет сделать?
15) Как проверяется адекватность модели?

2.7 Лабораторная работа «Временные ряды»

Цель работы

Знакомство с методами исследования модели временного ряда.

Форма проведения

Выполнение индивидуального задания средствами Excel.

Форма отчетности

Защита отчета. Отчет оформляется в виде файла Word.

Теоретические основы

Для подготовки к лабораторной работе рекомендуется [6]. Построение регрессионной модели проводится с помощью пакета анализа данных «Статистика» Excel. Полученные остатки исследуются с помощью трех методов.

Критерий **Голдфелда-Квандта** позволяет ответить на вопрос о наличии гетероскедастичности остатков. Гетероскедастичность означает, что остатки модели регрессии не являются постоянными, вследствие чего доверительные интервалы будут ненадежными. Проверка осуществляется в предположении о нормальном распределении остатков регрессии и пропорциональности их дисперсий значениям фактора. Последовательность остатков упорядочивается по возрастанию фактора и выделяются две группы остатков: третья часть в начале последовательности и третья часть – в конце. По этим группам рассчитываются суммы квадратов остатков: $S_1 = \sum_{i=1}^{k} \varepsilon_i^2$ и $S_3 = \sum_{i=n-k+1}^{n} \varepsilon_i^2$, где n – общее количество наблюдений, $k \approx n/3$.

Основная гипотеза $H_0 : S_1 = S_3$ означает отсутствие гетероскедастичности, альтернатива $H_1 : S_1 > S_3$ – наличие гетероскедастичности (рост дисперсий остатков с ростом значений фактора). Для проверки гипотезы рассматривается статистика, имеющая распределение Фишера $F_{(k-1,k-1)}$. Если $F_{факт} = \frac{S_3}{S_1} > F_{табл}$, где $F_{табл} = F_{(\alpha;k-1,k-1)}$, α – выбранный уровень значимости, то гипотеза H_0 об отсутствии гетероскедастичности отклоняется.

Критерий **серий** применяется для исследования вопроса о случайности остатков временного ряда. В ранговом критерии «восходящих» и «нисходящих» серий формируется последовательность знаков «+» или «–», на основе которой и принимается решение.

Пусть для временного ряда x_t, $t = \overline{1,n}$, построена модель и найдены остатки $\varepsilon_t, t = \overline{1,n}$ Опишем последовательность шагов критерия.

Шаг 1. Формулируем основную и альтернативную гипотезы:

H_0 : остатки ε_t случайны

H_1 : остатки ε_t не являются случайными

Шаг 2. Задаем уровень значимости α.

Шаг 3. Формируем последовательность знаков.

Для этого рассматриваем разности $\varepsilon_t - \varepsilon_{t-1}$ $(t = \overline{1,n})$ и ставим «+», если разность положительна, ставим «–», если она отрицательна. Если при некотором t оказалось, что $\varepsilon_t = \varepsilon_{t-1}$, то не ставим никакой знак.

Шаг 4. В качестве статистики рассматривается пара $(S, K_{\max})$, где S – количество серий в последовательности знаков (серия – набор идущих подряд одинаковых знаков), а $K_{\max}$ количество знаков в самой длинной серии. Наличие длинных серий в последовательности знаков – довод против гипотезы H_0.

Шаг 5. Принятие решения.

Если одновременно выполняются условия $\begin{cases} S > S_{\text{гр}}(n) \\ K_{\max} > K_{\text{гр}}(n) \end{cases}$, то гипотеза H_0 может быть принята с вероятностью ошибки первого рода α. Если хотя бы одно из неравенств нарушено, гипотеза H_0 отвергается, т.е. остатки $\varepsilon_t, t = \overline{1,n}$ нельзя считать статистически независимыми.

Границы критической области:

$$S_{\text{гр}}(n) = \frac{2n-1}{3} - 1{,}96 \cdot \sqrt{\frac{16n-29}{90}} \text{ при } \alpha = 0{,}05.$$

$$K_{\text{гр}}(n) = \begin{cases} 5 \text{ при } n \le 26, \\ 6 \text{ при } 26 < n \le 153, \\ 7 \text{ при } 153 < n \le 1170. \end{cases}$$

Критерий **Дарбина-Уотсона** отвечает на вопрос о наличии автокорреляции в остатках. Наблюдаемое значение статистики Дарбина-Уотсона рассчитывается по формуле

$$DW = \frac{\sum(\varepsilon_i - \varepsilon_{i-1})^2}{\sum \varepsilon_i^2}$$

и сравнивается с табличными границами d_L и d_U для заданного числа наблюдений *n*. Если наблюдаемое значение статистики $DW \in [0; d_L)$, то автокорреляция положительна; если $DW \in (4-d_L; 4]$, то автокорреляция отрицательна, в случае $DW \in (d_U; 4-d_U)$ – автокорреляция отсутствует. В остальных случаях вопрос об автокорреляции остатков остается открытым.

Порядок выполнения работы

В соответствии с номером варианта скопировать в Excel данные временного ряда и выполнить следующие задания.

1) Обратиться к пакету анализа данных «Статистика» Excel, подобрать с его помощью наилучшую модель, прокомментировать лист отчета, вывести остатки временного ряда.
2) Проанализировать остатки временного ряда графически, исследовать их на наличие гетероскедастичности с помощью теста Голдфелда-Квандта.
3) Проанализировать остатки временного ряда по критерию серий.
4) Проанализировать остатки временного ряда с помощью статистики Дарбина-Уотсона. Сделать вывод.
5) Написать отчет.

Варианты заданий

Приводятся варианты заданий для лабораторной работы (Приложение 3).

Контрольные вопросы

1) Как называется основной метод построения регрессионной модели? Опишите его суть.
2) Какие требования предъявляются к экспериментальным данным в методе наименьших квадратов?
3) Каков смысл требования $M(X^T \varepsilon) = 0$?
4) Каков смысл требования $D(\varepsilon_i) = \sigma_\varepsilon^2$?
5) Что означает условие $M(\varepsilon \varepsilon^T) = 0$?
6) В чем суть гетероскедастичности?
7) Каковы последствия гетероскедастичности остатков?

8) Какие методы применяются для выявления гетероскедастичности?
9) Опишите схему теста Голдфелда-Кванта.
10) Для чего применяется взвешенный метод наименьших квадратов и в чем его суть?
11) Каковы последствия автокорреляции?
12) Опишите алгоритм выявления автокорреляции остатков с использованием критерия Дарбина-Уотсона.
13) Для чего применяется критерий серий и в чем его суть?

2.8 Лабораторная работа «Цепи Маркова»

Цель работы

Закрепление теоретических знаний и получение навыка исследования с помощью марковских цепей.

Форма проведения

Выполнение индивидуального задания (решение ситуационной задачи) средствами MathCad.

Форма отчетности

Защита отчета. Отчет оформляется в виде файла MathCad.

Теоретические основы

Теоретический материал приведен в [7,8]. В электронном курсе опубликованы теоретические сведения о цепях Маркова, разобраны примеры решения задач.

Порядок выполнения работы

Задача 1.

1. Описать ситуацию первой задачи с помощью цепи Маркова. Объяснить, почему можно использовать такую модель.
2. Нарисовать орграф. Охарактеризовать состояния цепи Маркова.
3. Последовательно возводя матрицу перехода в 5, 10, 15, ... степень, исследовать поведение цепи при неограниченном увеличении времени наблюдения.
4. Если предельные вероятности существуют, найти их.

Задача 2.

1. По данной матрице перехода **поглощающей** цепи Маркова нарисовать орграф. Занумеровать состояния.
2. Записать матрицу перехода в каноническом виде.
3. Найти фундаментальную матрицу.
4. Указать, каким состояниям соответствуют строки и столбцы фундаментальной матрицы.
5. Сделать вывод о среднем времени нахождения цепи Маркова в каждом из непоглощающих состояний.
6. Матричным способом найти финальные вероятности поглощения в каждом из поглощающих состояний.

Пример решения системы линейных уравнений в пакете MathCad. Дана матрица перехода *P*

$$P=\begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix}.$$

Данная цепь Маркова регулярна, так как $p_{22}>0$. Следовательно, по теореме Маркова существует стационарный режим. Найдем вектор предельных вероятностей $U^T=\begin{bmatrix} u_1 & u_2 & u_3 \end{bmatrix}^T$ из уравнения $U^T=U^T P$.

$$\begin{bmatrix} u_1 & u_2 & u_3 \end{bmatrix}=\begin{bmatrix} u_1 & u_2 & u_3 \end{bmatrix}\cdot\begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix}.$$

Перейдем от матричной формы записи к записи системы линейных уравнений:

$$\begin{cases} u_1=\frac{1}{4}u_2+\frac{1}{4}u_3, \\ u_2=\frac{1}{2}u_1+\frac{1}{2}u_2+\frac{1}{4}u_3, \\ u_3=\frac{1}{2}u_1+\frac{1}{4}u_2+\frac{1}{2}u_3, \\ u_1+u_2+u_3=1. \end{cases}$$

Для решения данной системы линейных уравнений в пакете MathCad зададим начальные приближения:

u1:=0 u2:=0 u3:=0

Решим уравнение $U^T=U^T P$. Переходя от матричной формы записи к записи системы линейных уравнений, имеем:

Given

$u_1=0.25\cdot u_2+0.25\cdot u_3$

$u_2=0.5\cdot u_1+0.5\cdot u_2+0.25\cdot u_3$

$u_3=0.5\cdot u_1+0.25\cdot u_2+0.5\cdot u_3$

$u_1+u_2+u_3=1$

Используем вместо знака равенства «=» значок «=» Булевой алгебры (см. иконку Boolean).

Тогда искомый вектор предельных вероятностей

$$R:=Find\left(u_1,u_2,u_3\right)$$

Чтобы представить результат в обыкновенных дробях, щелкните правой кнопкой мыши по черному квадратику и выберите Format→Result→Fraction.

$$R = \begin{pmatrix} \frac{1}{5} \\ \frac{2}{5} \\ \frac{2}{5} \end{pmatrix}$$

Варианты заданий

Приводятся варианты заданий для лабораторной работы (Приложение И).

Контрольные вопросы

1) Что такое цепь Маркова?
2) В чем состоят свойства стохастичности матрицы перехода цепи Маркова?
3) Что такое поглощающее состояние цепи Маркова?
4) Как найти распределение вероятностей через *m* шагов от начала наблюдения?
5) Сформулируйте теорему о вероятности перехода в эргодическое состояние.
6) Что такое поглощающая цепь Маркова?
7) Как получить канонический вид матрицы перехода?
8) Что показывают элементы подматриц Q и R
9) Сформулируйте теорему о свойствах подматрицы *Q*.
10) Сформулируйте теорему об элементах фундаментальной матрицы.
11) Сформулируйте теорему о вероятности поглощения в данном поглощающем состоянии (что показывают элементы матрицы B)?
12) Что такое эргодическая цепь Маркова?
13) Как найти предельные вероятности?

3 МЕТОДИЧЕСКИЕ УКАЗАНИЯ К САМОСТОЯТЕЛЬНОЙ РАБОТЕ

3.1 Теоретическая подготовка

Самостоятельная работа с теоретическим материалом направлена на повторение понятий и методов теории вероятностей и освоение понятий и методов математической статистики. Теоретическая подготовка включает в себя не только проработку лекционного материала, но и самостоятельное изучение тем (вопросов) теоретической части дисциплины.

Проработка конспекта лекций играет важную роль при выстраивании структуры дисциплины и выделения важных аспектов изучаемого материала. Конспект оказывает помощь студенту при подготовке к защите лабораторных работ и последующим лекционным занятиям; в нем должны быть выделены основные положения, определения и формулы. Контроль изучения теоретического материала проводится в виде теоретического опроса или теста при защите лабораторных работ.

Примерные темы опросов

1) Различные определения вероятности.
2) Дискретная случайная величина.
3) Непрерывная случайная величина.
4) Законы распределения случайных величин.
5) Числовые характеристики и их свойства.
6) Системы случайных величин.
7) Независимость и некоррелированность случайных величин.
8) Свойства коэффициента корреляции.
9) Стохастическая и функциональная зависимости.
10) Функция регрессии и ее свойства.
11) Предельные теоремы теории вероятностей.
12) Оценки параметров. Свойства оценок.
13) Статистические гипотезы.
14) Генерация псевдослучайных чисел.
15) Вычисление интегралов методом Монте-Карло.
16) Однофакторный дисперсионный анализ.
17) Корреляционный анализ.
18) Регрессионный анализ.
19) Случайные процессы.
20) Стационарность и эргодичность.
21) Эргодические цепи Маркова.
22) Поглощающие цепи Маркова.
23) Временные ряды.

Некоторые темы изучаемой дисциплины рассматриваются на лекциях в обзорном порядке, затем выносятся для более детальной проработки на самостоятельное изучение.

Примерные темы для самостоятельной работы

1) Сравнение критериев согласия по мощности.
2) Способы повышения точности метода Монте-Карло.
3) Многомерный корреляционный анализ.
4) Случайные процессы: гармонический анализ.

Изучая темы, вынесенные для самостоятельной работы, необходимо на бумаге составить план прочитанного материала и решить задачи и упражнения, предложенные в учебнике. При работе с разными источниками важно привести в систему обозначения и термины, которые могут отличаться в разных учебниках. Если при изучении материала проводится сравнительный анализ условий применения или свойств характеристик, то

результаты такого сравнения желательно представить в виде таблицы, алгоритма или интеллект-карты. Контроль самостоятельного изучения материала осуществляется в форме проверки конспекта, собеседования с преподавателем или доклада на занятии.

Подготовка доклада начинается с выбора темы и обсуждения содержания доклада с преподавателем. Затем в течение недели студент осуществляет самостоятельный поиск в интернете или работает с литературой, рекомендованной преподавателем. Через неделю студент обсуждает с преподавателем план доклада и продолжает подготовку текста доклада и презентации. Оценка и рецензирование доклада проводится студентами и преподавателем совместно в устной или письменной форме.

Примерные темы докладов

1) Графическое представление информации в статистике.
2) Исследование свойств оценок распределения.
3) Ранговые методы статистики.
4) Гармонический анализ случайных процессов.

В случае самостоятельного изучения тем «Регрессионный анализ для временных рядов и случайные процессы общего вида», «Временные ряды», «Случайные процессы», обычно вызывающих затруднение у студентов, следует рассмотреть теоретический материал [3, 5, 6, 7]. Для лучшего усвоения указанных тем студентам предлагается ответить на ряд вопросов.

Вопросы по теме «Случайные процессы»

1) Как описать случайный процесс общего вида?
2) Что такое сечение случайного процесса? Реализация случайного процесса? Какие законы распределения необходимо задать для описания случайного процесса?
3) Какие основные характеристики рассматриваются для СП общего вида?
4) Какими свойствами обладают математическое ожидание и корреляционная функция случайного процесса общего вида?
5) Что такое стационарные (в узком и в широком смысле) случайные процессы?
6) Каковы свойства корреляционной функции стационарного случайного процесса?
7) Что такое нормированная корреляционная функция стационарного случайного процесса и какими свойствами она обладает?
8) Что такое интервал корреляции функции стационарного случайного процесса? Как его найти?
9) Что такое эргодические случайные процессы? Сформулируйте достаточные условия эргодичности.
10) Спектральное представление стационарного случайного процесса. Что показывает спектральная плотность?
11) Взаимосвязь спектральной плотности и корреляционной функции (уравнения Винера-Хинчина).
12) Комплексная форма записи спектрального разложения, уравнения Винера-Хинчина в комплексной форме.
13) Что такое эффективная ширина спектра стационарного случайного процесса?
14) Для чего используется частота Найквиста? Сформулируйте теорему Котельникова.

Вопросы по теме «Временные ряды»

1) Какие компоненты рассматриваются при построении модели временного ряда?
2) Что такое тренд временного ряда?
3) В чем разница между циклической и сезонной компонентами временного ряда?
4) Какими особенностями должны обладать критерии, применяющиеся при проверке гипотез о компонентах модели временного ряда?
5) Какую гипотезу проверяют с помощью критерия Фостера-Стюарта? Опишите суть критерия.
6) Какую гипотезу проверяют с помощью критерия поворотных точек? Опишите суть критерия.

7) Какую задачу решают с помощью критерия серий? Опишите суть критерия.
8) Какую задачу решают с помощью статистики Дарбина-Уотсона? В чем недостаток этого метода?
9) Какие методы используются при построении модели временного ряда?
10) Поясните термин «гетероскедастичность». Какие методы используются для выявления гетероскедастичности?
11) Для чего используется метод инструментальных переменных?
12) Для чего используется взвешенный метод наименьших квадратов?
13) Для чего используется авторегрессионное преобразование?
14) Какие методы применяются в задаче прогнозирования временных рядов?

3.2 Подготовка к лабораторным работам

Лабораторные работы позволяют получить навыки

- представления и обработки статистических данных;
- освоения алгоритма проверки статистических гипотез;
- анализа зависимостей в группах статистических данных;
- построения и анализа вероятностных и статистических моделей;
- интерпретации результатов исследования и написания отчета об исследовании.

Для подготовки к лабораторной работе необходимо изучить теоретический материал по теме работы, проработать основные понятия, ответить на контрольные вопросы, составить предварительный отчет, выписав туда основные формулы для расчетов по теме лабораторной работы.

По результатам выполнения лабораторной работы оформляется отчет, который защищается перед преподавателем. Отчет о лабораторной работе должен содержать:

- титульный лист, оформленный по стандарту ОС ТУСУР;
- номер варианта;
- условия всех задач;
- решение каждой задачи с необходимыми пояснениями и формулами.

При защите отчета студент должен знать используемые термины, уметь формулировать определения и теоремы, давать пояснения к решению.

3.3 Подготовка к промежуточной аттестации

Ниже приведены задачи и примеры вопросов для подготовки к промежуточной аттестации по дисциплине «Теория вероятностей и математическая статистика».

Задачи по теме «Основы математической статистики»

После выполнения двух лабораторных работ по темам «Основы математической статистики» и «Проверка статистических гипотез» студент должен знать основные понятия математической статистики, уметь пользоваться оценками параметров, знать их свойства, уметь формулировать статистические гипотезы и выбирать подходящие статистики для проверки гипотез, освоить алгоритм проверки статистических гипотез.

Задача 1 проверяет умение студента строить критическую область и пользоваться соответствующими таблицами распределения. Задача 2 проверяет умение студента формулировать математическую постановку данной ситуации и применять алгоритм проверки параметрической гипотезы. При решении задач 1 и 2 рекомендуется использовать справочный материал (Приложение А). Задача 3 требует применения критерия согласия (Пирсона) или критерия однородности (Вилкоксона). Примеры задач приведены ниже.

Задача 1. Нарисовать критическую область для проверки гипотезы $H_0: m_1 = m_2$ при альтернативе $H_1: m_1 \neq m_2$. Найти ее границы для уровня значимости $\alpha=0,02$ и

числе степеней свободы ν = 5. Предполагается нормальное распределение генеральной совокупности.

Задача 2. Точность наладки станка-автомата, выпускающего металлические стержни, характеризуется дисперсией длины стержня. Если эта величина будет больше *400 мкм*^{2}, станок останавливается для наладки. Выборочная дисперсия длины *15* случайно отобранных стержней из продукции станка оказалась равной $S^2 = 680$ *мкм*^{2}. Нужно ли производить наладку станка, если уровень значимости $\alpha = 0{,}01$?

Задача 3. Метод получения случайных чисел был применен 250 раз, при этом получены следующие результаты (i – цифра, n_i – частота ее появления)

i	0	1	2	3	4	5	6	7	8	9
n_i	27	18	23	31	21	23	28	25	22	32

Можно ли считать, что примененный метод действительно дает случайные числа?

Задачи по теме «Метод Монте-Карло и дисперсионный анализ»

После выполнения двух лабораторных работ по темам «Метод Монте-Карло» и «Дисперсионный анализ» предлагается решение следующих трех задач.

Первая задача является частью лабораторной работы «Метод Монте-Карло». При решении задачи требуется оценить количество опытов для вычисления интеграла, сформулировать теоремы, на которых основан метод и пояснить оценивание разброса с помощью характеристики «размах».

Во второй задаче необходимо применить указанный метод генерации псевдослучайных чисел и найти период полученной числовой последовательности. Начальные значения подобраны так, что период будет коротким (5 – 6 различных значений), поэтому если последовательность получается более длинной, следует вернуться и проверить расчеты с самого начала.

Третья задача описывает ситуацию, которую необходимо сформулировать как постановку задачи дисперсионного анализа. Поскольку в задаче приведен небольшой набор опытов и не делается предположения о нормальном распределении данных, следует применить ранговый метод дисперсионного анализа. Критические точки необходимо вычислить с помощью таблицы распределения «хи-квадрат».

Задача 1. Представить интеграл

$$I = \int_{-\frac{\pi}{6}}^{\frac{\pi}{6}} |\sin 3x| dx$$

в виде, удобном для применения метода Монте-Карло. Указать распределение случайной величины, которую необходимо генерировать для вычисления. Оценить количество опытов при доверительной вероятности $\beta = 0{,}94$ и требуемой точности $\varepsilon = 0{,}03$.

Задача 2. Определить период и записать последовательность различных значений псевдослучайных чисел, полученных методом вычетов с начальными значениями $m_0=1$, M=11, K=14.

Задача 3. На четырех малых предприятиях по одной технологии производятся комплектующие детали для основного производства. В таблице приведены данные о производительности труда (в условных единицах). Зависит ли производительность труда от номера предприятия?

Предприятие	Производительность труда					
1	50	53	58	62	60	57
2	54	46	50	64	59	63
3	52	48	51	70	62	61
4	60	55	56	58	54	51

Задачи по теме «Корреляционный и регрессионный анализ»

Задачи формулируются на основе соответствующих лабораторных работ, выполнение которых предваряет освоение необходимых навыков решения.

Первая задача проверяет умение студентов измерять корреляцию между двумя рядами данных. Здесь количество опытов мало и предположения о нормальности распределения не делается, поэтому следует рассчитать коэффициент корреляции Спирмена и проверить его значимость. Если уровень значимости в задании не указан, во всех задачах контрольной работы следует выполнять проверку при $\alpha = 0{,}05$.

Вторая задача требует знания свойств корреляционного отношения. В задаче требуется сформулировать математическую постановку для проверки гипотезы о линейности связи, подобрать статистику с известным законом распределения, построить критическую область и принять статистическое решение.

В третьей задаче проверяется знание основного метода построения регрессионных моделей и умение применить матричную форму записи для решения задачи регрессионного анализа. Пункты г) и д) можно считать дополнительными заданиями.

Задача 1. При приеме на работу семи кандидатам на вакантные должности было предложено два теста. Результаты тестирования (в баллах) приведены в таблице. Оценить корреляцию между тестами на уровне значимости 0,05.

Тест	Кандидат						
	1	2	3	4	5	6	7
1	31	82	25	26	53	30	29
2	21	55	8	27	32	42	26

Задача 2. По 56 выборочным данным (в корреляционной таблице 8 значений фактора и 6 значений отклика) была получена оценка коэффициента корреляции $r_{yx} = -0{,}64$ и оценка корреляционного отношения $\eta^2_{yx} = 0{,}54$. Можно ли описать зависимость y от x линейной моделью?

Задача 3. Предполагается, что зависимость y от x может быть описана линейной функцией.

x	–3	–2	–1	1	2
y	–5	–3	–1,5	1	3

Задание:

1) постройте диаграмму рассеяния по выборочным данным и проиллюстрируйте метод наименьших квадратов графически;
2) запишите уравнение регрессии для данной таблицы наблюдений в матричной форме;
3) найдите оценки коэффициентов матричным способом;
4) проверьте адекватность модели;
5) постройте 95% доверительный интервал для линии регрессии.

Задачи по теме «Временные ряды и цепи Маркова»

После выполнения двух лабораторных работ по темам «Временные ряды» и «Цепи Маркова» предлагается самостоятельное решение следующих двух задач.

Первая задача по теме «Временные ряды» может содержать задание на проверку гипотезы о существовании тренда, о наличии периодической составляющей в модели ВР, гипотезы о гетероскедастичности остатков, гипотезы о случайности остатков или гипотезы о

наличии автокорреляции в остатках временного ряда. Студенту необходимо повторить алгоритмы проверки перечисленных гипотез.

Вторая задача выбирается случайным образом из заданий на лабораторную работу по теме «Цепи Маркова» [8] и может содержать как задачу на исследование эргодической цепи Маркова, так и задачу на поглощающую цепь Маркова.

Задача 1. В таблице представлена динамика выпуска продукции Королевства кривых зеркал (у.е.).

Год	Выпуск	Год	Выпуск	Год	Выпуск
2001	11172	2006	13471	2011	23298
2002	14150	2007	13617	2012	26570
2003	14004	2008	16356	2013	23080
2004	13088	2009	20037	2014	23981
2005	12518	2010	21748	2015	23446

1) Проверить гипотезы о структуре модели временного ряда по критерию Фостера-Стюарта;
2) по критерию поворотных точек.

Задача 2. Частица на прямой может иметь координаты x=1, 2, 3, 4. Каждую секунду частица может совершать единичные скачки влево или вправо с вероятностями соответственно 0,2 и 0,8. Из положения x=1 частица с вероятностью 0,7 переходит в точку x=2 и с вероятностью 0,3 остается на месте; а из положения x=4 она с вероятностью 0,6 остается на месте, а с вероятностью 0,4 переходит в положение x=3. Задание: а) нарисовать орграф блужданий частицы; б) составить матрицу перехода цепи Маркова; в) найти предельные вероятности.

Промежуточная аттестация подводит итог курсу теории вероятностей и математической статистики. Для получения оценки «удовлетворительно» студенту достаточно ответить на вопросы тестового характера (темы вопросов приведены выше в п.3.1).

Примеры тестовых вопросов

Вопрос 1. Подброшены две монеты. Для события «выпал хотя бы один герб» противоположным является событие

А) выпала ровно одна цифра;
Б) выпала хотя бы одна цифра;
В) выпало две цифры;
Г) выпало два герба.
Д) выпало две цифры или выпало два герба

Вопрос 2. Если вероятность события А не меняется при наступлении события В, то эти события

А) невозможные;
Б) несовместные;
В) независимые;
Г) неполные.

Вопрос 3. В ряде распределения случайной величины X

X	2	3	4
P	0,1	***p***	0,3

пропущено значение ***p***, равное

А) 0; Б) 0,2; В) 0,4; Г) 0,6; Д) 0,8; Е) 1.

Вопрос 4. Функция распределения *F(x)* случайной величины *X* НЕ обладает свойством

А) стремится к 1 при $x \to \infty$;

Б) неотрицательна;
В) не убывает;
Г) площадь под кривой равна 1.
Вопрос 5. Случайная величина X – количество подбрасываний игральной кости до выпадения первой «шестерки» – распределена по закону
А) $Bin\left(6,\frac{1}{6}\right)$; Б) $Bin\left(6,\frac{1}{2}\right)$; В) $G(6)$; Г) $G(1/6)$; Д) $N(6,1/6)$.
Вопрос 6. Дана матрица распределения системы дискретных СВ (X,Y): $\left[p_{ij}\right]$, $i=\overline{1,n}; j=\overline{1,m}$. Сумма элементов j-ого столбца равна
А) единице;
Б) вероятности $P(X=x_i)$;
В) вероятности $P(Y=y_j)$;
Г) условной вероятности $P(X\left|Y=y_j\right.)$;
Д) условной вероятности $P(Y\left|X=x_i\right.)$.
Вопрос 7. Если случайные величины некоррелированы, то их ковариация
А) равна +1;
Б) равна –1;
В) равна 0;
Г) не может быть вычислена.
Вопрос 8. Закон больших чисел утверждает, что
А) при больших значениях случайной величины ее математическое ожидание постоянно;
Б) при большом значении среднего арифметического дисперсия случайной величины мала;
В) при большом количестве опытов значение среднего арифметического случайной величины примерно равно математическому ожиданию;
Г) при большом количестве опытов среднее арифметическое случайной величины можно описать нормальным законом распределения.
Вопрос 9. Гистограмма – это
А) выборка, упорядоченная по возрастанию;
Б) точечная оценка параметра генеральной совокупности;
В) интервальная оценка параметра генеральной совокупности;
Г) оценка функции распределения генеральной совокупности;
Д) оценка плотности распределения генеральной совокупности.
Вопрос 10. Оценка параметра называется несмещенной, если
А) ее математическое ожидание равно нулю;
Б) ее дисперсия равна нулю;
В) ее математическое ожидание равно значению параметра;
Г) ее дисперсия равна дисперсии параметра.
Вопрос 11. Однофакторный дисперсионный анализ отвечает на вопрос:
А) оказывает ли влияние признак X на фактор F?
Б) оказывает ли влияние фактор F на признак X?
В) какой уровень фактора F оказывает влияние на признак X?
Г) на какой уровень фактора F оказывает влияние признак X?
Вопрос 12. В методе Монте-Карло случайная величина – это случайная величина, распределенная
А) по стандартному нормальному закону;
Б) по закону Стьюдента;
В) по равномерному закону;

Г) по закону больших чисел.
Вопрос 13. Для оценки качества регрессионной модели НЕ применяется
А) коэффициент детерминации;
Б) интервал корреляции;
В) корреляционное отношение;
Г) остаточная дисперсия.
Вопрос 14. Если коэффициент корреляции Пирсона равен (–1), то корреляционное отношение
А) равно +1;
Б) равно –1;
В) меньше 1;
Г) больше 1;
Д) равно 0.
Вопрос 15. Если корреляционное отношение равно единице, то связь
А) условная;
Б) незначимая;
В) линейная;
Г) нелинейная;
Д) функциональная.
Вопрос 16. Корреляционная функция СТАЦИОНАРНОГО случайного процесса – это
А) случайная функция времени;
Б) неслучайная функция времени;
В) неслучайная функция, зависящая от промежутка между сечениями;
Г) неслучайная периодическая функция;
Д) постоянная величина.
Вопрос 17. Состояние цепи Маркова, вероятность выхода из которого равна нулю, называется
А) поглощающим;
Б) транзитивным;
В) стационарным;
Г) эргодическим.
Вопрос 18. Если на орграфе эргодической цепи Маркова существует хотя бы одна петля, то
А) имеется хотя бы одно поглощающее состояние;
Б) орграф такой цепи Маркова слабо связан;
В) условие регулярности цепи Маркова не выполняется;
Г) существуют предельные вероятности.
Вопрос 19. Для устранения гетероскедастичности используется
А) взвешенный метод наименьших квадратов;
Б) авторегрессионное преобразование;
В) метод инструментальных переменных;
Г) рекуррентный метод наименьших квадратов.
Вопрос 20. Для проверки существования периодической составляющей применяют критерий
А) Дарбина-Уотсона;
Б) Фостера-Стюарта;
В) критерий серий;
Г) поворотных точек.

Для получения оценки «хорошо» или «отлично» студент должен знать основные понятия теории вероятностей и математической статистики, уметь излагать их в корректной математической форме, пояснять на примерах. Студенту необходимо продемонстрировать умения и навыки, приобретенные при решении задач вероятностного и статистического

характера. В билете две задачи, аналогичные задачам для подготовки к промежуточной аттестации. Тематика задач приведена ниже.

Задача А

1) Применение неравенства Чебышева и центральной предельной теоремы для суммы и среднего арифметического независимых случайных величин.
2) Применение неравенства Чебышева и центральной предельной теоремы для относительной частоты события.
3) Применение теорем Муавра-Лапласа для расчета вероятностей биномиального закона распределения.
4) Построение интервальной оценки. Влияние метода отбора на точность оценки.
5) Определение количества опытов, необходимых для получения оценки с заданной точностью и заданной доверительной вероятностью
6) Проверка параметрических гипотез (о числовых характеристиках нормальной генеральной совокупности, о сравнении параметров нормальной генеральной совокупности, о генеральной доле).
7) Проверка гипотезы о виде распределения по критерию Пирсона.
8) Проверка гипотезы об однородности данных (критерий Вилкоксона).

Задача Б

1) Приведение интеграла к виду, удобному для применения метода Монте-Карло. Предварительная оценка количества опытов.
2) Генерация случайной величины методом вычетов; определение периода, проверка качества моделирования по критерию Пирсона.
3) Решение задачи однофакторного дисперсионного анализа (классическая схема).
4) Решение задачи непараметрического анализа (критерий Крускала-Уоллиса).
5) Выборочный коэффициент корреляции Пирсона. Проверка гипотезы о независимости.
6) Корреляционное отношение. Проверка гипотезы о линейности связи.
7) Коэффициент корреляции Спирмена. Проверка значимости.
8) Парная линейная регрессия. Нахождение оценок коэффициентов регрессии методом наименьших квадратов.
9) Проверка адекватности модели с помощью остаточной дисперсии.
10) Множественная линейная регрессия. Нахождение оценок параметров в матричной форме.
11) Компоненты модели временного ряда. Проверка гипотезы о существовании тренда.
12) Проверка гипотезы о наличии периодической составляющей.
13) Проверка гипотезы о случайности остатков временного ряда.
14) Проверка гипотезы об отсутствии автокорреляции остатков временного ряда.
15) Выявление гетероскедастичности данных.
16) Матрица перехода и орграф однородной цепи Маркова. Многошаговый переход в цепи Маркова. Распределение вероятностей через m шагов от начала наблюдения.
17) Поглощающие цепи Маркова. Канонический вид матрицы перехода. Нахождение среднего времени до момента поглощения.
18) Поглощающие цепи Маркова. Канонический вид матрицы перехода. Нахождение вероятности поглощения в заданном поглощающем состоянии.
19) Эргодические цепи Маркова. Нахождение предельных вероятностей.

Пример практической части билета промежуточной аттестации

Задача 1. Выборочное обследование распределения населения города по среднедушевому денежному доходу показало, что 45% обследованных в выборке имеют среднедушевой денежный доход не более 18 тыс. руб. В каких пределах находится доля населения, имеющая такой среднедушевой доход, во всей генеральной совокупности, если объем генеральной совокупности составляет 100000 человек, выборка не превышает 10%

объема генеральной совокупности и осуществляется по методу случайного бесповторного отбора, а доверительная вероятность принимается равной 0,92?

Задача 2 Предполагается, что зависимость между переменными *y* и *x*

x	2,5	4,5	5,0	1,5	3,5	6,0	6,5	4,0	3,5	2,0
y	0,5	1,2	1,7	0,3	0,8	2,7	3,3	1,0	0,7	0,4

описывается функцией $y = ax^2 + bx + c$, ошибки наблюдений независимы и имеют распределение $N(0, \sigma)$. С помощью замены переменных перейти к модели множественной линейной регрессии, записать уравнение в матричной форме и найти оценки параметров матричным способом.

Теоретические вопросы

1) Дисперсия и ее свойства.
2) Коэффициент корреляции и его свойства.
3) Функция регрессии.
4) Сходимость по распределению. Интегральная теорема Муавра-Лапласа как следствие центральной предельной теоремы.
5) Сходимость по вероятности. Закон больших чисел. Сформулировать теорему Чебышева.
6) Сходимость по вероятности. Закон больших чисел. Сформулировать теорему Бернулли.
7) Метод максимального правдоподобия. Пример получения оценок и исследование свойств.
8) Геометрический метод Монте-Карло и оценка его точности.
9) Однофакторный дисперсионный анализ. Основное дисперсионное тождество. Обоснование перехода к гипотезе о сравнении дисперсий.
10) Метод наименьших квадратов. Вывести формулы для простой линейной регрессии.
11) Цепи Маркова. Сформулировать теорему о вероятности перехода в эргодическое множество.
12) Канонический вид матрицы перехода поглощающей цепи Маркова. Сформулировать теорему о свойствах элементов фундаментальной матрицы.

4 РЕКОМЕНДУЕМЫЕ ИСТОЧНИКИ

1. Хрущева, И. В. Основы математической статистики и теории случайных процессов : учебное пособие / И. В. Хрущева, В. И. Щербаков, Д. С. Леванова. — Санкт-Петербург : Лань, 2009. — 336 с. — ISBN 978-5-8114-0914-3. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/426 (дата обращения: 24.01.2022). — Режим доступа: для авториз. пользователей.
2. Гмурман В.Е. Теория вероятностей и математическая статистика.– М.: Высш.шк., 2003.– 480 с.
3. Хрущева, И. В. Теория вероятностей : учебное пособие / И. В. Хрущева. — Санкт-Петербург : Лань, 2021. — 304 с. — ISBN 978-5-8114-0915-0. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/167789 (дата обращения: 24.01.2022). — Режим доступа: для авториз. пользователей.
4. Буре, В. М. Теория вероятностей и математическая статистика : учебник / В. М. Буре, Е. М. Парилина. — Санкт-Петербург : Лань, 2021. — 416 с. — ISBN 978-5-8114-1508-3. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/168536 (дата обращения: 24.01.2022). — Режим доступа: для авториз. пользователей.
5. Кремер Н.Ш. Теория вероятностей и математическая статистика: Учебник для вузов.– 2-е изд., перераб. и доп. – М.: ЮНИТИ-ДАНА, 2004.– 573 с.
6. Потахова, И. В. Эконометрика: Методические указания к лабораторным работам и самостоятельной работе [Электронный ресурс] / И. В. Потахова. — Томск: ТУСУР, 2018. — 60 с. — Режим доступа: https://edu.tusur.ru/publications/8138 (дата обращения: 24.01.2022). — Режим доступа: для авториз. пользователей.
7. Вентцель Е.С., Овчаров Л.А Теория случайных процессов и ее инженерные приложения. – М.: Академия, 2003 г.
8. Смыслова З.А. Цепи Маркова [Электронный ресурс]. — Режим доступа: https://edu.tusur.ru/lecturer/distance_courses/466 (дата обращения: 24.01.2022). — Режим доступа: для авториз. пользователей.

ПРИЛОЖЕНИЕ А

Параметрические гипотезы

Гипотеза о параметре генеральной совокупности

Гипотеза	$H_0 : a = a_0$	$H_0 : a = a_0$	$H_0 : \sigma = \sigma_0$
Предпо-ложения	НГС, σ известно	НГС, σ неизвестно	НГС *a* неизвестно
Оценки	$\hat{a} = \bar{x}$	$\hat{a} = \bar{x}\ ;\ \hat{\sigma} = s$	$\hat{a} = \bar{x}\ ;\ \hat{\sigma} = s$
Статис-тика	$\frac{\bar{X} - a_o}{\sigma} \cdot \sqrt{n}$	$\frac{\bar{X} - a_o}{s} \cdot \sqrt{n}$	$\frac{(n-1)S^2}{\sigma_o^2}$
Распре-деление	$N(0,\ 1)$	$T_{(n-1)}$	$\chi^2_{(n-1)}$

Сравнение параметров генеральной совокупности

Гипотеза	$H_0 : a_1 = a_2$	$H_0 : a_1 = a_2$	$H_0 : \sigma_1 = \sigma_2$
Предположения	НГС, σ_1, σ_2 известны	НГС, $\sigma_1 = \sigma_2$, σ_1, σ_2 неизвестны	НГС *a* неизвестно
Оценки	$\hat{a}_1 = \bar{x};$ $\hat{a}_2 = \bar{y}$	$\hat{a}_1 = \bar{x}; \hat{a}_2 = \bar{y}$ $\hat{\sigma}_1 = s_1; \hat{\sigma}_2 = s_2$ $s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$	$\hat{a}_1 = \bar{x}; \hat{a}_2 = \bar{y}$ $\hat{\sigma}_1 = s_1; \hat{\sigma}_2 = s_2$
Статистика	$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	$\frac{\bar{X} - \bar{Y}}{S} \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}$	$\frac{S_1^2}{S_2^2}$
Распределение	$N(0,\ 1)$	$T_{(n_1+n_2-2)}$	$F_{(n_1-1, n_2-1)}$

Гипотезы о генеральной доле и о сравнении генеральных долей

Гипотеза	$H_0 : p = p_0$	$H_0 : p_1 = p_2$
Предположения	Схема испытаний Бернулли	Схема испытаний Бернулли
Оценки	$p^* = \frac{m}{n}$	$p_1^* = \frac{m_1}{n_1};\ p_2^* = \frac{m_2}{n_2}\quad \tilde{p} = \frac{m_1 + m_2}{n_1 + n_2}$
Статистика	$\frac{p^* - p_o}{\sqrt{p_o(1-p_o)}} \cdot \sqrt{n}$	$\frac{p_1^* - p_2^*}{\sqrt{\tilde{p}(1-\tilde{p})}} \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}$
Распределение	$N(0,\ 1)$	$N(0,\ 1)$

ПРИЛОЖЕНИЕ Б

Варианты задания «Проверка статистических гипотез»

Задача 1.

ВАРИАНТ 1

Компания, производящая средства для потери веса, утверждает, что прием таблеток в сочетании со специальной диетой позволяет сбросить в среднем 400 г веса. Случайным образом отобраны 25 человек, использующих эту терапию, и обнаружено, что в среднем еженедельная потеря в весе составила 430 г со с.к.о. 110 г. Проверьте гипотезу о том, что средняя потеря в весе составляет 400 г. Уровень значимости $\alpha = 0{,}05$.

ВАРИАНТ 2

Инженер по контролю качества проверяет среднее время горения нового вида электроламп. Для проверки случайным образом было отобрано 20 ламп, среднее время горения которых составило 1075 часов. Предположим, что среднее квадратичное отклонение времени горения для генеральной совокупности известно и составляет 100 часов. На уровне значимости $\alpha = 0{,}05$ проверьте гипотезу о том, что среднее время горения ламп более 1000 часов.

Предположим, что инженер по контролю качества не имеет информации о генеральной дисперсии и использует выборочное среднеквадратичное отклонение. Изменится ли ответ?

ВАРИАНТ 3

Компания утверждает, что новый вид зубной пасты для детей лучше предохраняет зубы от кариеса, чем зубные пасты, производимые другими фирмами. Для проверки была отобрана случайным образом группа из 400 детей, которые пользовались новым видом зубной пасты. Другая группа из 300 детей, также случайно выбранных, в это же время пользовалась другими видами зубной пасты. Было выявлено, что у 30 детей, использующих новую пасту, и 25 детей из контрольной группы появились новые признаки кариеса.

Имеются ли у компании достаточные основания для утверждения о том, что новый сорт зубной пасты эффективнее предотвращает кариес, чем другие виды зубной пасты?

ВАРИАНТ 4

Компания по производству безалкогольных напитков предполагает выпустить на рынок новую модификацию популярного напитка, в котором сахар заменен препаратом на основе стевии. Компания хотела бы быть уверенной в том, что не менее 70 % ее потребителей предпочтут новую модификацию напитка. Новый напиток был предложен на пробу 2000 человек, и 1422 из них сказали, что он вкуснее старого.

Может ли компания отклонить предложение о том, что только 70 % всех ее потребителей предпочтут новую модификацию напитка старой? Уровень значимости 0,05.

Задача 2. Критерий согласия Пирсона.

По результатам наблюдений определены частоты n_j попадания случайной величины X в заданные интервалы $[a_j; a_{j+1}), j = 1,2,...,k$. Рассчитать по данному статистическому ряду оценки параметров $a = \bar{x}$ и $\sigma = s$, пользуясь формулами $\bar{x} = \frac{1}{n}\sum_{j=1}^{k} n_j z_j$, $s^2 = \frac{1}{n-1}\sum_{j=1}^{k} n_j \left(z_j - \bar{x}\right)^2$,

где n — объем выборки;

k — число интервалов группировки;

$z_j = \frac{a_j + a_{j+1}}{2}$ – середина j-го интервала.

С помощью критерия согласия Пирсона на уровне значимости $\alpha = 0.05$ выяснить, можно ли считать случайную величину X нормально распределенной с параметрами $\bar{x}$ и s, рассчитанными по выборке.

Вариант 1

$[a_j; a_{j+1})$	[1.2; 1.5)	[1.5; 1.8)	[1.8; 2.1)	[2.1; 2.4)	[2.4; 2.7)	[2.7; 3.0)
n_j	2	5	9	7	4	3

Вариант 2

$[a_j; a_{j+1})$	[2.3; 2.5)	[2.5; 2.7)	[2.7; 2.9)	[2.9; 3.1)	[3.1; 3.3)	[3.3; 3.5)
n_j	3	6	9	8	5	2

Вариант 3

$[a_j; a_{j+1})$	[3.5; 3.8)	[3.8; 4.1)	[4.1; 4.4)	[4.4; 4.7)	[4.7; 5.0)	[5.0; 5.3)
n_j	3	4	8	10	5	3

Вариант 4

$[a_j; a_{j+1})$	[1.3; 1.5)	[1.5; 1.7)	[1.7; 1.9)	[1.9; 2.1)	[2.1; 2.3)	[2.3; 2.5)
n_j	2	4	11	8	5	3

Задачи 3, 4. Критерии однородности.

Для задачи 3, 4 преподавателем генерируются две выборки заданного объема из генеральной совокупности с известными параметрами.

Даны две выборки X и Y. Проверьте гипотезу об однородности двух выборок **Задача 3** используя критерий знаков. Уровень значимости принять α=0,05; **Задача 4** используя критерий Вилкоксона. Уровень значимости принять α=0,02.

ПРИЛОЖЕНИЕ В

Варианты задания «Метод Монте-Карло»

Вариант	Интеграл I	β	ε
1.	$\int_{2}^{5} \sqrt{x^3 + 3x}\, dx$	0,92	0,02
2.	$\int_{1}^{3} \sqrt{x^3 + 8}\, dx$	0,94	0,02
3.	$\int_{-1}^{2} \sqrt{x^2 + 1}\, dx$	0,98	0,05
4.	$\int_{0,5}^{3} \sqrt{2 + x^2}\, dx$	0,94	0,03
5.	$\int_{3}^{5} \sqrt{x^3 + 2}\, dx$	0,96	0,03
6.	$\int_{0}^{2} \sqrt{8 - x^3}\, dx$	0,95	0,03
7.	$\int_{-1}^{2} \sqrt{x^4 + 1}\, dx$	0,96	0,05

ПРИЛОЖЕНИЕ Г

Варианты задания «Дисперсионный анализ»

ВАРИАНТ 1

25	21	17	28	29
31	35	23	22	26
26	25	24	27	23
21	20	22	30	33
24	23	21	32	20

ВАРИАНТ 2

53	52	45	43	48	54
49	55	38	47	46	49
48	50	51	54	52	46
44	47	46	49	50	42

ВАРИАНТ 3

34	41	42	35	39
40	43	44	43	47
34	37	40	38	32
40	38	39	36	37
41	42	45	48	50

ВАРИАНТ 4

53	69	67	62	61	62
51	64	56	57	66	67
53	58	52	51	53	50
59	59	61	60	60	69

ПРИЛОЖЕНИЕ Д

Варианты задания «Корреляционный анализ»

Часть 1. Несгруппированные данные (фактор – отклик)

ВАРИАНТ 1	ВАРИАНТ 2	ВАРИАНТ 3
0.944 4.467	1.353 -1.799	1.133 -5.82
0.97 4.05	3.075 -5.325	2.211 -10.3
2.273 5.685	4.688 -10.2	2.728 -14.89
4.107 7.325	5.017 -7.426	4.929 -33.42
4.516 7.864	5.189 -9.349	5.419 -34.05
5.978 9.382	5.687 -12.38	6.662 -45.74
6.248 11.61	6.045 -10.14	6.789 -46.05
6.281 9.913	7.505 -12.3	7.174 -45.35
8.548 12.84	7.941 -11.33	7.497 -48.97
8.946 14.79	9.181 -13.91	7.537 -50.1
9.315 15.06	9.317 -13.17	10.26 -69.67
9.4 13.9	10.44 -18.34	10.74 -73.05
9.54 14.8	10.5 -15.5	11.18 -75.34
9.6 15.2	10.6 -16.2	11.3 -72.3
9.72 14.56	10.3 -17.3	11.6 -70.2
ВАРИАНТ 4	**ВАРИАНТ 5**	**ВАРИАНТ 6**
1.353 1.527	1.133 -5.82	1.133 8.652
3.075 0.7278	2.211 -10.3	2.211 11.5
4.688 0.6984	2.728 -14.89	2.728 14.14
5.017 0.06758	4.929 -33.42	4.929 26.59
5.189 0.2915	5.419 -34.05	5.419 28.41
5.687 0.2237	6.662 -45.74	6.662 32.78
6.045 -0.147	6.789 -46.05	6.789 33.93
7.505 -0.2601	7.174 -45.35	7.174 37
7.941 -1.42	7.497 -48.97	7.497 35.86
9.181 -1.865	7.537 -50.1	7.537 35.89
9.317 -1.353	10.26 -69.67	10.26 50.86
10.44 -2.455	10.74 -73.05	10.74 49.93
10.52 -1.63	11.18 -75.34	11.18 52.02
10.7 -1.82	11.34 -70.26	11.24 46.92
10.82 -1.74	11.68 -71.22	11.56 48.64

ПРИЛОЖЕНИЕ Е

Варианты задания «Корреляционный анализ»

Часть 2. Сгруппированные данные (корреляционная таблица)

ВАРИАНТ 1

	Y					
X	4	9	14	19	24	29
10	2	3				
20		7	3			
30			2	50	2	
40				10	6	
50				4	7	3

ВАРИАНТ 2

	Y					
X	10	15	20	25	30	35
30	4	2				
40		6	4			
50			6	45	2	
60			2	8	6	
70				4	7	4

ВАРИАНТ 3

	Y					
X	15	20	25	30	35	40
5	2	6				
10		4	4			
15			7	35	8	
20			2	10	8	
25				5	6	3

ВАРИАНТ 4

	Y					
X	6	11	16	21	26	31
22	1	5				
32		5	3			
42			9	40	2	
52			4	11	6	
62				4	7	3

ВАРИАНТ 5

	Y					
X	10	15	20	25	30	35
6	4	2				
12		6	2			
18			5	40	5	
24			2	8	7	
30				4	7	8

ВАРИАНТ 6

	Y					
X	4	9	14	19	24	29
20				4	7	3
30				9	7	
40			4	45	5	
50		4	6			
60	1	4				

ПРИЛОЖЕНИЕ Ж

Варианты задания «Регрессионный анализ»

По 12 предприятиям региона изучается зависимость выработки продукции на одного работника y (тыс. руб.) от ввода в действие новых основных фондов x_1 (% от стоимости фондов на конец года) и от удельного веса рабочих высокой квалификации в общей численности рабочих x_2 (%) (смотри таблицу своего варианта). Построить модель множественной линейной регрессии, найти оценки параметров, определить доверительные интервалы для параметров, проверить адекватность модели. Ошибки наблюдений независимы и имеют распределение $N(0,\sigma)$.

Вариант 1				**Вариант 2**			
Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	6	3,6	9	1	6	3,5	10
2	6	3,6	12	2	6	3,6	12
3	6	3,9	14	3	7	3,9	15
4	7	4,1	17	4	7	4,1	17
5	7	3,9	18	5	7	4,2	18
6	7	4,5	19	6	8	4,5	19
7	8	5,3	19	7	8	5,3	19
8	8	5,3	19	8	9	5,3	20
9	9	5,6	20	9	9	5,6	20
10	10	6,8	21	10	10	6	21
11	9	6,3	21	11	10	6,3	21
12	11	6,4	22	12	11	6,4	22

ПРИЛОЖЕНИЕ 3

Варианты задания «Временные ряды»

В таблице представлены сведения о доходах Y (одинаковые для всех вариантов), расходах на промышленные товары X (по вариантам) в течение 22 месяцев

Y	Вариант №				
	1	2	3	4	5
	X	X	X	X	X
91,76	16,34	10,90	24,25	3,03	7,73
38,68	10,49	6,99	8,28	7,81	6,07
34,14	5,30	4,08	4,21	1,63	0,28
30,77	13,79	10,61	12,95	4,49	9,18
50,02	2,03	1,57	2,40	0,43	2,63
34,33	9,65	7,43	2,42	6,31	8,51
42,63	13,91	10,70	11,80	5,05	17,05
63,47	3,24	2,16	1,01	3,94	0,85
19,86	2,20	1,47	1,86	0,29	0,91
58,87	12,82	11,65	4,28	5,37	2,33
72,45	29,44	26,77	29,97	6,54	10,96
29,70	8,03	7,30	1,25	0,93	6,98
93,74	33,44	22,29	39,73	1,82	32,73
17,77	0,60	0,40	0,74	0,51	0,72
78,84	32,66	23,33	41,47	15,87	7,89
39,73	6,24	4,46	2,40	1,78	1,48
93,87	26,48	18,91	24,48	25,53	20,85
86,15	25,31	16,87	20,51	31,97	27,76
25,95	2,27	2,06	1,85	2,28	0,19
36,95	12,05	8,03	10,88	10,92	13,03
45,78	20,65	17,20	3,11	12,76	3,41
12,36	0,23	0,15	0,26	0,05	0,34

ПРИЛОЖЕНИЕ И

Варианты задания «Цепи Маркова»

ВАРИАНТ 1

1. Частица на прямой может иметь координаты x=1,2,3,4. Каждую секунду частица может совершать единичные скачки влево или вправо с вероятностями соответственно 0,3 и 0,7. Из положения x=1 частица с вероятностью 0,7 переходит в точку x=2 и с вероятностью 0,3 остается на месте, а из положения x=4 она с вероятностью 0,7 остается на месте, а с вероятностью 0,3 перейдет в положение x=3. Составить матрицу перехода блужданий и орграф.

Является ли ЦМ регулярной? Найти предельные вероятности.

2. Найти фундаментальную матрицу и матрицу B. Какую информацию о цепи Маркова содержат эти матрицы?

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ \frac{1}{5} & \frac{2}{5} & 0 & \frac{2}{5} \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

ВАРИАНТ 2

1. На окружности расположены точки $A_1, A_2, ..., A_4$ – вершины правильного четырехугольника. Частица движется из точки в точку следующим образом: из данной точки она перемещается в одну из ближайших соседних точек с вероятностью 0,5. Построить матрицу перехода и орграф данной цепи Маркова. Какова вероятность частице оказаться в этой же точке через два шага? Является ли ЦМ регулярной? Исследовать поведение системы при $t \to \infty$.

2. Дана матрица перехода *P*. Найти фундаментальную матрицу и матрицу B. Какую информацию о цепи Маркова содержат эти матрицы?

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{2} & \frac{1}{4} \\ \frac{2}{3} & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$