Федеральное государственное бюджетное образовательное учреждение
высшего образования
«КАЛИНИНГРАДСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Е. Ю. Скоробогатых

ПРИКЛАДНАЯ СТАТИСТИКА

Учебно-методическое пособие по изучению дисциплины
для студентов всех специальностей и направлений
(*модуль саморазвития*)

Калининград
Издательство ФГБОУ ВО «КГТУ»
2024

Рецензенты:
кандидат физико-математических наук, доцент кафедры прикладной математики и информационных технологий ФГБОУ ВО «Калининградский государственный технический университет» А. И. Руденко
кандидат технических наук, доцент кафедры теории машин и механизмов и деталей машин ФГБОУ ВО «Калининградский государственный технический университет» О. С. Витренко

Скоробогатых, Е. Ю.
Прикладная статистика: учебно-методическое пособие по изучению дисциплины для студентов всех специальностей и направлений (модуль саморазвития) / Е. Ю. Скоробогатых. – Калининград: Издательство ФГБОУ ВО «КГТУ». – 2024. – 67 с.

Учебно-методическое пособие является руководством по изучению дисциплины «Прикладная статистика» модуля саморазвития для студентов всех специальностей и направлений подготовки, выбравших данную дисциплину для изучения. Содержит характеристику дисциплины (цель и планируемые результаты изучения дисциплины, место дисциплины в структуре основной профессиональной образовательной программы, описание видов и процедур текущего контроля и промежуточной аттестации), тематический план с описанием для каждой темы форм проведения занятия, вопросов для изучения, методических материалов к занятию, методических указаний по выполнению самостоятельной работы.
Табл. 4, рис. 1, список лит. – 3 наименования

Учебно-методическое пособие по изучению дисциплины рекомендовано к использованию в учебном процессе в качестве локального электронного методического материала методической комиссией ИЦТ 3 декабря 2024 г., протокол № 8

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Учебно-методическое пособие представляет комплекс систематизированных материалов для изучения дисциплины «Прикладная статистика» модуля саморазвития для студентов всех специальностей и направлений, выбравших данную дисциплину для изучения.

Прикладная статистика является одной из ключевых дисциплин, играющих важную роль в современном техническом образовании. В условиях стремительного развития науки и технологий, а также увеличения объема данных, с которыми сталкиваются специалисты, умение эффективно анализировать и интерпретировать статистическую информацию становится необходимым навыком для будущих инженеров и ученых.

Пособие охватывает основные концепции и методы статистического анализа, которые могут быть применены в различных областях техники и науки, включая инженерные исследования, управление качеством, экономику и другие прикладные направления. Значительное внимание уделяется использованию современных статистических программных средств (Excel, Loginom, R, Python), что позволяет студентам развивать навыки работы с данными и проводить анализ с использованием актуальных инструментов.

Целью освоения дисциплины является получение систематизированных знаний об основных закономерностях и особенностях математической и прикладной статистики, ее практических приложений; об инструментарии, связанном с анализом данных в области будущей профессиональной деятельности; выработка навыков получения, анализа и обобщения информации, построения математических моделей объектов и процессов профессиональной деятельности.

В результате освоения дисциплины обучающийся должен:

знать:

– терминологию прикладной статистики;

– общие методы поиска, критического анализа и синтеза информации, в том числе специальные методы, применяемые в статистике;

уметь:

– анализировать условия и ограничения поставленной задачи, интерпретировать и ранжировать информацию по ней на основе имеющейся статистической информации;

– выбирать методы и средства решения задачи, оценивая их достоинства и недостатки;

– анализировать методологические проблемы, возникающие при решении;

владеть:

– навыками анализа условий и ограничений поставленной задачи;

– практическим навыком применения методов и концепций прикладной статистики для построения математических моделей процессов и явлений и проведения расчетов, а также интерпретации полученных результатов.

Дисциплина «Прикладная статистика» относится к модулю Б1.О.ДЭ.1 саморазвития (элективные дисциплины) (Б1.О.ДЭ.1.35, Б1.О.ДЭ.1.36) основной профессиональной образовательной программы высшего образования по всем специальностям и направлениям подготовки КГТУ.

При изучении дисциплины используются знания, умения и навыки довузовской подготовки по математике, элементы матричной алгебры, основные понятия и инструменты дифференциального и интегрального исчисления.

Дисциплина является базой при изучении дисциплин математического и естественнонаучного модуля, инженерно-технического модуля.

Общая трудоемкость дисциплины составляет 2 зачетные единицы (з.е.), т. е. 72 академических часа контактной и самостоятельной учебной работы студента; работы, связанной с текущей и промежуточной (заключительной) аттестацией по дисциплине. Трудоемкость и структура дисциплины при обучении представлены в таблице 1.

Таблица 1 – Объем (трудоёмкость освоения) и структура дисциплины

Наименование	**Семестр**	**Форма контроля**	**З.е.**	**Акад. часов**	**Контактная работа**					**СРС**
					лек.	лаб.	пр.	РЭ	КА	
Прикладная статистика	2 или 3	Зачет	2	72	16	-	16	3	0,15	36,85

Обозначения: Э – экзамен; К – контрольная работа, РГР – расчетно-графическая работа; КР – курсовая работа; лек. – лекционные занятия; лаб. - лабораторные занятия; пр. – практические занятия; РЭ – контактная работа с преподавателем в ЭИОС; КА – контактная работа, включающая консультации, инд. занятия, практики и аттестации; СРС – самостоятельная работа студентов

Основными видами аудиторных учебных занятий по дисциплине являются: лекции, лабораторные и практические занятия.

Формирование знаний, обучающихся обеспечивается проведением лекционных занятий.

Изучение дисциплины сопровождается практическими занятиями, в ходе которых происходит закрепление теоретических знаний, формирование и совершенствование умений, навыков и компетенций.

В ходе изучения дисциплины предусматривается применение эффективных методик обучения, которые предполагают постановку вопросов проблемного характера с разрешением их, как непосредственно в ходе занятий, так и в ходе самостоятельной работы. Обучающимся рекомендуется широкое использование ПЭВМ и средств компьютерного моделирования. В этом плане роль консультаций сводится, в основном, к помощи в изучении методов решения статистических задач.

Контроль знаний в ходе изучения дисциплины осуществляется в виде текущего контроля и промежуточной аттестации в форме зачета.

Текущий контроль (контроль выполнения заданий на практических занятиях и заданий для самостоятельной работы) предназначен для проверки хода и качества усвоения курсантами/студентами учебного материала и стимулирования их учебной работы. Он может осуществляться в ходе всех видов занятий в форме, избранной преподавателем или предусмотренной рабочей программой дисциплины.

Текущий контроль предполагает постоянный контроль преподавателем качества усвоения учебного материала, активизацию учебной деятельности курсантов/студентов на занятиях, побуждение их к самостоятельной систематической работе. Он необходим обучающимся для самоконтроля на разных этапах обучения. Их результаты учитываются выставлением преподавателем оценок в журнале учета успеваемости и в ходе ежемесячной аттестации.

При текущем контроле успеваемости учитывается:

– выполнение обучающимся всех работ и заданий, предусмотренных рабочей программой дисциплины, а именно:

– выполнение заданий на практических занятиях;

– самостоятельную работу обучающихся;

– посещаемость аудиторных занятий (занятий с применением ДОТ).

Промежуточная аттестация.

Промежуточная аттестация в форме зачета (второй или третий семестр) проходит по результатам прохождения всех видов текущего контроля успеваемости. В отдельных случаях (при не прохождении всех видов текущего контроля) зачет может быть проведен в виде тестирования.

1. ТЕМАТИЧЕСКИЙ ПЛАН ДИСЦИПЛИНЫ

Тематический план и трудоёмкость освоения дисциплины представлена в таблице 2.

Таблица 2 – Трудоёмкость освоения дисциплины

№ п/п	Модуль (тема)	Контактная работа с преподавателем					СРС
		ЛК	ЛР	ПР	РЭ	КА	
1	Введение в статистику. Основные понятия и определения. Основные сведения из теории вероятностей	2		2			4
2	Выборочные исследования. Предобработка статистических данных, визуализация. Описательная статистика	2		2	0,5		4
3	Статистическое оценивание параметров. Точечные и интервальные оценки	2		2			4
4	Проверка статистических гипотез	4		4	1		6
5	Дисперсионный анализ (ANOVA)	2		2	0,5		6
6	Анализ зависимостей	2		2	0,5		6
7	Линейный регрессионный анализ	2		2	0,5		7
ИТОГО:		16		16	47	0,15	36,85

2. СОДЕРЖАНИЕ И МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ИЗУЧЕНИЮ ДИСЦИПЛИНЫ

2.1. Тема 1. Введение в прикладную статистику. Основные понятия и определения. Основные сведения из теории вероятностей

Вопросы для изучения

1. Цели задачи прикладной статистики
2. Основные понятия: генеральная совокупность и выборка; переменные: качественные и количественные; статистические данные
3. Методы сбора и предобработки данных.
4. Вероятность, основные теоремы и формулы теории вероятностей.
5. Случайная величина, закон распределения случайной величины.
6. Некоторые модельные распределения (биномиальное, экспоненциальное, нормальное)
7. Закон больших чисел.
8. Компьютерные технологии анализа данных

Методические указания

Математическая статистика – раздел математики, в котором изучаются методы сбора, систематизации и обработки результатов наблюдений массовых случайных явлений для выявления существующих закономерностей.

Математическая статистика тесно связана с теорией вероятностей. Оба эти раздела изучают массовые случайные явления. Связующим звеном между ними являются предельные теоремы теории вероятностей. При этом теория вероятностей выводит из математической модели свойства реального процесса, а математическая статистика устанавливает свойства математической модели, исходя из данных наблюдений реального процесса (из статистических данных).

Средства математической статистки позволяют решать следующие задачи:

– провести предварительную обработку статистических данных и представить их в удобном для дальнейшего изучения и анализа виде;

– оценить неизвестные характеристики наблюдаемой случайной величины (например, неизвестные вероятность события, функцию распределения, математическое ожидание, дисперсию, параметры неизвестного распределения);

– осуществить проверку статистических гипотез, то есть дать обоснованные выводы о согласовании результатов оценивания с опытными данными.

Результаты исследования статистических данных методами математической статистики используются для принятия решений в задачах планирования, управления, прогнозирования в экономических и технических системах. Говорят, что математическая статистика – это теория принятия решений в условиях неопределенности.

Рекомендуемые источники: [1, гл. 1–4]; [2, т. 1, гл. 1–2].

Основные теоретические сведения и решение типовых задач

При *классическом определении* за вероятность события A принимают отношение числа благоприятных этому событию исходов m к числу всех возможных исходов опыта n.

$$P(A) = \frac{m}{n}.$$

При нахождении числа благоприятных или всех возможных исходов опыта используются формулы *комбинаторики*.

Пусть дано множество из n элементов.

Перестановками называются комбинации, составленные из всех n элементов данного множества, которые отличаются только порядком следования в них элементов. Общее число перестановок из n элементов определяется по формуле

$$P_n = n!.$$

Размещениями из n элементов по m называются комбинации, которые отличаются составом элементов и порядком их следования. Их общее число находится по формуле

$$A_n^m = \frac{n!}{(n-m)!}.$$

Сочетаниями из n элементов по m называются комбинации, которые отличаются только составом элементов. Общее число сочетаний определяется по формуле

$$C_n^m = \frac{n!}{m!(n-m)!}.$$

Теоремы сложения и умножения вероятностей

Суммой событий называется событие, состоящее в появлении хотя бы одного из рассматриваемых событий.

События называются *несовместными*, если они не могут появиться одновременно в одном опыте. В противном случае события называются *совместными*.

Произведением событий называется событие, состоящее в появлении всех рассматриваемых событий.

События называются *независимыми*, если появление одного события не влияет на вероятность появления другого. В противном случае события называются *зависимыми*.

Вероятность события B, вычисленная при условии, что произошло событие A, называется условной вероятностью и обозначается $P_A(B)$

Теорема сложения вероятностей несовместных событий

$P(A_1 + A_2 + \cdots + A_n) = P(A_1) + P(A_2) + \cdots + P(A_n).$

Теорема сложения вероятностей двух совместных событий

$P(A_1 + A_2) = P(A_1) + P(A_2) - P(A_1 \cdot A_2).$

Теорема умножения вероятностей двух событий

$P(A \cdot B) = P(A) \cdot P_A(B).$

Если события независимы, то

$P(A \cdot B) = P(A) \cdot P(B).$

Замечание. Теорему умножения вероятностей можно обобщить на любое конечное число событий.

$P(A_1 \cdot A_2 \cdot \ldots \cdot A_n) = P(A_1)P_{A_1}(A_2)P_{A_1,A_2}(A_3) \ldots P_{A_1,A_2,\ldots A_{n-1}}(A_n),$

или для несовместных событий

$P(A_1 \cdot A_2 \cdot \ldots \cdot A_n) = P(A_1)P(A_2) \ldots P(A_n).$

Формула полной вероятности. Формула Байеса

Пусть $H_1, H_2, \ldots, H_n$ - единственно возможные попарно несовместные события (гипотезы). Событие B может произойти только с одним из $H_1, H_2, \ldots, H_n$.

Для нахождения вероятности события B используется *формула полной вероятности*

$$P(B) = \sum_{i=1}^{n} P(H_i) \cdot P_{H_i}(B).$$

Для определения вероятности H_i при условии, что событие B наступило, используется *формула Байеса*

$$P_B(H_i) = \frac{P(H_i) \cdot P_{H_i}(B)}{\sum_{i=1}^{n} P(H_i) \cdot P_{H_i}(B)}.$$

Формула Бернулли

Пусть производится n независимых испытаний, в каждом из которых событие A наступает с вероятностью p. Вероятность того, что событие A наступит ровно m раз в n испытаниях определяется по *формуле Бернулли*

$$P_n(m) = C_n^m p^m q^{n-m}, q = 1 - p$$

вероятность того, что событие наступит

а) менее m раз: $P = P_n(0) + P_n(1) + \cdots P_n(m-1)$,

б) более m раз: $P = P_n(m+1) + P_n(m+2) + \cdots P_n(n)$,

в) не менее m раз: $P = P_n(m) + P_n(m+1) + \cdots P_n(n)$,

г) не более m раз: $P = P_n(0) + P_n(1) + \cdots P_n(m)$.

Дискретные случайные величины

Случайной величиной называется величина, которая в результате опыта может принять одно из множества своих возможных значений (заранее не известно какое). Различают дискретные и непрерывные случайные величины.

Дискретной случайной величиной называется величина множество всех возможных значений которой есть конечное или счетное множество фиксированных величин.

Законом (рядом) распределения вероятностей дискретной случайной величины называют последовательность возможных значений случайной величины и соответствующих им вероятностей:

X	x_1	x_2	…	x_n
p	p_1	p_2	…	p_n

причем

$$\sum_{i=1}^{n} p_i = 1$$

Функцией распределения случайной величины называется функция $F(x)$, которая равна вероятности случайного события, состоящего в том, что

дискретная случайная величина X примет значение меньшее некоторого значения *x*, т. е.

$F(x) = P(X < x)$.

Для дискретной случайной величины функцию распределения можно задать в виде

$$F(x) = \begin{cases} 0, x \le x_1 \\ p_1, x_1 < x \le x_2 \\ p_1 + p_2, x_2 < x \le x_3 \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ 1, x > x_n \end{cases}$$

Основными числовыми характеристиками случайных величин являются математическое ожидание, дисперсия, среднее квадратическое отклонение.

Математическим ожиданием дискретной случайной величины называется величина

$$M(X) = \sum_{i=1}^{n} x_i \cdot p_i.$$

Математическое ожидание характеризует среднее значение случайной величины.

Дисперсией случайной величины X называется математическое ожидание квадрата отклонения случайной величины от ее математического ожидания:

$$D(X) = M(X - M(X))^2$$

Дисперсию целесообразно вычислять по формуле

$$D(X) = M(X^2) - (M(X))^2$$

Средним квадратическим отклонением называют величину

$\sigma(X) = \sqrt{D(X)}$.

Дисперсия и среднее квадратическое отклонение характеризуют рассеяние значений случайной величины около ее среднего значения.

Непрерывные случайные величины

Непрерывной называется случайная величина множество всех возможных значений которой есть непрерывный конечный или бесконечный интервал.

Функция распределения вероятностей определяется формулой

$F(x) = P(X < x)$.

Функция распределения непрерывной случайной величины является непрерывной функцией.

Плотностью распределения или плотностью вероятностей называется производная от функции распределения:

$f(x) = F'(x)$.

Плотность распределения должна удовлетворять следующим свойствам:

$$1) f(x) \ge 0$$

$$2)\int_{-\infty}^{+\infty} f(x)dx = 1$$

Вероятность попадания непрерывной случайной величины в заданный интервал определяется формулой

$P(a < X < b) = \int_a^b f(x)dx.$

Математическое ожидание непрерывной случайной величины определяется формулой

$M(X) = \int_{-\infty}^{+\infty} x \cdot f(x)dx.$

Дисперсия вычисляется по формулам

$D(X) = \int_{-\infty}^{+\infty} \big(x - M(X)\big)^2 f(x)dx$ или $D(X) = \int_{-\infty}^{+\infty} x^2 f(x)dx - \big(M(X)\big)^2.$

Нормальное распределение

Нормальным распределением называют распределение непрерывной случайной величины, плотность распределения которой имеет вид

$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}},$

где a – математическое ожидание, σ – среднее квадратическое отклонение нормальной случайной величины.

Вероятность попадания нормальной случайной величины в интервал (α, β)находится по формуле

$P(\alpha < X < \beta) = \Phi\left(\frac{\beta-a}{\sigma}\right) - \Phi\left(\frac{\alpha-a}{\sigma}\right),$

где Ф(х) – интегральная функция Лапласа.

Вероятность отклонения от среднего на величину меньшую δ выражается равенством

$P(|X - a| < \delta) = 2\Phi\left(\frac{\delta}{\sigma}\right).$

Решение типовых задач

Непосредственное вычисление вероятностей

Задача 1. Из 30 студентов 10 имеют спортивные разряды. Какова вероятность того, что выбранные наудачу 3 студента – разрядники?

Решение.

Пусть событие A – 3 выбранных наудачу студента являются разрядниками. Общее число случаев выбора трех студентов из тридцати равно $n = C_{30}^3$, так как комбинации представляют собой сочетания, ибо отличаются только составом студентов. Точно также число студентов, благоприятствующих событию A, равно $m = C_{10}^3$. Итак,

$$P(A) = \frac{m}{n} = \frac{C_{10}^3}{C_{30}^3} = \frac{61}{203} \approx 0{,}030.$$

Задача 2. В лифт на первом этаже девятиэтажного дома вошли 4 человека, каждый из которых может выйти независимо друг от друга на любом этаже. Какова вероятность того, что все пассажиры выйдут: а) на 6-ом этаже; б) на одном этаже?

Решение.

а) Пусть событие *A* – все пассажиры выйдут на 6-ом этаже. Каждый пассажир может выйти со 2-го по 9-ый этаж 8 способами. По правилу произведения общее число способов выхода четырех пассажиров из лифта равно $n = 8 \cdot 8 \cdot 8 \cdot 8$. Число случаев, благоприятствующих событию *A,* равно $m = 1$. Следовательно,

$$P(A) = \frac{1}{8^4} = 0{,}00024.$$

б) Пусть событие *B* – все пассажиры выйдут на одном этаже. Теперь событию *B* будут благоприятствовать $m = 8$ случаев (все выйдут на 2 этаже, 3-м,…, 9-м этаже). Поэтому $P(B) = {}^{8}/_{8^4} = 0{,}00195$.

Формула полной вероятности. Формула Байеса

Задача 3. Частица пролетает мимо трех счетчиков, причем она может попасть в каждый из них с вероятностью 0,3, 0,2, 0,4. Если частица попадает в первый счетчик, то она регистрируется с вероятностью 0,6, во второй – с вероятностью 0,5 и в третий – с вероятностью 0,55. Найти вероятность того, что частица будет зарегистрирована (событие А).

Решение. Выдвигаем гипотезы:

H_1 – частица попадает в первый счетчик $P(H_1) = 0{,}3$,

H_2 – частица попадает во второй счетчик $P(H_2) = 0{,}2$,

H_3 – частица попадает в третий счетчик $P(H_3) = 0{,}4$.

Эти события не пересекаются, но не составляют полной группы. Чтобы получить полную группу добавим событие H_4 – частица не попадает ни в один счетчик

$$P(H_4) = 1 - 0{,}3 - 0{,}2 - 0{,}4 = 0{,}1.$$

Условные вероятности равны:

$$P(A/H_1) = 0{,}6;\ P(A/H_2) = 0{,}5;\ P(A/H_3) = 0{,}55;\ P(A/H_4) = 0.$$

По формуле полной вероятности имеем:

$$P(A) = 0{,}3 \cdot 0{,}6 + 0{,}2 \cdot 0{,}5 + 0{,}4 \cdot 0{,}55 + 0{,}1 \cdot 0 = 0{,}5.$$

Задача 4. Три завода выпускают одинаковые изделия, причем первый завод производит 50, второй – 20, третий – 30 % всей продукции. Первый завод

выпускает 1 % брака, второй – 8 %, третий – 3 %. Наудачу выбранное изделие оказалось бракованным (событие А). Найти вероятность того, что оно изготовлено на втором заводе.

Решение.

Гипотезы:

H_1 – изделие изготовлено на первом заводе, $P(H_1) = 0,5$;

H_2 – изделие изготовлено на втором заводе, $P(H_2) = 0,2$;

H_3 – изделие изготовлено на третьем заводе, $P(H_3) = 0,3$.

По условию задачи: $P(A/H_1) = 0,01$, $P(A/H_2) = 0,08$, $P(A/H_3) = 0,03$.

Окончательно имеем:

$$P(H_2/A) = 0,2\cdot0,08/(0,5\cdot0,01+0,2\cdot0,08+0,3\cdot0,03) = 8/15.$$

Схема Бернулли: формула Бернулли, приближенные формулы Муавра-Лапласа и Пуассона

Задача 5. Вероятность попадания в мишень при одном выстреле для данного стрелка равна 0,8 и не зависит от номера выстрела. Требуется найти вероятность того, что при 5 выстрелах произойдет ровно 2 попадания в мишень.

Решение.

В этом примере *n=5, p=0,8 и m=2;* по формуле Бернулли находим:

$$P_5(2) = C_5^2 0,8^2 0,2^3 = 0,0512.$$

Задача 6. Вероятность наступления события А в каждом из 900 независимых испытаний равна *p=0,8.* Найдите вероятность того, что событие А произойдет: а) 750 раз; б) от 710 до 740 раз.

Решение.

а) Воспользуемся локальной формулой Муавра-Лапласа.

$x = \frac{750-900\cdot0,8}{\sqrt{900\cdot0,8\cdot0,2}} = 2,5$. По приложению 1 пособия 2 находим: $\phi(2,5) = 0,0175$Тогда

$$P_{900}(750) \approx \frac{0,0175}{\sqrt{900\cdot0,8\cdot0,2}} = 0,00146.$$

б) $x_1 = \frac{710-900\cdot0,8}{\sqrt{900\cdot0,8\cdot0,2}} = -0,83;$ $x_2 = \frac{740-900\cdot0,8}{\sqrt{900\cdot0,8\cdot0,2}} = 1,67$

По приложению 2 пособия 2 находим:

$$\Phi(-0,83) = -\Phi(0,83) = -0,2967; \quad \Phi(1,67) = 0,4527.$$

Окончательно имеем:

$$P_{900}(710 \le m \le 740) = 0,4527 + 0,2967 = 0,7492.$$

Задача 7. В тираже «Спортлото 6 из 49» участвуют 10 000 000 карточек. Найти вероятность события А – хотя бы в одной из них зачеркнуты все 6 выигрышных номеров.

Решение. Перейдем к противоположному событию – ни на одну карточку не выпал максимальный выигрыш $\bar{A}$. В каждой карточке номера зачеркиваются случайным образом и не зависят от других карточек, поэтому применима схема Бернулли с параметрами $n = 10000000, p = \frac{C_6^6 C_{43}^0}{C_{49}^6} = 7 \cdot 10^{-8}$. Поскольку λ = np = 0,7, то для определения воспользуемся формулой Пуассона. Тогда

$P(\bar{A}) = P_{10000000}(0) \approx P(0; 0{,}7) = 0{,}49659; P(A) = 0{,}50341$. Таким образом, вероятность, что из 10 000 000 карточек, хотя бы одна окажется с максимальным выигрышем чуть больше ½.

Случайные величины. Основные законы распределения. Числовые характеристики случайных величин

Задача 8. На зачете студент получил 4 задачи. Вероятность решить правильно каждую задачу равна 0,8. Определить ряд распределения случайной величины – числа правильно решенных задач и построить многоугольник распределения. Найти функцию распределения и построить ее график.

Решение.

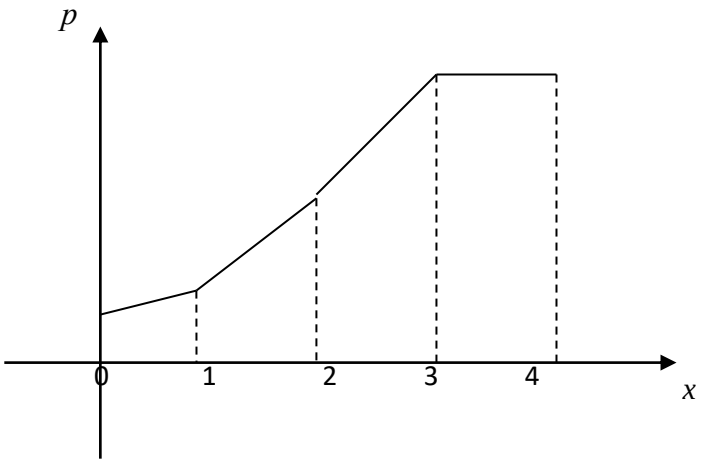
Возможные значения случайной величины X: 0, 1, 2, 3, 4. Соответствующие вероятности вычисляем по формуле Бернулли:

$$P(X = 0) = C_4^0 0,8^0 0,2^4 = 0{,}0016;$$
$$P(X = 1) = C_4^1 0,8^1 0,2^3 = 0{,}0256;$$
$$P(X = 2) = C_4^2 0,8^2 0,2^2 = 0{,}1536;$$
$$P(X = 3) = C_4^3 0,8^3 0,2^1 = 0{,}4096;$$
$$P(X = 4) = C_4^4 0,8^4 0,2^0 = 0{,}4096.$$

x	0	1	2	3	4
p	0,0016	0,0256	0,1536	0,4096	0,4096

Проверка: 0,0016 + 0,0256 + 0,1536 + .0,4096+ 0,4096 = 1

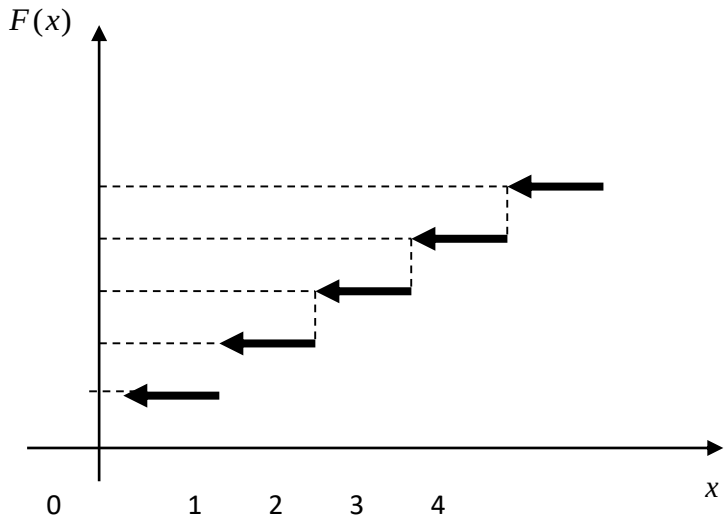
Многоугольник распределения имеет вид:



Используя данные из таблицы и формулу для функции распределения $F(x) = P(X < x). = \sum_{x_i<x} p_i$, получим функцию распределения:

$$F(x) = \begin{cases} 0, x \le 0, \\ 0{,}0016, 0 < x \le 1, \\ 0{,}0272, 1 < x \le 2, \\ 0{,}1808, 2 < x \le 3, \\ 0{,}5904, 3 < x \le 4, \\ 1, x > 4. \end{cases}$$

Построим ее график.



Задача 9. Задан закон распределения дискретной случайной величины:

x	-1	0	1	2
p	0,1	0,3	0,5	0,1

Найти математическое ожидание и дисперсию.

Решение.

$$M(x) = -1 \cdot 0{,}1 + 0 \cdot 0{,}3 + 1 \cdot 0{,}5 + 2 \cdot 0{,}1 = 0{,}6$$

$$D(x) = (-1)^2 \cdot 0{,}1 + 0^2 \cdot 0{,}3 + 1^2 \cdot 0{,}5 + 2^2 \cdot 0{,}1 - 0{,}6^2 = 0{,}64$$

Закон больших чисел. Предельные теоремы

Задача 10. Среднее количество вызовов, поступающих на коммутатор завода в течение часа, равно 300. Оценить вероятность того, что в течение следующего часа число вызовов на коммутатор превысит 400.

Решение. По условию *M(X) = 300*. Согласно неравенству Маркова

$$P(X > 400) \leq \frac{300}{400},$$ т. е.

вероятность того, что число вызовов превысит 400, будет не более 0,75.

Контрольные вопросы

1. Что такое прикладная статистика?
2. Каковы основные цели прикладной статистики?
3. Каковы основные этапы проведения статистического исследования?
4. Какие ошибки могут возникнуть при интерпретации статистических данных?
5. Что такое вероятность?
6. Каковы основные правила вычисления вероятности?
7. Что такое независимые события?
8. Что такое зависимые события?
9. Как рассчитывается условная вероятность?
10. Какие основные теоремы теории вероятностей необходимо знать?
11. Что такое нормальное распределение?
12. Почему нормальное распределение важно в статистике?

2.2 Тема 2 Выборочные исследования. Предобработка статистических данных, визуализация. Описательная статистика

Вопросы для изучения

1. Генеральная и выборочная совокупности. Вариационный ряд. Группировка данных. Статистическое распределение выборки. Эмпирическая функция распределения и ее свойства.

2. Визуализация данных. Полигон, гистограмма, диаграмма, коробчатые графики.

3. Описательная статистика: меры центральной тенденции (средняя, мода, медиана), меры разброса (размах, дисперсия, стандартное отклонение, коэффициент вариации), асимметрия и эксцесс

4. Компьютерная реализация рассмотренных методов.

Методические указания

Для освоения темы «Описательная статистика, группировка и визуализация данных», начните с базовых понятий статистики, таких как среднее значение, медиана, мода, дисперсия и стандартное отклонение. Узнайте о различных мерах центральной тенденции и разброса. Изучи методы группировки данных, включая построение полигонов, гистограмм и частотных таблиц. Визуализируйте данные с помощью графиков (диаграмм, столбчатых диаграмм, коробчатых диаграмм и др.). Для практики используйте компьютерные инструменты анализа данных.

Рекомендуемые источники: [1, гл. 8]; [2, т. 1, гл. 6].

Программное обеспечение

– Python (библиотеки scipy.stats, statsmodels);

– R (base R: основные функции для описательной статистики (mean, median, sd, var, summary) встроены прямо в ядро R; визуализация: ggplot2 - поддерживает широкий спектр типов диаграмм, включая гистограммы, диаграммы рассеяния, линейные графики и многое другое.);

– Excel (инструмент «Анализ данных», описательная статистика);

– SPSS (встроенные функции для описательной статистики и визуализации).

Основные теоретические сведения и решение типовых задач

Основными понятиями математической статистики являются генеральная совокупность и выборка.

Генеральная совокупность – это совокупность всех мысленно возможных объектов данного вида, над которыми проводятся наблюдения с целью получения конкретных значений определенной случайной величины.

Генеральная совокупность может быть *конечной* или *бесконечной* в зависимости от того, конечна или бесконечна совокупность составляющих ее объектов.

Выборкой (выборочной совокупностью) называется совокупность случайно отобранных объектов из генеральной совокупности.

Выборка должна быть *репрезентативной (представительной)*, то есть ее объекты должны достаточно хорошо отражать свойства генеральной совокупности.

Выборка может быть *повторной*, при которой отобранный объект (перед отбором следующего) возвращается в генеральную совокупность, и *бесповторной*, при которой отобранный объект не возвращается в генеральную совокупность.

Число N объектов генеральной совокупности и число n объектов выборки называют объемами генеральной и выборочной совокупностей соответственно. При этом предполагают, что $N >> n$ (значительно больше).

Вариационные ряды

Полученные различными способами отбора данные образуют выборку, обычно это множество чисел, расположенных в беспорядке. По такой выборке трудно выявить какую-либо закономерность их изменения (*варьирования*).

Для обработки данных используют операцию *ранжирования*, которая заключается в том, что результаты наблюдений над случайной величиной, то есть наблюдаемые значения случайной величины располагают в порядке возрастания.

Пример 1. Дана выборка: 2,4,7,3,1,1,3,2,7,3

Проведем ранжирование выборки: 1,1,2,2,3,3,3,4,7,7

После проведения операции ранжирования значения случайной величины объединяют в группы, то есть группируют так, что в каждой отдельной группе значения случайной величины одинаковы. Каждое такое значение называется *вариантом.* Варианты обозначаются строчными буквами латинского алфавита с индексами, соответствующими порядковому номеру группы x_i , y_j, ...

Изменение значения варианта называется *варьированием***.**

Последовательность вариантов, записанных в возрастающем порядке, называется *вариационным рядом*.

Число, которое показывает, сколько раз встречаются соответствующие значения вариантов в ряде наблюдений, называется *частотой* или *весом варианта* и обозначается n_i, где i - номер варианта.

Отношение частоты данного варианта к общей сумме частот называется относительной частотой или *частостью* (*долей*) соответствующего варианта и обозначается $\mathrm{p}_i^* = \left(\frac{n_i}{n}\right)$ или $\mathrm{p}_i^* = \frac{n_i}{\sum_{i=1}^{m} n_i}$, где m – число вариантов. Частость является статистической вероятностью появления варианта x_i. Естественно считать частость p_i^* аналогом вероятности p_iпоявления значения x_i случайной величины X.

Дискретным статистическим рядом называется ранжированная совокупность вариантов (x_i) с соответствующими им частотами (n_i) или частостями (p_i^*).

Дискретный статистический ряд удобно записывать в виде таблицы

$\boldsymbol{x_i}$	1	2	3	4	7
$\boldsymbol{n_i}$	2	2	3	1	2
$\frac{\boldsymbol{n_i}}{\boldsymbol{n}}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{2}{10}$

$\sum_{i=1}^{5} n_i = 10; \sum_{i=1}^{5} p_i^* = 1.$

Характеристики дискретного статистического ряда:

1. *Размах варьирования* $R = x_{max} - x_{min}$.

2. *Мода* (M_0^*) – вариант, имеющий наибольшую частоту (в примере 1 $M_0^* = 3$).

3. *Медиана* (M_e^*) – значение случайной величины, приходящееся на середину ряда.

Пусть n – объем выборки.

Если $n = 2k$, то есть ряд имеет четное число членов, то $M_e^* = \frac{x_k + x_{k+1}}{2}$. Если $n = 2k + 1$, т. е. ряд имеет нечетное число членов, то $M_e^* = x_{k+1}$ (в примере 1. $M_e^* = 3$).

Если изучаемая случайная величина X является непрерывной или число значений ее велико, то составляют *интервальный статистический ряд.*

Для его составления необходимо:

1. Определить число интервалов статистического ряда по формуле Стерджеса:

$$m \approx (1 + 3{,}322\, lg\, n)$$

2. Определить длину частичного интервала (шаг)h:

$$h = \frac{x_{max} - x_{min}}{m} \text{ или } h = \frac{x_{max} - x_{min}}{1 + 3{,}322\, lg\, n}.$$

Если шаг окажется дробным, то за длину интервала берут ближайшее целое число или ближайшую простую дробь (обычно берут интервалы одинаковые по длине, но могут быть интервалы и разной длины).

3. Определить границы интервалов.

Начало первого интервала рекомендуется определять по формуле:

$x_{нач} = x\frac{h}{2_{min}}$;

конец последнего должен удовлетворять условию $x_{кон} - h \leq xкон_{max}$; промежуточные интервалы получают, прибавляя к концу предыдущего интервала шаг.

4. Определить частоты.

Просматривая результаты наблюдений, определяют, сколько значений случайной величины попало в каждый конкретный интервал. При этом в интервал включают значения, большие или равные нижней границе интервала, и меньшие – верхней границы.

5. Составить таблицу интервального статистического ряда.

В первую строку таблицы вписывают частичные промежутки $[x_0, x_1)$, $[x_1, x_2)$, …, $[x_{m-1}, x_m)$, во вторую – количество наблюдений n_i (где $i = \overline{1, m}$) попавших в каждый интервал; то есть частоты соответствующих интервалов.

$[x_0 - x_1)$	$[x_1 - x_2)$	$[x_2 - x_3)$	…	$[x_{m-1} - x_m)$
n_1	n_2	n_3	…	n_m

$$\sum_{i=1}^{m} n_i = 1.$$

Иногда интервальный статистический ряд, для простоты исследований условно заменяют дискретным. В этом случае серединное значение -го интервала принимают за вариант x_i, а соответствующую интервальную частоту n_i - за частоту этого варианта.

Графическое изображение статистических данных

Статистическое распределение изображается графически с помощью полигона и гистограммы.

Полигоном частот называют ломаную, отрезки которой соединяют точки с координатами (x_i, n_i); *полигоном частостей* – с координатами $\left(x_i, p_i^*\right)$, где $p_i^* = \frac{n_i}{n}$, $i = \overline{1, m}$.

Полигон служит для изображения дискретного статистического ряда.

Полигон частостей является аналогом многоугольника распределения дискретной случайной величины в теории вероятностей.

Гистограммой частот (частостей) называют ступенчатую фигуру, состоящую из прямоугольников, основания которых расположены на оси Ox и длины их равны длинам частичных интервалов (h), а высоты равны отношению $\frac{n_i}{h}$ – для гистограммы частот; $\frac{n_i}{n \cdot h}$ – для гистограммы частостей.

Гистограмма является графическим изображением интервального ряда.

Площадь гистограммы частот равна n, а гистограммы частостей равна 1.

Можно построить полигон для интервального ряда, если его преобразовать в дискретный ряд. В этом случае интервалы заменяют их серединными значениями и ставят в соответствие интервальные частоты (частости). Полигон получим, соединив отрезками середины верхних оснований прямоугольников гистограммы.

Пример. Дана выборка значений случайной величины X объема 20:

12, 14, 19, 15, 14, 18, 13, 16, 17, 12, 18, 17, 15, 13, 17, 14, 14, 13, 14, 16

Требуется:

– построить дискретный вариационный ряд;

– найти размах варьирования R, моду M_0, медиану M_e;

– построить полигон частостей.

1) Ранжируем выборку:

12, 12, 13, 13, 13, 14, 14, 14, 14, 14, 15, 15, 16, 16, 17, 17, 17, 18, 18, 19.

2) Находим частоты вариантов и строим дискретный вариационный ряд:

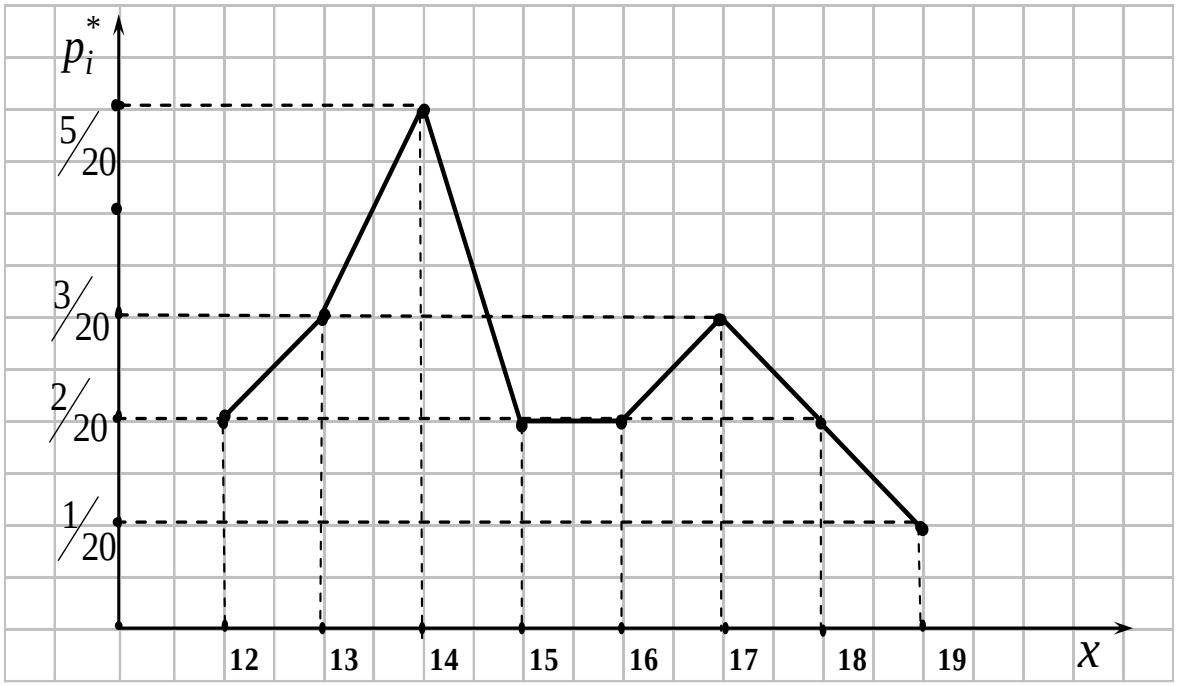
Значения вариантов x_i	12	13	14	15	16	17	18	19
Частоты n_i	2	3	5	2	2	3	2	1
Частости $p_i^* = \frac{n_i}{n}$	$\frac{2}{20}$	$\frac{3}{20}$	$\frac{5}{20}$	$\frac{2}{20}$	$\frac{2}{20}$	$\frac{3}{20}$	$\frac{2}{20}$	$\frac{1}{20}$

$\sum_{i=1}^{8} n_i = 20, \sum_{i=1}^{8} p_i = 1.$

3) По результатам таблицы 3 находим:

$R = 19 - 12 = 7,\ M_0 = 14,\ \ M_e = \frac{x_{10}+x_{11}}{2} = \frac{14+15}{2} = 14{,}5$

4) Строим полигон частостей.



Пример. Результаты измерений отклонений от нормы диаметров 50 подшипников дали численные значения (в мкм).

-1,760	-0,291	-0,110	-0,450	0,512
-0,158	1,701	0,634	0,720	0,490
1,531	-0,433	1,409	1,740	-0,266
-0,058	0,248	-0,095	-1,488	-0,361
0,415	-1,382	0,129	-0,361	-0,087
-0,329	0,086	0,130	-0,244	-0,882
0,318	-1,087	0,899	1,028	-1,304
0,349	-0,293	0,105	-0,056	0,757
-0,059	-0,539	-0,078	0,229	0,194
0,123	0,318	0,367	-0,992	0,529

Для данной выборки:

– построить интервальный вариационный ряд;

– построить гистограмму и полигон частостей.

1. Строим интервальный ряд.

По данным таблицы определяем: $x_{min};\ x_{max}$

Для определения длины интервала h используем формулу Стерджеса:

$$h = \frac{x_{max} - x_{min}}{1 + 3{,}322\, lg\ 50}.$$

Число интервалов $m \approx 1 + 3{,}322\, lg\ 5\,0.$

$$h = \frac{x_{max} - x_{min}}{1 + 3{,}322\, lg\, 50} = \frac{1{,}74 - (-1{,}76)}{1 + 3{,}322\, lg\, 50} \approx \frac{3{,}5}{1 + 3{,}322\, lg\, 50} \approx \frac{3{,}5}{6{,}644} \approx 0{,}526$$

Примем h =0,6 , $m = 7$.

За начало первого интервала примем величину

$x_{нач} = x\frac{h}{2}_{min}$.

Конец последнего интервала должен удовлетворять условию:

$x_{кон} - h \leq xкон_{max}$.

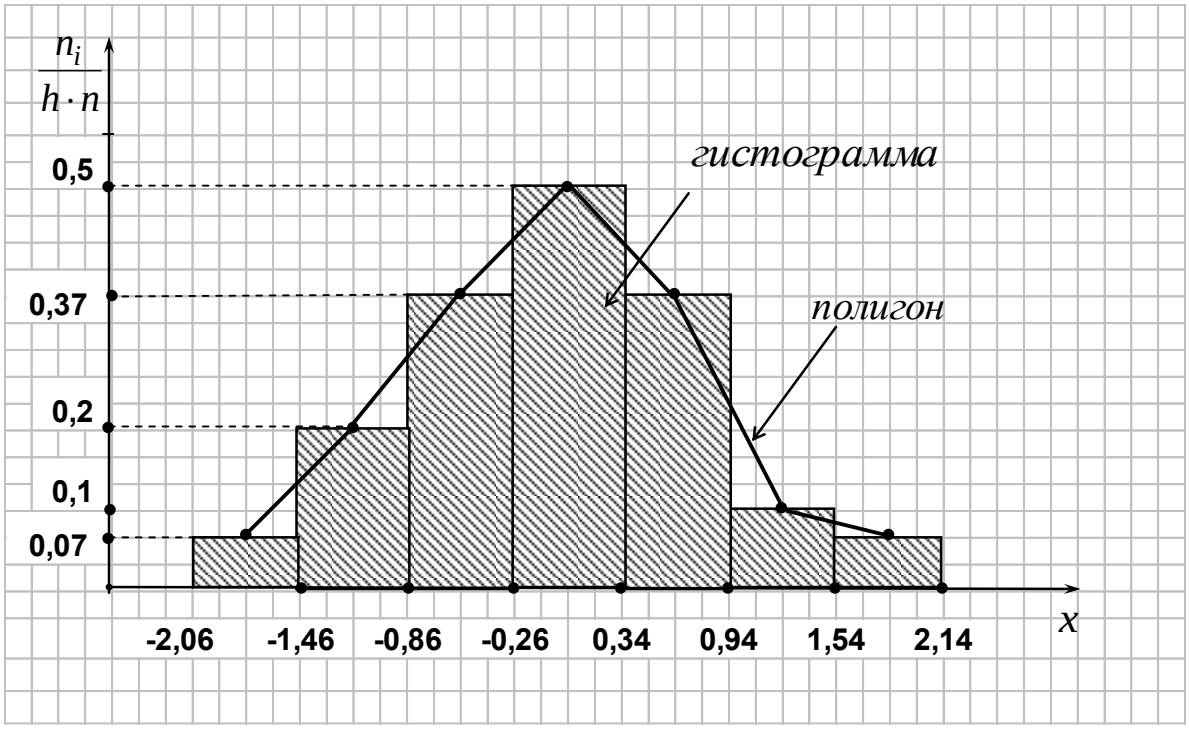
Действительно, $2{,}14 - 0{,}6 \leq 1{,}74 < 2{,}14$; $1{,}54 \leq 1{,}74 < 2{,}14$.

Строим интервальный ряд

Интервалы	Частоты n_i	Частости p_i
$[-2{,}06; -1{,}46)$	2	$\frac{2}{50}$
$[-1{,}46; -0{,}86)$	6	$\frac{6}{50}$
$[-0{,}86; -0{,}26)$	11	$\frac{11}{50}$
$[-0{,}26; 0{,}34)$	15	$\frac{15}{50}$
$[0{,}34; 0{,}94)$	11	$\frac{11}{50}$
$[0{,}94; 1{,}54)$	3	$\frac{3}{50}$
$[1{,}54; 2{,}14)$	2	$\frac{2}{50}$

$\sum_{i=1}^{7} n_i = 50, \sum_{i=1}^{7} p_i = 1$

Строим гистограмму частостей.



Вершинами полигона являются середины верхних оснований прямоугольников гистограммы.

Убедимся, что площадь гистограммы равна 1.

$$S = h \cdot \left(\frac{n_1 + n_2 + \cdots + n_m}{n \cdot h}\right)$$

$$S = 0{,}6(0{,}07 + 0{,}2 + 0{,}37 + 0{,}5 + 0{,}37 + 0{,}1 + 0{,}07) = 0{,}6 \cdot 1{,}68 = 1{,}008 \approx 1$$

Выборочное среднее. Выборочная дисперсия. Выборочное среднее квадратическое отклонение

В теории вероятностей определяются числовые характеристики для случайных величин, с помощью которых можно сравнивать однотипные случайные величины. Аналогично можно определить ряд числовых характеристик и для выборки. Поскольку эти характеристики вычисляются по статистическим данным (по данным, полученным в результате наблюдений), их называют *статистическими характеристиками.*

Пусть дано статистическое распределение выборки объема n:

x_i	x_1	x_2	x_3	x_4	..	x_m
n_i	n_1	n_2	n_3	n_4	..	n_m

где m – число вариантов.

Выборочным средним $\overline{x_в}$ называется среднее арифметическое всех значений выборки:

$\overline{x_в} = \frac{1}{n}\sum_{i=1}^{m} x_i n_i.$

В случае интервального статистического ряда в качестве x_i берут середины интервалов, а n_i – соответствующие им частоты.

Выборочной дисперсией $D_в$ называется среднее арифметическое квадратов отклонений значений выборки от выборочного среднего $\overline{x_в}$:

$$D_в = \frac{1}{n}\sum_{i=1}^{m} (x_i - \overline{x_в})^2 \cdot n_i$$

Выборочное среднее квадратическое выборки определяется формулой:

$$\sigma_в = \sqrt{D_в}.$$

Особенность $\sigma_в$ состоит в том, что оно измеряется в тех же единицах, что и данные выборки.

Если объем выборки мал ($n \leq 30$), то пользуются *исправленной выборочной дисперсией*: $S^2 = \frac{n}{n-1} D_в$.

Величина $S = \sqrt{S^2}$ называется *исправленным средним квадратическим отклонением.*

Выборочные начальные и центральные моменты

Асимметрия. Эксцесс

Приведем краткий обзор характеристик, которые наряду с уже рассмотренными применяются для анализа статистических рядов и являются аналогами соответствующих числовых характеристик случайной величины.

Среднее выборочное и выборочная дисперсия являются частным случаем более общего понятия – *момента* статистического ряда.

Начальным выборочным моментом порядка l называется среднее арифметическое l - х степеней всех значений выборки:

$\nu_l^* = \frac{1}{n}\sum_{i=1}^{m} x_i^l \cdot n_i$

Из определения следует, что начальный выборочный момент первого порядка: $\nu_1^* = \frac{1}{n}\sum_{i=1}^{m} x_i \cdot n_i = \overline{x_в}$.

Центральным выборочным моментом порядка l называется среднее арифметическое l - х степеней отклонений наблюдаемых значений выборки от выборочного среднего $\overline{x_в}$:

$\mu_l^* = \frac{1}{n}\sum_{i=1}^{m} (x_i - \overline{x_в})^l \cdot n_i.$

Из определения следует, что *центральный выборочный момент второго порядка*:

$\mu_2^* = \frac{1}{n}\sum_{i=1}^{m}(x_i - \overline{x_в})^2 \cdot n_i = D_в = \sigma_в^2$.

Выборочным коэффициентом асимметрии называется число $\boldsymbol{A_s^*}$, определяемое формулой: $A_s^* = \frac{\mu_3^*}{\sigma_в^3}$.

Выборочный коэффициент асимметрии служит для характеристики асимметрии полигона вариационного ряда. Если полигон асимметричен, то одна из ветвей его, начиная с вершины, имеет более пологий «спуск», чем другая.

Если $A_s^* < 0$, то более пологий «спуск» полигона наблюдается слева, если $A_s^* > 0$ – справа. В первом случае асимметрию называют *левосторонней*, а во втором – *правосторонней*.

Выборочным коэффициентом эксцесса или *коэффициентом крутости* называется число E_k^*, определяемое формулой: $E_k^* = \frac{\mu_4^*}{\sigma_в^4} - 3$.

Выборочный коэффициент эксцесса служит для сравнения на «крутость» выборочного распределения с нормальным распределением.

Коэффициент эксцесса для случайной величины, распределенной по нормальному закону, равен нулю.

Поэтому за стандартное значение выборочного коэффициента эксцесса принимают $E_k^* = 0$.

Если $E_k^* < 0$, то полигон имеет более пологую вершину по сравнению с нормальной кривой; если $E_k^* > 0$, то полигон более крутой по сравнению с нормальной кривой.

Контрольные вопросы

1. Что такое генеральная совокупность?
2. Что такое выборка и как она используется в статистике?
3. Какие существуют типы выборок? Приведите примеры.
4. Что такое описательная статистика?
5. Каковы основные методы анализа описательной статистики?
6. Каковы определения среднего, медианы и моды?
7. Что такое дисперсия и стандартное отклонение?
8. Каковы методы визуализации данных? Приведите примеры.

2.3 Тема 3. Статистическое оценивание параметров. Точечные и интервальные оценки

Вопросы для изучения

1. Оценка неизвестных параметров распределения. Точечные оценки.
2. Свойства точечных оценок: несмещенность, состоятельность, эффективность.

3. Методы нахождения точечных оценок

4. Точечные оценки математического ожидания и дисперсии; исправленная дисперсия.

5. Интервальные оценки. Доверительный интервал, доверительная вероятность.

6. Доверительные интервалы для параметров нормального распределения.

7. Компьютерные инструменты для статистического оценивания

Методические указания

При изучении темы «Точечные и интервальные оценки неизвестных параметров» потребуются такие понятия теории вероятностей как случайная величина, закон распределения, математическое ожидание, дисперсия. Изучите различные виды оценок (точечная оценка, интервальная оценка) и методы их получения (метод моментов, метод максимального правдоподобия), достоинства и недостатки этих методов. Освойте понятие доверительного интервала, доверительной вероятности и их связь с уровнем значимости. Практикуйтесь в решении задач, используя статистические пакеты, чтобы лучше усвоить материал.

Рекомендуемые источники: [1, гл. 9]; [2, т. 1, гл. 7]; [3, гл. 1].

Программное обеспечение

– Python (библиотеки scipy.stats и statsmodels: функция t.interval позволяет рассчитать доверительный интервал для нормального распределения);

– R (библиотеки stats и tibble содержат функции для расчета доверительных интервалов, такие как confint и CI.);

– Excel (функции ДОВЕРИТ.НОРМ и ДОВЕРИТ.СТЬЮДЕНТ);

– SPSS (встроенные функции для построения доверительных интервалов).

Основные теоретические сведения и решение типовых задач

Одной из центральных задач математической статистики является задача оценивания теоретического распределения случайной величины на основе выборочных данных.

При этом часто предполагается, что вид закона распределения генеральной совокупности известен, но неизвестны параметры этого распределения, такие как математическое ожидание, дисперсия. Требуется найти приближенные значения этих параметров, то есть получить статистические оценки указанных параметров.

Статистической оценкой $\overline{\theta}$ параметра θ теоретического распределения называют его приближенное значение, зависящее от данных выборки.

Рассматривая выборочные значения x_1 , x_2, …, x_n как реализации случайных величин X_1 , X_2, …, X_n, получивших конкретные значения в результате опытов, можно представить оценку $\overline{\theta}$ как функцию этих случайных величин: $\overline{\theta} = \phi(X_1, X_2, \ldots, X_n)$. Это означает, что оценка тоже является случайной величиной.

Если для оценки θ взять несколько (k) выборок, то получим столько же случайных оценок $\overline{\theta_1}$, $\overline{\theta_2}$, …, $\overline{\theta_k}$.

Если число наблюдений невелико, то замена неизвестного параметра θ оценкой $\overline{\theta}$ приводит к ошибке, которая тем больше, чем меньше число опытов.

Точечные оценки

Статистические оценки могут быть *точечными* и *интервальными*.

Точечные оценки представляют собой число или точку на числовой оси. Чтобы оценка $\overline{\theta}$ была близка к значению параметра θ, она должна обладать свойствами состоятельности, несмещенности и эффективности.

Оценка $\overline{\theta}$параметра θ называется *состоятельной,* если она сходится по вероятности к оцениваемому параметру, то есть для любого $\varepsilon > 0$:

$$\lim_{n\to\infty} P\left(\left|\overline{\theta} - \theta\right| < \varepsilon\right) = 1.$$

Поясним смысл этого равенства.

Пусть ε – очень малое положительное число. Тогда данное равенство означает, что чем больше объем выборки n, тем ближе оценка $\overline{\theta}$ приближается к оцениваемому параметру θ.

Свойство состоятельности нужно проверять в первую очередь. Оно *обязательно* для любого правила оценивания. Несостоятельные оценки не используются.

Оценка $\overline{\theta}$ параметра θ называется *несмещенной,* если $M\left(\overline{\theta}\right) = \theta$, то есть математическое ожидание оценки равно оцениваемому параметру. Если$M\left(\overline{\theta}\right) \neq \theta$, то оценка θ называется *смещенной.*

Это свойство оценки желательно, но не обязательно. Часто полученная оценка бывает смещенной, но ее можно поправить так, чтобы она стала несмещенной.

Иногда, оценка бывает *асимптотически несмещенной*, то есть $M\left(\overline{\theta}\right) \to \theta$.

Требования несмещенности особенно важно при малом числе опытов.

Несмещенная оценка $\overline{\theta}$ параметра θ называется *эффективной,* если она среди всех несмещенных оценок, в определенном классе оценок данного параметра, обладает наименьшей дисперсией.

Можно показать, что:

– $\overline{x_{\text{в}}}$ является состоятельной, несмещенной и эффективной оценкой $M(X)$ в классе линейных оценок;

– $D_{\text{в}}$ является состоятельной, смещенной оценкой $D(X)$;

– $S^2 = \frac{n}{n-1} D_{\text{в}}$ является состоятельной, несмещенной оценкой $D(X)$ (при больших n разница между S^2 и $D_{\text{в}}$ мала; S^2 используется при малых выборках, обычно при $n \leq 30$);

– относительная частота $\frac{n_A}{n}$ появления события A в n независимых испытаниях является состоятельной, несмещенной и эффективной оценкой, в классе линейных оценок, неизвестной вероятности $p = P(A)$ (p – вероятность появления события A в каждом испытании);

– эмпирическая функция распределения выборки $F^*(x)$ является состоятельной, несмещенной оценкой функции распределения $F(x)$ случайной величины X.

Для нахождения оценок неизвестных параметров используют различные методы. Наиболее распространенными являются: метод моментов, метод максимального правдоподобия (ММП), метод наименьших квадратов (МНК).

Интервальные оценки

При выборке малого объема точечная оценка может существенно отличаться от оцениваемого параметра. В этом случае целесообразно использовать интервальные оценки.

Интервальной называют оценку, которая определяется двумя числами – концами интервала.

Пусть найденная по данным выборки величина $\overline{\theta}$ служит оценкой неизвестного параметра θ. Оценка $\overline{\theta}$ определяет θ тем точнее, чем меньше $\left|\theta - \overline{\theta}\right|$, то есть чем меньше δ в неравенстве $\left|\theta - \overline{\theta}\right| < \delta$ ($\delta > 0$).

Поскольку $\overline{\theta}$ – случайная величина, то и разность $\left|\theta - \overline{\theta}\right|$ – случайная величина. Поэтому неравенство $\left|\theta - \overline{\theta}\right| < \delta$, при заданном δ может выполняться только с некоторой вероятностью.

Доверительной вероятностью (надежностью) оценки $\overline{\theta}$ параметра θ называется вероятность γ, с которой выполняется неравенство $\left|\theta - \overline{\theta}\right| < \delta$.

Обычно задается надежность γ и определяется δ. Чаще всего надежность задается значениями от 0,95 и выше, в зависимости от конкретно решаемой задачи.

Неравенство $\left|\theta - \overline{\theta}\right| < \delta$ можно записать $\overline{\theta} - \delta < \theta < \overline{\theta} + \delta$.

Доверительным интервалом называется интервал $\left(\overline{\theta}-\delta;\ \overline{\theta}+\delta\right)$, который покрывает неизвестный параметр с заданной надежностью γ.

Доверительный интервал для оценки математического ожидания нормального распределения при известной дисперсии

Пусть случайная величина Xимеет нормальное распределение $N(a;\ \sigma)$. Известно значение σ и задана доверительная вероятность (надежность) γ.

Доверительный интервал для параметра a по выборочному среднему $\overline{x_в}$ имеет вид:

$$\left(\overline{X_в}-t\cdot\frac{\sigma}{\sqrt{n}};\ \overline{X_в}+t\cdot\frac{\sigma}{\sqrt{n}}\ \right)$$

Пример. Случайная величина имеет нормальное распределение с известным средним квадратическим отклонением $\sigma=3$. Найти доверительный интервал оценки неизвестного математического ожидания по выборочной средней $\overline{x_в}$, если объем выборки $n=36$, а надежность оценки $\gamma=0{,}95$.

1. Находим t: $2\Phi(t)=0{,}95\ \ \Phi(t)=0{,}475$.

По таблице значений функции Лапласа $t=1{,}96$.

2. Определяем $\delta=t\cdot\frac{\sigma}{\sqrt{n}}=1{,}96\cdot\frac{3}{\sqrt{36}}=0{,}98$.

Доверительный интервал запишется в виде: $(\overline{x_в}-0{,}98;\ \overline{x_в}+0{,}98\ )$. ●

Доверительный интервал для оценки математического ожидания при неизвестной дисперсии

Пусть случайная величина Химеет нормальное распределение: $N(a;\ \sigma)$, причем σ – неизвестно, γ – задана.

Доверительный интервал для оценки $a=M(X)$ имеет вид:

$$\left(\overline{X_в}-t_j\cdot\frac{S}{\sqrt{n}};\ \overline{X_в}+t_j\cdot\frac{S}{\sqrt{n}}\ \right),$$

где $\overline{X_в}$ – выборочное среднее; S – исправленное среднее квадратическое отклонение; t_j – находим по таблице квантилей распределения Стьюдента (приложение 4) в зависимости от числа степеней свободы и доверительной вероятности γ.

Пример. Произведено пять независимых наблюдений над случайной величиной $X\sim N(a;\sigma)$. Результаты наблюдений таковы:

$x_1=35,\ x_2=20,\ x_3=15,\ x_4=-12,\ x_5=42.$

Построить для неизвестного $M(x)=a$ доверительный интервал, если $\gamma=0{,}95$.

1. Находим $\overline{x_в}$: $\overline{x_в}=\frac{1}{5}(-35+20+15-12+42)=\frac{1}{5}\cdot 30=6$

$\underline{\overline{x_в}=6}$

2. Находим S^2:

$S^2=\frac{1}{4}((-35-6)^2+(20-6)^2+(15-6)^2+(-12-6)^2+(42-6)^2)=$

$= \frac{1}{4}((-41)^2 + 16^2 + 9^2 + (-18)^2 + 36^2) = \frac{1}{4}(1681 + 256 + 81 + 324 + 1296) = \frac{1}{4}3638 = 909{,}5$

$\underline{S = \sqrt{909{,}5} \approx 30{,}2}$

3. По таблице квантилей распределения Стьюдента (Приложение 4) для $\gamma = 0{,}95$ и $n - 1 = 4$ находим t_j:

$\underline{t_j = 2{,}78}$

Доверительный интервал:

$\left(6 - 2{,}78\frac{30{,}2}{2{,}24};\ 6 + 2{,}78\frac{30{,}2}{2{,}24}\right)$ или $(31{,}5;\ 43{,}5)$.

Доверительный интервал для оценки среднего квадратического отклонения нормального распределения

Если $M(X) = a$ неизвестно, то доверительный интервал для оценки $\sigma(X)$имеет вид:

$$\left(\frac{\sqrt{n-1}\cdot S}{\chi_2};\ \frac{\sqrt{n-1}\cdot S}{\chi_1}\right)$$

где n – объем выборки; S – исправленное среднее квадратическое отклонение:

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X_в}\right)^2$,

$\chi_1^2 = \chi^2_{\frac{1+\gamma}{2};n-1}$, $\chi_2^2 = \chi^2_{\frac{1-\gamma}{2};n-1}$ – квантилиχ^2 – распределения, определяемые по таблице $\chi^2_{\alpha,k}$ при $k = n - 1$ и $\alpha = \frac{1+\gamma}{2}$, $\alpha = \frac{1-\gamma}{2}$.

Пример. Для оценки параметра $\sigma(X)$ нормально распределенной случайной величины была сделана выборка объема в 25 единиц и вычислено $S = 0{,}8$. Найти доверительный интервал, покрывающий σ с вероятностью $\gamma = 0{,}95$.

Имеем $n = 25$, $\gamma = 0{,}95$.

$$\chi_1^2 = \chi^2_{\frac{1+0{,}95}{2};25-1} = \chi^2(0{,}975;\ 24) = 12{,}4$$

$$\chi_2^2 = \chi^2_{\frac{1-0{,}95}{2};-1} = \chi^2(0{,}025;\ 24) = 39{,}4$$

Доверительный интервал имеет вид:

$\left(\frac{\sqrt{24}\cdot 0{,}8}{\sqrt{39{,}4}};\ \frac{\sqrt{24}\cdot 0{,}8}{\sqrt{12{,}4}}\right)$ или $(0{,}79;\ 1{,}4)$.

Контрольные вопросы

1. Что такое точечная оценка параметра статистической совокупности?
2. Каковы основные свойства хорошей точечной оценки?
3. Какие существуют методы нахождения точечных оценок?
4. Какие статистические величины являются наилучшими оценками неизвестных математического ожидания и дисперсии генеральной совокупности?
5. Что такое интервальная оценка параметра?

6. Что такое уровень доверия и как он влияет на ширину доверительного интервала?

7. Каковы основные методы построения доверительных интервалов?

8. Что такое стандартная ошибка и как она используется в оценивании параметров?

9. Как интерпретировать доверительный интервал в контексте статистических данных?

10. Каковы основные ошибки, которые могут возникнуть при оценивании параметров?

2.4 Тема 4. Статистические гипотезы

Вопросы для изучения

1. Понятие статистической гипотезы. Основная и конкурирующая гипотезы.
2. Ошибки первого и второго рода.
3. Уровень значимости, досягаемый уровень значимости (p-value).
4. Статистический критерий и критическая область.
5. Классический алгоритм проверки статистической гипотезы
6. Параметрическая проверка гипотез
7. Непараметрическая проверка гипотез
8. Компьютерная реализация методов проверки статистических гипотез.

Методические указания

Для изучения темы «Статистические гипотезы» потребуется понимание основ теории вероятностей и базовых понятий математической статистики. Следует изучить типы гипотез (нулевая и альтернативная), критерии проверки гипотез (например, t-критерий Стьюдента, критерий хи-квадрат Пирсона) и методы их применения. Важно освоить практическое использование статистических пакетов для анализа данных, таких как R или Python. Регулярная практика решения задач поможет закрепить теоретический материал.

Рекомендуемые источники: [1, гл.10]; [2, т. 1, гл. 8]; [3, гл. 2].

Программное обеспечение

– Python (библиотеки scipy.stats, statsmodels);

– R (функции t.test(), chisq.test() и другие);

– Excel (инструмент «Анализ данных» имеет базовые функции для некоторых видов тестирования гипотез, таких как t-тесты.);

– SPSS (имеет функции, позволяющие тестировать различные гипотезы, включая t-тесты, F-тесты, хи-квадрат и многие другие).

Основные теоретические сведения и решение типовых задач

Статистической гипотезой называется всякое высказывание о генеральной совокупности (случайной величине), проверяемое по выборке (т. е. по результатам наблюдений).

Примеры статистических гипотез:

– математическое ожидание случайной величины равно конкретному числовому значению;

– генеральная совокупность распределена по нормальному закону.

Гипотезы могут быть *параметрические* (гипотезы о параметрах распределения известного вида) и *непараметрические* (гипотезы о виде неизвестного распределения).

Процедура сопоставления гипотезы с выборочными данными называется *проверкой гипотезы*. Для проверки гипотез используют *аналитические* и *статистические* методы.

Классический метод проверки гипотез

В соответствии с поставленной задачей и на основании выборочных данных формулируется (выдвигается) гипотеза H_0, которая называется основной или *нулевой*. Одновременно с выдвинутой гипотезой H_0, рассматривается противоположная ей гипотеза H_1, которая называется конкурирующей или *альтернативной*.

Для проверки нулевой гипотезы вводят специально подобранную случайную величину K, распределение которой известно и называют ее *критерием.*

Поскольку гипотеза H_0 для генеральной совокупности принимается по выборочным данным, то она может быть ошибочной. При этом возможны ошибки двух родов.

Ошибка первого рода состоит в том, что отвергается гипотеза H_0, когда она на самом деле верна.

Ошибка второго рода состоит в том, что отвергается альтернативная гипотеза H_1, когда она на самом деле верна.

1) Для определения вероятности ошибки первого рода вводится параметр α: $\alpha = P_{H_0}(H_1)$ – вероятность того, что будет принята гипотеза H_1, при условии, что H_0 верна. Величину α называют *уровнем значимости*. Обычно α выбирают в пределах $0{,}001 - 0{,}1$.

2) Вероятность ошибки второго рода определяется параметром β: $\beta = P_{H_1}(H_0)$ – вероятность того, что будет принята гипотеза H_0, при условии, что H_1

верна. Величину $(1-\beta)$, т. е. недопустимость ошибки второго рода (отвергнуть неверную и принять верную гипотезу H_1) называют *мощностью критерия*.

Суть метода.

Множество всех значений критерия разбивают на два непересекающихся подмножества: одно из них содержит значения критерия, при которых нулевая гипотеза H_0 отвергается; другое – при которых она принимается.

Критической областью называется совокупность значений критерия, при которых нулевую гипотезу отвергают. Обозначим критическую область ω.

Областью принятия гипотезы (областью допустимых значений) называется совокупность значений критерия, при которых нулевую гипотезу принимают.

Если вычисленное по выборке значение критерия K попадает в критическую область ω, то гипотеза H_0 отвергается и принимается гипотеза H_1. В этом случае можно совершить ошибку первого рода, вероятность которой равна α. Иначе, вероятность того, что критерий K примет значение из критической области ω, должна быть равна заданному значению α, то есть$P(K \in \omega) = \alpha$.

Критическая область ω определяется неоднозначно. Возможны три случая расположения ω. Они определяются видом нулевой и альтернативной гипотез и законом распределения критерияK.

Правосторонняя критическая область (рисунок 1, а) состоит из интервала $\left(k_{\text{пр.}\alpha}^{\text{кр}};\ +\infty\right)$, где $k_{\text{пр.}\alpha}^{\text{кр}}$ определяется из условия $P\left(K > k_{\text{пр.}\alpha}^{\text{кр}}\right) = \alpha$ и называется правосторонней критической точкой, отвечающей уровню значимости α.

Левосторонняя критическая область (рисунок 1, б) состоит из интервала $\left(-\infty; k_{\text{л.}\alpha}^{\text{кр}}\right)$, где $k_{\text{л.}\alpha}^{\text{кр}}$ определяется из условия $P\left(K < k_{\text{л.}\alpha}^{\text{кр}}\right) = \alpha$ и называется левосторонней критической точкой, отвечающей уровню значимости α.

Двусторонняя критическая область (рисунок 1, в) состоит из следующих двух интервалов: $\left(-\infty; k_{\text{л.}\alpha/2}^{\text{кр}}\right)$ и $\left(k_{\text{пр.}\alpha/2}^{\text{кр}}; +\infty\right)$, где точки $k_{\text{л.}\alpha/2}^{\text{кр}}$ и $k_{\text{пр.}\alpha/2}^{\text{кр}}$ определяются из условий $P\left(K < k_{\text{л.}\alpha/2}^{\text{кр}}\right) = \frac{\alpha}{2}$ и $P\left(K > k_{пр.\alpha/2}^{кр}\right) = \frac{\alpha}{2}$ и называются двусторонними критическими точками.

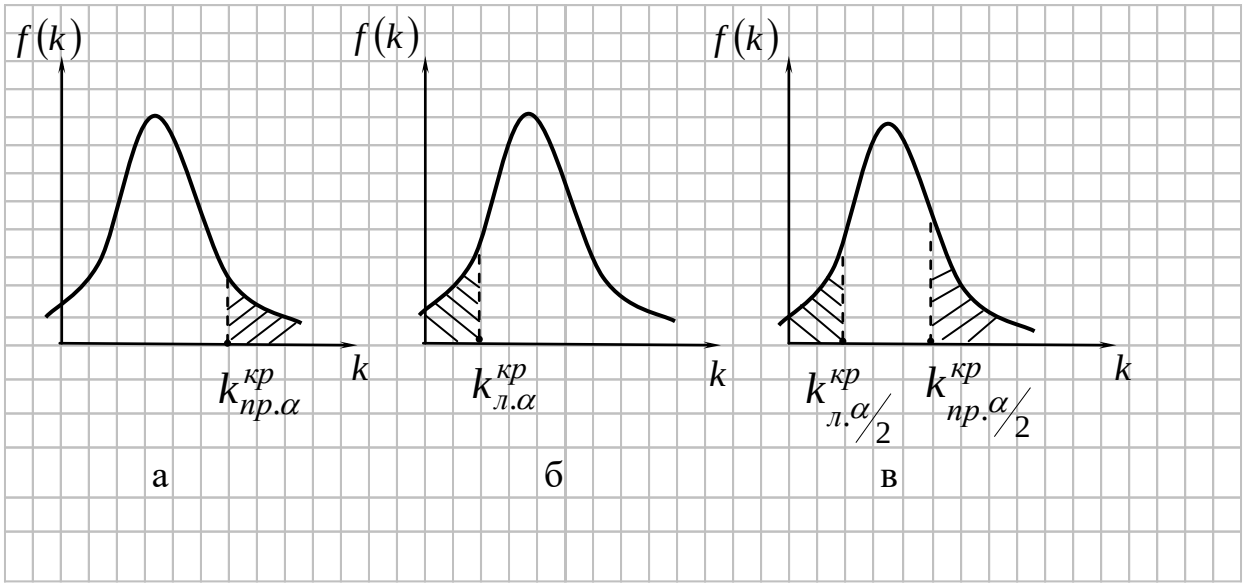


Рисунок 1

Алгоритм проверки нулевой гипотезы

1. Располагая выборкой, формулируют нулевую гипотезу H_0 и альтернативную гипотезу H_1.

2. Выбирают критерий проверки гипотезы H_0, зависящий от выборочных данных и условий рассматриваемой задачи. Наиболее часто используют случайные величины, имеющие следующие законы распределения: нормальный, Стьюдента, Фишера-Снедекора, хи-квадрат.

3. Задают уровень значимости выбранного критерия и определяют соответствующую ему критическую область. Для определения критической области достаточно найти критическую точку $t_{кр}$ – ее границу. Для каждого критерия имеются таблицы, по которым находят критическую точку.

4. Вычисляют значение критерия по результатам произведенных измерений и сравнивают с критической точкой.

5. Нулевую гипотезу *отвергают*, если вычисленное значение критерия попадает в критическую область, или считают *справедливой*, если оно окажется внутри области допустимых значений.

Проверка гипотез о законе распределения

Во многих случаях закон распределения изучаемой случайной величины X неизвестен, но есть основания предположить, что он имеет вполне определенный вид: нормальный, экспоненциальный или какой-либо другой.

Пусть выдвинута гипотеза H_0 о каком-либо законе распределения. Для проверки этой гипотезы H_0 требуется по выборке сделать заключение, согласуются ли результаты наблюдений с высказанным предположением.

Статистический критерий проверки гипотезы о предполагаемом законе неизвестного распределения называется *критерием согласия.* Он используется для проверки согласия предполагаемого вида распределения с опытными данными на основании выборки.

Существуют различные критерии согласия: Пирсона, Колмогорова, Фишера и другие. Наиболее часто применяется критерий Пирсона.

Пример. Проверка гипотезы о нормальном распределении генеральной совокупности по критерию Пирсона.

Пусть выборка из генеральной совокупностиXзадана в виде статистического интервального ряда:

$[x_1, x_2)$	$[x_2, x_3)$	...	$[x_m, x_{m+1})$
n_1	n_2	...	n_m

где n_i – интервальные частоты; $\sum_{i=1}^{m} n_i = n$ – объем выборки; m – число интервалов; h – длина интервала; x_i – середина интервала.

Требуется проверить гипотезу H_0 о том, что генеральная совокупность X распределена по нормальному закону, применяя критерий Пирсона.

Правило проверки

1. Вычисляем $\overline{x_в}$ и $\sigma_в$

2. Находим теоретические частоты n_i': $n_i' = P_i \cdot n$, где $P_i = F(x_{i+1}) - F(x_i)$ – вероятность попадания рассматриваемой случайной величины в интервал $[x_i, x_{i+1})$;

$F(x)$ – функция распределения случайной величины, гипотеза о котором проверяется

3. Сравниваем эмпирические (n_i) и теоретические (n_i') частоты с помощью критерия Пирсона.

$$\chi^2_{набл} = \sum_{i=1}^{m} \frac{(n_i - n_i')^2}{n_i'}$$

Находим число степеней свободы k: $k = m - r - 1$, где m – число интервалов; r – число параметров предполагаемого распределения

Для нормального распределения $k = m - 3$, так как $r = 2$ (нормальный закон распределения характеризуется двумя параметрами a и σ).

4. В таблице критических точек (*квантилей*) распределения Пирсона по заданному уровню значимости α и числу степеней свободы находим $\chi^2_{кр}(\alpha; k)$ правосторонней критической области.

5. Если $\chi^2_{\text{набл}} < \chi^2_{\text{кр}}$ – нет оснований отвергнуть гипотезу H_0о нормальном распределении генеральной совокупности.

Если $\chi^2_{набл} > \chi^2_{кр}$ – гипотезу отвергаем.

Замечание.

1) Объем выборки должен быть достаточно велик $(n \geq 50)$.

2) Малочисленные частоты $(n_i < 5)$ следует объединить. В этом случае и соответствующие им теоретические частоты также надо сложить. Если производилось объединение частот, то при определении числа степеней свободы по формуле $k = m - 3$ следует в качестве m принять число интервалов, оставшихся после объединения частот.

С помощью критерия Пирсона можно проверить гипотезу о любом виде распределения, при этом алгоритм его применения не изменяется.

Пример. Пусть из генеральной совокупности X задана выборка объемом 50. Требуется проверить гипотезу H_0 о нормальном распределении генеральной совокупности по данной выборке.

Интервалы	$[-2{,}06; -1{,}46)$	$[-1{,}46; -0{,}86)$	$[-0{,}86; -0{,}26)$	$[-0{,}26; 0{,}34)$
Частоты n_i	2	6	11	15

Интервалы	$[0{,}34; 0{,}94)$	$[0{,}94; 1{,}54)$	$[1{,}54; 2{,}14)$	
Частоты n_i	11	3	2	$\sum_{i=1}^{7} n_i = 50.$

Проверим гипотезу H_0 по критерию Пирсона.

1) $\overline{x_в} = -0{,}032$, $\sigma_в = 0{,}8195$.

2) Найдем теоретические частоты n_i'

Интервальный ряд содержит интервалы с частотами меньшими 5. Следовательно, два первых и два последних интервала объединяем, при этом соответствующие частоты суммируем.

Составим расчетную таблицу.

i	Границы интервала		n_i	Границы интервала		$\Phi(z_i)$	$\Phi(z_{i+1})$	P_i	n_i'
	x_i	x_{i+1}		z_i	z_{i+1}				
1	-2,06	-0,86	8	$-\infty$	-1,01	-0,5	-0,3438	0,1562	7,81
2	-0,86	-0,26	11	-1,01	-0,28	-0,3438	-0,1103	0,2335	11,675
3	-0,26	0,34	15	-0,28	0,45	-0,1103	0,1736	0,2839	14,195
4	0,34	0,94	11	0,45	1,19	0,1736	0,3830	0,2094	10,47
5	0,94	2,14	5	1,19	$+\infty$	0,3830	0,5	0,1170	5,85
								1	**50**

3) Сравним эмпирические (n_i) и теоретические (n_i') частоты. Для этого составляем расчетную таблицу.

i	n_i	n_i'	$n_i - n_i'$	$(n_i - n_i')^2$	$\frac{(n_i - n_i')^2}{n_i'}$	n_i^2	$\frac{n_i^2}{n_i'}$
1	8	7,810	0,190	0,0361	0,0046	64	8,1946
2	11	11,675	-0,675	0,4556	0,0390	121	10,3640
3	15	14,195	0,805	0,6480	0,0457	225	15,8507
4	11	10,470	0,530	0,2809	0,0268	121	11,5568
5	5	5,850	-0,850	0,7225	0,1235	25	4,2735
					0,2396		**50,2396**

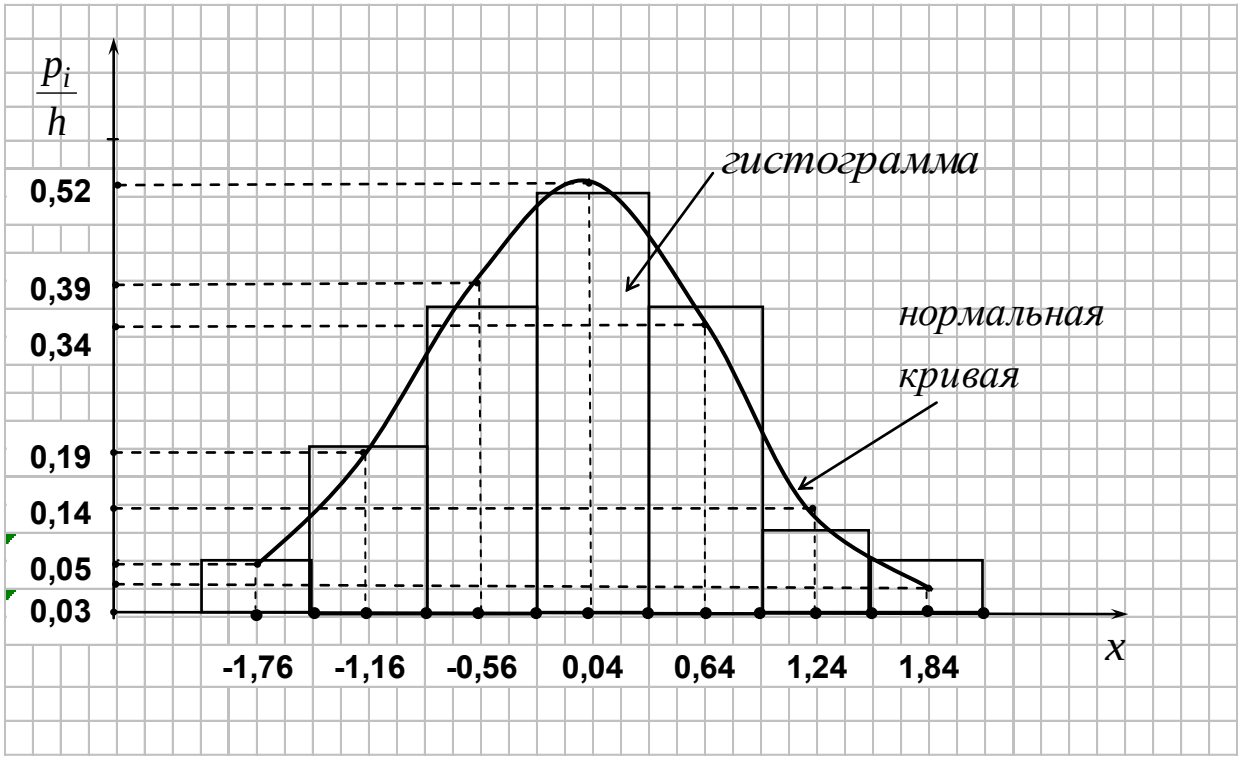
$\chi^2_{набл} = 0{,}2396$.

4) Зададим $\boldsymbol{\alpha = 0{,}05}$. Вычислим число степеней свободы $k = 5 - 3 = 2$ и найдем $\chi^2_{\text{кр}}(0{,}05; 2) = 6{,}0$ (таблица критических точек распределения «хи-квадрат»). Получим $\chi^2_{набл} < \chi^2_{кр}$.

Следовательно, нет оснований отвергать гипотезу H_0 о нормальном распределении генеральной совокупности X.

Другими словами различие между эмпирическими (n_i) и теоретическими (n_i') частотами незначительное (случайное), которое можно объяснить малым объемом выборки.

Проведем визуальную проверку согласования опытных данных с нормальным законом распределения. Для этого построим нормальную кривую и гистограмму относительных частот на одном чертеже.



Так как гипотеза о нормальном распределении не отвергается, то нормальная кривая хорошо сглаживает гистограмму.

Контрольные вопросы

1. Что такое статистическая гипотеза?
2. Каковы основные виды статистических гипотез?
3. В чем разница между нулевой и альтернативной гипотезами?
4. Что такое уровень значимости и как он определяется?
5. Каковы основные шаги в тестировании статистических гипотез?
6. Что такое ошибка первого рода и ошибка второго рода?
7. Как интерпретировать p-значение в контексте тестирования гипотез?
8. Какие существуют методы тестирования гипотез (например, t-тест, z-тест)?
9. Каковы критерии для выбора подходящего теста для проверки гипотезы?
10. Что такое мощность теста и как она влияет на результаты тестирования гипотез?

2.5 Тема 5. Дисперсионный анализ (ANOVA)

Вопросы для изучения

1. Понятие дисперсионного анализа (ANOVA).
2. Однофакторный дисперсионный анализ.
3. Предпосылки для применения однофакторного дисперсионного анализа.
4. Многофакторный дисперсионный анализ
5. Основные этапы проведения дисперсионного анализа.
6. Интерпретация результатов дисперсионного анализа.
7. Компьютерные инструменты для дисперсионного анализа.

Методические указания

Для изучения темы «Дисперсионный анализ» (ANOVA) важно хорошее знание таких понятий статистики, как средние значения, дисперсии, проверка гипотез. Требуется изучить основные понятия однофакторного и многофакторного ANOVA, а также пост-хок тестов. Практика решения задач с использованием статистического программного обеспечения, например, SPSS, R или Python, поможет закрепить теорию. Особое внимание уделите интерпретации результатов и пониманию ограничений метода.

Рекомендуемые источники: [1, гл. 11].

Программное обеспечение

– Python (библиотеки statsmodels и scikit-learn);

– R (пакеты aov и car);

– Excel (инструмент «Анализ данных» позволяет проводить только простой однофакторный анализ данных);

– SPSS (набор функций для однофакторного и многофакторного дисперсионного анализа, а также пост-хок тесты).

Основные теоретические сведения и решение типовых задач

Дисперсионный анализ – метод, направленный на поиск зависимостей в экспериментальных данных путём исследования значимости различий в средних значениях. Дисперсионный анализ позволяет сравнивать средние значения двух и более групп.

Основную задачу дисперсионного анализа можно сформулировать следующим образом: оказывает ли значимое влияние на значение некоторой количественной переменной интересующий нас признак, измеренный на номинальном или порядковом уровне?

В терминах метода дисперсионного анализа та переменная, которая, как мы считаем, должна оказывать влияние на конечный результат, называется фактором. Например, если мы хотим объяснить различия в средних доходах респондентов тем, что респонденты проживают в различных населенных пунктах, то переменная «место проживания респондента» – будет выступать фактором. Конкретное значение фактора (например, определенный населенный пункт) называют уровнем фактора. Значение измеряемого признака (в нашем примере – величину среднего дохода) называют откликом.

Если исследуется зависимость отклика только от одного фактора, то такой дисперсионный анализ называется однофакторным, если исследуется зависимость от двух и более факторов, то такой дисперсионный анализ называется многофакторным.

Само название – дисперсионный анализ (analysis of variance – сокращенно ANOVA) происходит из того, что метод проверки статистической гипотезы о равенстве средних значений в нескольких непересекающихся группах, основан на сопоставлении двух оценок дисперсии, анализируемой количественной переменной.

В ***однофакторной модели дисперсионного анализа*** исходят из следующей модели порождения данных:

$$x_{ij} = \mu_j + \varepsilon_{ij} = \mu + \alpha_j + \varepsilon_{ij}, \quad i = \overline{1, n_j}, \quad j = \overline{1, k},$$

где x_{ij} – i-ое наблюдаемое значение отклика в j-ой группе (для j-го уровня фактора); μ – среднее значение отклика по всем уровням фактора (среднее по всей совокупности); μ_j – среднее значение отклика для j-го уровня фактора; $\alpha_j = \mu_j - \mu$ – дифференциальный эффект среднего, соответствующий j-му уровню фактора; ε_{ij} – независимые случайные величины с математическим ожиданием равным нулю и одинаковой дисперсией σ^2.

Выражение $x_{ij} = \mu + \alpha_j + \varepsilon_{ij}$ можно представить в виде

$$x_{ij} = \mu + (\mu_j - \mu) + (x_{ij} - \mu_j),$$

или:

$$x_{ij} - \mu = (\mu_j - \mu) + (x_{ij} - \mu_j).$$

Данное соотношение говорит о том, что отклонение наблюдаемого значения отклика для j-ой группы складывается из суммы двух слагаемых: отклонения отклика от среднего значения j-ой группы: $(x_{ij} - \mu_j)$, и отклонения среднего значения j-ой группы от среднего значения всей совокупности: $(\mu_j - \mu)$. Что, по сути, означает, что дисперсия отклика может быть представлена в виде суммы двух дисперсий, одна из которых характеризует внутригрупповую изменчивость, а вторая межгрупповую.

Разложение общей дисперсии на составляющие для выборочных данных обычно записывается в виде равенства сумм квадратов соответствующих отклонений:

$$SS_T = SS_B + SS_R,$$

Где $SS_T = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{\mu})^2$ – общая, или полная, сумма квадратов отклонений; $SS_B = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\bar{\mu}_j - \bar{\mu})^2 = \sum_{j=1}^{k} n_j (\bar{\mu}_j - \bar{\mu})^2$ – сумма квадратов отклонений групповых средних от общего среднего, или межгрупповая (межуровневая факторная) сумма квадратов отклонений, также называемая суммой квадратов эффекта фактора или просто эффектом фактора; $SS_R = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{\mu}_j)^2$– сумма квадратов отклонений наблюдений от групповых средних, или внутригрупповая (остаточная) сумма квадратов отклонений, также называемая остаточным эффектом или эффектом ошибок; k – число уровней фактора, n_j – число наблюдений для j-го уровня фактора; $n = \sum_{j=1}^{k} n_j$ – общее число наблюдений.

В разложении дисперсии на составляющие заключена основная идея дисперсионного анализа: общая вариация переменной, порожденная влиянием фактора и измеренная суммой SS_T, складывается из двух компонент: SS_B и SS_R, характеризующих изменчивость этой переменной между уровнями фактора (SS_B) и внутри уровней фактора (SS_R).

В дисперсионном анализе анализируются не сами суммы квадратов отклонений, а так называемые средние квадраты, которые получаются делением сумм квадратов отклонений на соответствующее число степеней свободы. Число степеней свободы для суммы квадратов случайных величин определяется как общее число линейно независимых слагаемых.

Для полной суммы квадратов $SS_T = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{\mu})^2$ число степеней свободы $\nu_T = n - 1$, так как при ее расчете используются n наблюдений, связанных между собой одним уравнением для общего выборочного среднего всей совокупности.

Для суммы квадратов эффекта фактора $SS_\text{B} = \sum_{j=1}^{k} n_j(\bar{\mu}_j - \bar{\mu})^2$ число степеней свободы $\nu_B = k - 1$, так как при ее расчете используются k групповых средних, связанных между собой также одним уравнением для общего выборочного среднего всей совокупности.

Для суммы квадратов ошибок $SS_R = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{\mu}_j)^2$ число степеней свободы $\nu_R = n - k$, ибо при его расчете используются n наблюдений, связанных между собой k уравнениями для выборочных средних k групп.

Соответственно выражения для средних квадратов отклонений, которые являются несмещенными оценками соответствующих дисперсий, имеют вид:

$$MS_T = \frac{1}{n-1}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{\mu})^2,$$

$$MS_B = \frac{1}{k-1}\sum_{j=1}^{k} n_j(\bar{\mu}_j - \bar{\mu})^2,$$

$$SS_R = \frac{1}{n-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{\mu}_j)^2.$$

В случае нормального распределения остатков ε_{ij}, при условии истинности $H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$ (что равносильно: $\mu_1 = \mu_2 = \cdots = \mu_k$), статистика

$$F = \frac{MS_B}{MS_R} = \frac{n-k}{k-1}\frac{SS_B}{SS_R}$$

имеет распределение Фишера с $\nu_1 = k - 1$ и $\nu_2 = n - k$ числом степеней свободы.

Если наблюдаемое значение статистики $F_{\text{набл}} \geq F_{\text{кр}}$, где $F_{\text{кр}}$ – критическая точка распределения Фишера уровня α (или квантиль уровня $1 - \alpha$) с числом

степеней свободы $\nu_1 = k - 1$ и $\nu_2 = n - k$, то нулевая гипотеза отклоняется и считается, что средние для различных уровней фактора значимо различаются.

Условия применимости данной модели дисперсионного анализа:

1) нормальность распределения данных для каждого уровня фактора;

2) однородность (равенство) дисперсий для различных уровней фактора.

Рассмотренная модель дисперсионного анализа предполагает, что данные измерены в количественной шкале.

Для порядковых данных непараметрической альтернативой однофакторного дисперсионного анализа являются ранговый дисперсионный анализ Краскела–Уоллиса и медианный тест.

Если анализируется одновременное влияние двух и более различных факторов на результаты наблюдений, то используется ***многофакторный дисперсионный анализ***. Например, двухфакторная модель нам потребуется, если мы будем строить модель объяснения различий в средних доходах респондентов не только с учетом места проживания респондента, но и с учетом пола респондента.

Пусть мы исследуем влияние на величину X двух факторов A и B, имеющих, соответственно k и m уровней. В двухфакторной модели дисперсионного анализа обычно исходят из следующей модели порождения данных:

$$x_{ijl} = \mu_{ij} + \varepsilon_{ijl} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijl}, \quad l = \overline{1, n_{ij}}, \quad i = \overline{1, k}, \quad j = \overline{1, m},$$

где x_{ijl} – l-ое наблюдаемое значение отклика для i-го уровня фактора A и j-го уровня фактора B; μ – среднее значение отклика по всей совокупности (генеральное среднее); μ_{ij} – среднее значение отклика для i-го уровня фактора A и j-го уровня фактора B; $\alpha_i = \mu_{i*} - \mu$ – главный эффект i-го уровня фактора A (μ_{i*} – среднее значение отклика для i-го уровня фактора A); $\beta_j = \mu_{*j} - \mu$ – главный эффект j-го уровня фактора B (μ_{*j} – среднее значение отклика для j-го уровня фактора B); $\gamma_{ij} = \mu_{ij} - \mu_{i*} - \mu_{*j} + \mu$ – эффект взаимодействия i-го уровня фактора A и j-го уровня фактора B; ε_{ijl} – независимые случайные величины с математическим ожиданием равным нулю и одинаковой дисперсией σ^2.

Заметим, что эффекты α_i, β_j, γ_{ij} удовлетворяют условиям: $\sum_{i=1}^{k} \alpha_i = 0$, $\sum_{j=1}^{m} \beta_i = 0$, $\sum_{i=1}^{k} \gamma_{ij} = 0$, $\sum_{j=1}^{m} \gamma_{ij} = 0$.

Выражение $x_{ijl} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijl}$ можно представить в виде:

$$x_{ijl} - \mu = (\mu_{i*} - \mu) + (\mu_{*j} - \mu) + (\mu_{ij} - \mu_{i*} - \mu_{*j} + \mu) + (x_{ijl} - \mu_{ij}).$$

Данное соотношение говорит о том, что отклонение наблюдаемого значения отклика складывается из суммы четырех слагаемых: отклонения отклика от среднего значения для i, j-го набора уровней факторов A и B ($x_{ijl} - \mu_{ij}$), главных эффектов i-го уровня фактора A и j-го уровня фактора B и эффекта взаимодействия. Что, означает, с учетом указанных выше условий на эффекты, что

дисперсия отклика может быть представлена в виде суммы четырех дисперсий, одна из которых характеризует внутригрупповую изменчивость для i, j-го набора уровней факторов A и B, а остальные соответствующие эффекты.

Разложение общей дисперсии на составляющие для выборочных данных обычно записывается в виде равенства сумм квадратов соответствующих отклонений (которое, вообще говоря, справедливо только в случае выполнения условия пропорциональности $n_{ij} = n_{i*}n_{*j}/n$):

$$SS_T = SS_A + SS_B + SS_{AB} + SS_R,$$

Где $SS_T = \sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{l=1}^{n_{ij}}(x_{ijl} - \bar{\mu})^2$ – общая, или полная, сумма квадратов отклонений; $SS_A = \sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{l=1}^{n_{ij}}(\bar{\mu}_{i*} - \bar{\mu})^2 = \sum_{i=1}^{k} n_{i*}(\bar{\mu}_{i*} - \bar{\mu})^2$ – сумма квадратов отклонений средних по уровням фактора A от общей средней, или сумма квадратов главных эффектов A; $SS_B = \sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{l=1}^{n_{ij}}(\bar{\mu}_{*j} - \bar{\mu})^2 = \sum_{j=1}^{m} n_{*j}(\bar{\mu}_{*j} - \bar{\mu})^2$– сумма квадратов отклонений средних по уровням фактора B от общей средней, или сумма квадратов главных эффектов B; $SS_{AB} = \sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{l=1}^{n_{ij}}(\bar{\mu}_{ij} - \bar{\mu}_{*j} - \bar{\mu}_{i*} + \bar{\mu})^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij}(\bar{\mu}_{ij} - \bar{\mu}_{*j} - \bar{\mu}_{i*} + \bar{\mu})^2$ – сумма квадратов взаимодействия эффектов A и B; $SS_R = \sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{l=1}^{n_{ij}}(x_{ijl} - \bar{\mu}_{ij})^2$– остаточная сумма квадратов отклонений.

Число степеней свободы сумм квадратов SS_A и SS_{B} равно соответственно $\nu_A = k - 1$ и $\nu_B = m - 1$.

Число степеней свободы сумм квадратов взаимодействия эффектов SS_{AB} равно $\nu_{AB} = km - (\kappa - 1) - (m - 1) - 1 = (\kappa - 1)(m - 1)$.

Число степеней свободы сумм квадратов остатков SS_R равно $\nu_R = n - km$.

Соответственно средние суммы квадратов будут равны:

$MS_A = \frac{SS_A}{k-1}$, $MS_{\text{B}} = \frac{SS_B}{m-1}$, $MS_A = \frac{SS_{AB}}{(k-1)(m-1)}$, $MS_R = \frac{SS_R}{n-km}$.

Поскольку двухфакторная модель учитывает различные эффекты влияния факторов, то и статистический анализ для двухфакторной модели предполагает проверку гипотез о значимости различных эффектов. В качестве статистик критериев проверки гипотез о значимости соответствующих эффектов используются отношения средней суммы квадратов эффектов к средней сумме квадратов остатков. При условии истинности H_0: «эффект незначим» и нормальном распределении остатков данные статистики имеют распределение Фишера с параметрами степеней свободы, определяемыми числами степеней свободы соответствующих сумм, участвующих в отношении. В таблице 3 приведены основные рассматриваемые гипотезы, статистики критериев для проверки данных гипотез и соответствующие числа степеней свободы данных статистик.

Таблица 3 – Статистики для проверки гипотез двухфакторного дисперсионного анализа

Основная гипотеза:	Все $\alpha_i = 0$	Все $\beta_j = 0$	Все $\gamma_{ij} = 0$
Статистика критерия	MS_A/MS_R	MS_B/MS_R	MS_{AB}/MS_R
Числа степеней свободы	$\nu_1 = k-1$ $\nu_2 = n-nk$	$\nu_1 = m-1$ $\nu_2 = n-nk$	$\nu_1 = (k-1)(m-1)$ $\nu_2 = n-nk$

Если наблюдаемое значение статистики $F_{\text{набл}} \geq F_{\text{кр}}$, где $F_{\text{кр}}$ – критическая точка распределения Фишера уровня α (или квантиль уровня $1-\alpha$) с числом степеней свободы ν_1 и ν_2, то нулевая гипотеза отклоняется и считается, что средние для различных уровней фактора значимо различаются.

Контрольные вопросы

1. Что такое дисперсионный анализ (ANOVA)?
2. Какова основная цель дисперсионного анализа?
3. Какие типы дисперсионного анализа существуют? Приведите примеры.
4. Что такое однофакторный дисперсионный анализ?
5. Каковы предпосылки для применения однофакторного дисперсионного анализа?
6. Что такое многофакторный дисперсионный анализ?
7. Каковы основные этапы проведения дисперсионного анализа?
9. Как интерпретировать результаты дисперсионного анализа?
10. Что такое пост-хок тесты и когда они применяются?
11. Каковы основные ограничения дисперсионного анализа?
12. Каковы альтернативы дисперсионному анализу, если его предпосылки не выполняются?

2.6 Тема 6. Анализ зависимостей

Вопросы для изучения

1. Функциональная, статистическая, корреляционная зависимости
2. Корреляция Пирсона
3. Ранговая и частная корреляция
4. Матрицы сопряженности

Методические указания

Основными понятиями корреляционного анализа зависимостей являются: положительная и отрицательная корреляция, зависимость и независимость признаков, связь между этими понятиями. Основным инструментом анализа корреляции является расчет коэффициентов корреляции (Пирсона, Спирмена и Кен-

далла), частных коэффициентов корреляции, корреляционных матриц и матриц сопряженности. Важно уметь интерпретировать результаты корреляционного анализа и понимать ограничения метода.

Рекомендуемые источники: [1, гл. 12]; [2, т. 1, гл. 10–11].

Программное обеспечение

Для выполнения расчетов и визуализации данных рекомендуется использовать:

– Python (библиотеки pandas, numpy, scipy и seaborn);

– R (пакеты cor и corrplot);

– Excel (инструмент «Анализ данных»);

– SPSS (набор функций для корреляционного анализа).

Основные теоретические сведения и решение типовых задач

Линейная корреляция

Если система состоит из двух случайных величин X и Y, связанных линейной зависимостью, ее характеристиками являются:

– начальные моменты $\nu_{10} = M(X)$, $\nu_{01} = M(Y)$,

– центральные моменты $\mu_{20} = D(X)$, $\mu_{02} = D(Y)$,

– центральный момент $\mu_{11} = M\left(\mathring{X}\mathring{Y}\right) = k_{xy}$ – *корреляционный момент*,

– *коэффициент линейной корреляции* $\rho_{xy} = \frac{k_{xy}}{\sigma_x \sigma_y}$ – показатель силы линейной связи между X и Y.

В статистическом анализе используются их соответствующие выборочные оценки.

Если объем выборки невелик, выборочные моменты определяются по следующим формулам:

– выборочные средние: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$;

– выборочные (исправленные) дисперсии:

$$\hat{S}_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n}{n-1}(\overline{x^2} - (\bar{x})^2),$$

$$\hat{S}_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{n}{n-1}(\overline{y^2} - (\bar{y})^2);$$

– выборочный корреляционный момент:

$$k_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \overline{xy} - \bar{x}\,\bar{y}\ ;$$

– выборочный коэффициент линейной корреляции:

$$r_{xy} = \frac{k_{xy}}{\hat{S}_x\,\hat{S}_y}\ .$$

1. Выборочный коэффициент корреляции не зависит от выбора начала отсчета и единицы измерения; иными словами, при любых a_1, a_2, b_1 и b_2

$$r(a_1x + b_1, a_2y + b_2) = r_{xy}.$$

2. Выборочный коэффициент корреляции не превышает единицы, $|r_{xy}| \le 1$.

3. Если $Y = aX + b$, то $r_{xy} = \frac{a}{|a|}$, т. е. $r_{xy} = \pm 1$.

Чем ближе $|r_{xy}|$ к единице, тем сильнее связь, тем меньше представлены в ней случайные факторы. При $|r_{xy}| = 1$ случайные величины связаны *линейной функциональной* зависимостью.

Как всякая выборочная оценка, выборочный коэффициент корреляции является величиной случайной, достоверность (значимость) которой следует проверить с помощью того или иного критерия.

Если система (X, Y) распределена нормально, то вопрос о значимости коэффициента корреляции решается с помощью случайной величины $T = r\sqrt{\frac{n-2}{1-r^2}}$, распределенной по закону Стьюдента с $\nu = n - 2$ степенями свободы. Высказывается гипотеза H_o: $\rho_{xy} = 0$. Если $t_{эмп} = r_{xy}\sqrt{\frac{n-2}{1-r_{xy}^2}}$ по абсолютной величине не превышает критического значения $t_{кр} = t_{\beta,\nu}$, полученного из таблицы t-распределения, то гипотеза H_o принимается. Если $|t_{эмп}| > t_{\beta,\nu}$, гипотеза отвергается, корреляционная связь между признаками X и Y признается значимой.

При больших объемах выборки n выборочный коэффициент корреляции r_{xy} распределен асимптотически нормально с параметрами

$$m_r = \rho_{xy} \quad \text{и} \quad \sigma_r = \frac{1 - r_{xy}^2}{\sqrt{n}}\ .$$

Криволинейная корреляция

Если связь между случайными величинами X и Y нелинейна, то для ее

оценки используется общая методика, основанная на сравнении двух дисперсий.

Пусть результаты выборочного обследования системы (*X*,*Y*) сведены в корреляционную таблицу, и пусть *X* является факторным признаком, а *Y* – результативным.

Общая дисперсия S_y^2 случайного признака Y равна сумме дисперсии $S_{\bar{y}/x}^2$ условных средних этого признака («межгрупповой дисперсии») и средней внутригрупповых дисперсий $\bar{S}_{y/x}^2$ («остаточной дисперсии»):

$$S_y^2 = S_{\bar{y}/x}^2 + \bar{S}_{y/x}^2.$$

Степень влияния признака *X* на изменчивость признака *Y* характеризуется *корреляционным отношением*

$$\eta_{y/x}^2 = \frac{\sigma_{\bar{y}/x}^2}{\sigma_y^2},$$

точечной оценкой которого служит *выборочное корреляционное отношение*

$$\hat{\eta}_{y/x}^2 = \frac{S_{\bar{y}/x}^2}{S_y^2}.$$

Величина корреляционного отношения заключена между 0 и 1, причем, если $\hat{\eta}_{y/x}^2 = 1$, то случайные признаки связаны функциональной зависимостью, а если $\hat{\eta}_{y/x}^2 = 0$, случайные признаки *X* и *Y* некоррелированы.

Корреляционное отношение характеризует тесноту связи между случайными признаками *независимо от ее формы*. Соотношение между линейным коэффициентом корреляции и корреляционным отношением таково: $\rho \leq \eta$. Разность $\eta - \rho$ является мерой нелинейности корреляционной связи.

Понятие о множественной корреляции

Корреляционный анализ системы случайных величин (X_1, X_2, …, X_m) предусматривает прежде всего исследование взаимного влияния каждой пары составляющих. В результате получают оценки *парных коэффициентов корреляции* $r_{i,j}$, $i = 1,2,\ldots,m$, $j = 1,,\ldots,m$, совокупность которых составляет *корреляционную матрицу*

$$q_m = \begin{pmatrix} 1 & r_{12} & \ldots & r_{1m} \\ r_{21} & 1 & \ldots & r_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ r_{m1} & r_{m2} & \ldots & 1 \end{pmatrix}$$

Известно, что на результативный признак воздействуют не только от-

дельно взятые факториальные признаки X_i, но и их взаимное влияние. Чтобы выявить связь между двумя из них, исключив влияние остальных признаков, вводят понятие *частного коэффициента корреляции.*

Коэффициент линейной корреляции между составляющими X_k и X_j системы $(X_1, X_2, \ldots, X_m)$, вычисленный при условии, что l из оставшихся m–2 факториальных признаков зафиксированы, называется *частным коэффициентом корреляции* l-го порядка. Его выборочная оценка обозначается $r_{kj\cdot n_1 n_2 \ldots n_l}$, где $n_1, n_2, \ldots, n_l$ $(l \leq n-2)$ – номера фиксированных признаков.

Для вычисления частного коэффициента корреляции l-го порядка составляют вспомогательную матрицу порядка l+2 из элементов матрицы q_m , индексы которых соответствуют индексам коэффициентов частной корреляции (при $l = m - 2$ используется вся матрица q_m).

В частности, выборочный коэффициент корреляции $r_{12\cdot 34\ldots m}$ является мерой линейной связи между составляющими X_1 и X_2 при фиксированных X_3, …, X_m и вычисляется по формуле

$$r_{12\cdot 3\ldots m} = -\frac{d_{12}}{\sqrt{d_{11}d_{22}}},$$

где d_{12}, d_{11}, d_{22} – алгебраические дополнения элементов r_{12}, r_{11}, r_{22} корреляционной матрицы q_m.

Общую корреляционную связь одного из признаков X_k системы $(X_1, X_2, \ldots, X_m)$ со всеми остальными ее составляющими можно определить с помощью *множественного (совокупного) коэффициента корреляции*, оценкой которого является величина

$$R_k = \sqrt{1 - \frac{det\, q_m}{d_{kk}}},$$

где $det\, q_m$ – определитель корреляционной матрицы q_m; d_{kk} – алгебраическое дополнение элемента r_{kk} этой матрицы.

Пример. Даны результаты 15 наблюдений над системой случайных величин (X_1, X_2, X_3, X_4):

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X_1	9,8	8,2	7,6	8,5	5,9	9,3	12,	12,	11,	9,9	13,	14,	14,	7,5	8,1
X_2	57,	44,	40,	44,	30,	47,	78,	62,	60,	55,	72,	73,	80,	38,	42,
X_3	50,	28,	15,	21,	2,0	35,	98,	67,	59,	52,	83,	87,	10	14,	23,
X_4	12,	7,1	4,3	5,1	1,4	7,4	15,	11,	12,	9,8	15,	14,	18,	4,7	5,5

Требуется: а) составить выборочную корреляционную матрицу; б) вычислить выборочный частный коэффициент корреляции $r_{12\cdot3}$ между признаками X_1 и X_2 при фиксированном X_3; в) определить множественный коэффициент корреляции R_1.

Решение. Вычислим выборочные средние, несмещенные средние квадратичные отклонения данных признаков и их парные корреляционные моменты:

$$\bar{x}_1 = 10{,}24; \quad \bar{x}_2 = 55{,}38; \quad \bar{x}_3 = 49{,}28; \quad \bar{x}_4 = 9{,}75;$$
$$\hat{S}_1 = 2{,}62; \quad \hat{S}_2 = 15{,}25; \quad \hat{S}_3 = 31{,}43; \quad \hat{S}_4 = 7{,}95;$$
$$\overline{x_1x_2} = 606{,}26; \quad \overline{x_1x_3} = 585{,}31; \quad \overline{x_1x_4} = 112{,}39;$$
$$\overline{x_2x_3} = 3207; \quad \overline{x_2x_4} = 614{,}33; \quad \overline{x_3x_4} = 633{,}57;$$
$$k_{12} = 39{,}165; \quad k_{13} = 80{,}685; \quad k_{14} = 12{,}58;$$
$$k_{23} = 477{,}73; \quad k_{24} = 74{,}56; \quad k_{34} = 153{,}26.$$

Теперь вычислим парные коэффициенты корреляции и составим корреляционную матрицу.

$$r_{12} = 0{,}979; \quad r_{13} = 0{,}979; \quad r_{14} = 0{,}969; \quad r_{23} = 0{,}997;$$
$$r_{24} = 0{,}988; \quad r_{34} = 0{,}985;$$

$$q_4 = \begin{pmatrix} 1 & 0{,}979 & 0{,}979 & 0{,}969 \\ 0{,}979 & 1 & 0{,}997 & 0{,}988 \\ 0{,}979 & 0{,}997 & 1 & 0{,}985 \\ 0{,}969 & 0{,}988 & 0{,}985 & 1 \end{pmatrix}.$$

Чтобы вычислить выборочный частный коэффициент корреляции $r_{12\cdot3}$, составим вспомогательную матрицу, состоящую из элементов корреляционной матрицы, индексы которых имеют цифры 1, 2 и 3:

$$\begin{pmatrix} 1 & 0{,}979 & 0{,}979 \\ 0{,}979 & 1 & 0{,}997 \\ 0{,}979 & 0{,}997 & 1 \end{pmatrix}.$$

Вычисляем алгебраические дополнения элементов r_{12}, r_{11}, r_{22} этой матрицы.

$$d_{12} = \begin{vmatrix} 0{,}979 & 0{,}997 \\ 0{,}979 & 1 \end{vmatrix} = 0{,}002937; \quad d_{11} = \begin{vmatrix} 1 & 0{,}997 \\ 0{,}997 & 1 \end{vmatrix} = 0{,}005991;$$

$$d_{22} = \begin{vmatrix} 1 & 0{,}979 \\ 0{,}979 & 1 \end{vmatrix} = 0{,}041559.$$

Наконец, вычисляем частный коэффициент корреляции $r_{12\cdot3}$:

$$r_{12\cdot3} = \frac{d_{12}}{\sqrt{d_{11}\cdot d_{22}}} = \frac{0{,}002937}{\sqrt{0{,}005991\cdot 0{,}041559}} = \frac{0{,}002937}{0{,}015779} = 0{,}1861.$$

Множественный коэффициент корреляции R_1 характеризует влияние факториальных признаков X_2, X_3, X_4 на результативный признак X_1:

$$R_1 = \sqrt{1 - \frac{det\, q_4}{d_{11}}};$$

где $det\, q_4 = 5{,}715 \cdot 10^{-6}, \quad d_{11} = 5{,}991 \cdot 10^{-3}$.

Отсюда $R_1 = \sqrt{1 - \frac{5{,}715\cdot 10^{-6}}{5{,}991\cdot 10^{-3}}} = 0{,}9995.$

Как видно, обобщенный коэффициент корреляции весьма близок к единице, что говорит об очень сильной корреляционной связи между данными признаками.

Контрольные вопросы

1. Что такое корреляция и как она измеряется?
2. Что такое коэффициент корреляции Пирсона и как он интерпретируется?
3. Что такое коэффициент корреляции Спирмена и как он интерпретируется?
4. В чем разница между коэффициентом корреляции Пирсона и Спирмена?
5. Что такое частная корреляция и как она рассчитывается и как интерпретируется?
6. Как визуализировать корреляцию между переменными?
7. Что такое матрицы сопряженности?
8. Каковы ограничения корреляционного анализа?

2.7 Тема 7. Линейный регрессионный анализ

Вопросы для изучения

1. Классическая модель парной линейной регрессии.
2. Оценка параметров модели парной линейной регрессии
3. Множественная регрессия: расширение модели, интерпретация результатов.
4. Оценка точности и достоверности модели.
5. Теорема Маркова-Гаусса. Анализ регрессионных остатков.
6. Проблема мультиколлинеарности и гетероскедастичности.

Методические указания

Линейная регрессия – это метод анализа зависимости между одной или несколькими независимыми переменными (предикторами) и одной зависимой переменной (откликом или результатом). Целью является построение модели, которая описывает эту зависимость и позволяет делать прогнозы. Для изучения темы «Линейный регрессионный анализ» потребуется понимания основ линейной алгебры.

Изучите принципы построения модели линейной регрессии, включая определение уравнения регрессии и коэффициентов. Изучите методы оценки качества модели, такие как коэффициент детерминации (R^2) и стандартные ошибки. Уделите внимание интерпретации результатов и пониманию ограничений метода.

При построении регрессионных моделей важно учитывать выполнение условий Маркова-Гаусса, так как они обеспечивают корректность использования метода наименьших квадратов. Обратите внимание на проблему мультиколлинеарности, которая возникает при высокой корреляции между независимыми переменными и может привести к недостоверным результатам. Анализ регрессионных остатков помогает выявить наличие гетероскедастичности и автокорреляции, что необходимо для правильной интерпретации модели.

Изучите возможности компьютерных пакетов анализа данных и программных библиотек для решения задач моделирования и прогнозирования с помощью линейной регрессии.

Рекомендуемые источники: [1, гл. 13]; [2, т. 2, гл. 2]; [3, гл. 3].

Программное обеспечение

Для выполнения расчетов и визуализации данных рекомендуется использовать:

– Python (библиотеки numpy, pandas, scipy, statsmodels)

– R (пакеты lm, ggplot2)

– Excel (инструмент «Анализ данных»)

Основные теоретические сведения и решение типовых задач

Задача регрессии – выбор модели зависимости между переменными и определение оценок неизвестных параметров этой модели.

Выбор модели регрессионных зависимостей осуществляется исходя из теоретических представлений о возможной взаимосвязи между переменными или из визуального анализа графиков наблюдений.

В зависимости от количества включенных в модель факторов *X* модели делятся на *однофакторные* (парная модель регрессии) и *многофакторные* (модель множественной регрессии).

В зависимости от вида функции $f(x_1, x_2, \dots, x_p)$ модели делятся на *линейные* и *нелинейные*.

Линейная регрессия

Однофакторная регрессия (на наблюдаемую переменную *Y* влияет один фактор *X*) $y = \beta_0 + \beta_1 x + \varepsilon$ ***Замечание***. *Уравнение называется простой линейной регрессией или парной линейной регрессией*	***Множественная регрессия*** (на наблюдаемую переменную *Y* влияют несколько факторов $X_1, X_2, \dots, X_p$) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$
Для оценки неизвестных параметров β применяют метод наименьших квадратов (МНК), суть которого состоит в минимизации суммы квадратов отклонений фактических значений результатного признака y_i от его расчетных значений $\hat{y}_i$, т. е.: $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \to min$	
Алгоритм МНК в форме обобщенного обращения матрицы	
1. Ввести исходные данные – массивы *Y* и *X*. 2. Составить матрицу $A_{n\text{x}2}$, *n* – число наблюдений, 2- число неизвестных параметров $A = \begin{pmatrix} 1 & x_1 \\ \dots & \dots \\ 1 & x_n \end{pmatrix}$.	1. Ввести исходные данные – массивы *Y* и $X_1, X_2, \dots, X_p$. 2. Составить матрицу $A_{n\text{x}(\text{p}+1)}$, *n* – число наблюдений, *p+1*- число неизвестных параметров $A = \begin{pmatrix} 1 & x_{11} \dots & x_{p1} \\ \dots & \dots \dots & \dots \\ 1 & x_{1n} \dots & x_{pn} \end{pmatrix}$.
3. Матрица неизвестных параметров $\beta = (A^T A)^{-1} A^T Y$.	
В среде MathCAD для определения параметров простой линейной регрессии можно использовать встроенные функции 1. $\beta_1 = slope(X, Y)$, $\beta_0 = intercept(X, Y)$;	

2. line(X,Y)	

Качество уравнения регрессии определяется по величине средней ошибки аппроксимации

$$\overline{A} = \frac{1}{n}\sum\left|\frac{y_i - f(x_i)}{y_i}\right| \cdot 100\%$$

(уравнение можно использовать как прогностическую модель, если $\overline{A} \leq 15\ \%$).

Влияние совокупности факторов на результат Y

Выборочный линейный коэффициент корреляции *(характеризует степень взаимосвязи пары случайных величин, если зависимость между ними соответствует прямой линии)* $r_{xy} = r_{yx} = \frac{\sum(x_i - \overline{x}_в)(y_i - \overline{y}_в)}{s_x s_y}$ $-1 \leq r \leq 1$; $\lvert r\rvert = 1$ – линейная функциональная связь, $\beta_1 \neq 0$; $r = 0$ - *Y* и *X* некоррелированы. В среде MathCAD используют встроенную функцию *corr(X,Y)*	***Выборочный сводный коэффициент корреляции R_s*** *(характеризует связь Y со всеми факторами, входящими в уравнение)* $R_s = \sqrt{1 - \frac{\sum(y_i - \hat{y}_i)^2}{nD_в(y)}}$ $0 \leq R_s \leq 1$; $R_s = 1$ - *Y* имеет функциональную связь с совокупностью факторов; $R_s = 0$ - *Y* некоррелирован ни с одним из факторов

Проверка значимости выборочного коэффициента корреляции (t-критерий Стьюдента)

H_0 – изучаемый фактор (факторы) не оказывает существенного влияния на результат, т. е. коэффициент корреляции генеральной совокупности равен 0;

H_1 – коэффициент корреляции генеральной совокупности отличен от 0.

$t_{\text{критическое}} = qt(1 - \frac{\alpha}{2}, n - p - 1)$;

p – число факторов, влияющих на результат;

n – число измерений;

w – критическая область двусторонняя

$t_{\text{набл}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$	$t_{\text{набл}} = \frac{R_s\sqrt{n-p-1}}{\sqrt{1-{R_s}^2}}$

Проверка значимости уравнения регрессии (F-критерий Фишера) – *установить, соответствует ли математическая модель экспериментальным данным и достаточно ли включенных в уравнение факторов (од-*

ного или нескольких) для описания зависимой переменной

H_0 – уравнение регрессии не надежное;

H_1 – уравнение регрессии надежное

$$F_{\text{критическое}} = qF(1-\alpha, p, n-p-1)$$

p – число факторов, влияющих на результат;

n – число измерений;

w – критическая область левосторонняя

$F_{\text{наблюдаемое}} = \frac{r_{xy}^2(n-2)}{1-r_{xy}^2}$	$F_{\text{наблюдаемое}} = \frac{R^2(n-p-1)}{1-R^2}$

Значимость отдельных (кроме свободного члена) коэффициентов регрессии (t-критерий Стьюдента)

H_0 – коэффициент статистически не значим;

H_1 – коэффициент статистически значим

$$t_{\text{критическое}} = qt(1-\frac{\alpha}{2}, n-p-1)$$

p – число факторов, влияющих на результат;

n – число измерений;

w – критическая область двусторонняя.

Если коэффициент статистически не значим, то фактор, соответствующий этому коэффициенту следует исключить из модели (при этом ее качество не ухудшится)

Коэффициенты эластичности и детерминации

1. Коэффициент эластичности $E_i = \beta_i \frac{\overline{x}_{\text{в}i}}{\overline{y}_{\text{в}}}$ показывает, на сколько процентов в среднем изменяется результативный признак Y при изменении факторного признака X_i на 1 %. Высокий уровень эластичности означает сильное влияние независимой переменной на объясняемую переменную.

2. Коэффициент детерминации r_{xy}^2 (R^2) показывает долю вариации результативного признака, объясненную вариацией факторного признака. Чаще всего, давая интерпретацию коэффициента детерминации, его выражают в процентах

Особенности практического применения линейных множественных регрессионных моделей

Одним из условий регрессионной модели является предположение о линейной независимости объясняющих переменных, т. е., решение задачи возможно лишь тогда, когда столбцы и строки матрицы исходных данных линейно

независимы. Для экономических показателей это условие выполняется не всегда.

Под ***мультиколлинеарностью*** понимается высокая взаимная коррелированность объясняющих переменных (факторов), которая приводит к линейной зависимости нормальных уравнений. Существует несколько способов для определения наличия или отсутствия мультиколлинеарности. Один из подходов заключается в анализе коэффициентов парной корреляции.

1. Факторные признаки, у которых $r_{yx_i} < 0{,}5$ исключают из модели.

2. Считают явление мультиколлинеарности в исходных данных установленным, если коэффициент парной корреляции между двумя переменными (факторами) больше 0,8. В этом случае одну переменную исключают из рассмотрения. При этом какую переменную оставить, а какую удалить из анализа, решают в первую очередь на основании экономических соображений. Если с экономической точки зрения ни одной из переменных нельзя отдать предпочтение, то оставляют ту из двух переменных, которая имеет больший коэффициент корреляции с зависимой переменной.

Контрольные вопросы

1. Что такое линейная регрессия и для чего она используется?
2. Какова формула модели линейной регрессии?
3. Что такое зависимая и независимая переменные в контексте линейной регрессии?
4. Какие методы лежат в основе оценки параметров модели линейной регрессии?
5. Как интерпретировать коэффициенты линейной регрессии?
6. Что такое R-квадрат и как он оценивает качество модели?
7. Каковы предпосылки для применения линейной регрессии?
8. Что такое остатки в линейной регрессии и как их анализировать?
9. Как проверить значимость коэффициентов линейной регрессии?
10. Что такое мультиколлинеарность и как она влияет на модель линейной регрессии?
11. Как можно улучшить модель линейной регрессии?

3 МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО САМОСТОЯТЕЛЬНОЙ РАБОТЕ

Внеаудиторная самостоятельная работа в рамках данной дисциплины включает в себя:

– подготовку к аудиторным занятиям (лекциям, практическим занятиям, лабораторным работам) и выполнение соответствующих заданий;

– самостоятельную работу над отдельными темами учебной дисциплины в соответствии с тематическим планом;

– подготовка к текущему контролю в виде контрольных срезов по разделам дисциплины;

– выполнение курсовой работы;

– выполнение контрольных работ для студента заочной формы обучения;

– подготовку к экзамену.

Подготовка к лекционным занятиям

При подготовке к лекции рекомендуется повторить ранее изученный материал, что дает возможность получить необходимые разъяснения преподавателя непосредственно в ходе занятия. Рекомендуется вести конспект, главное требование к которому быть систематическим, логически связанным, ясным и кратким. По окончанию занятия обязательно в часы самостоятельной подготовки, по возможности в этот же день, повторить изучаемый материал и доработать конспект.

Подготовка к практическим занятиям

Подготовка к практическим занятиям предусматривает:

– изучение теоретических положений, лежащих в основе решения типовых задач и выполнения практических заданий;

– проработку учебного материала, рекомендованной литературы и методической разработки на предстоящее занятие.

Задачи и практические задания, предназначенные для выполнения на практических занятиях под руководством преподавателя и самостоятельно в рамках домашнего задания для дополнительной проработки тем дисциплины, и представляют собой подборки практических задач.

4. ОЦЕНОЧНЫЕ СРЕДСТВА ДЛЯ ТЕКУЩЕЙ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ

К оценочным средствам текущего контроля успеваемости относятся:

– тестовые задания открытого и закрытого типов.

Промежуточная аттестация в форме зачета (второй или третий семестр) проходит по результатам прохождения всех видов текущего контроля успеваемости. В отдельных случаях (при непрохождении всех видов текущего контроля) зачет может быть проведен в виде тестирования.

Универсальная система оценивания результатов обучения включает в себя системы оценок: 1) «отлично», «хорошо», «удовлетворительно», «неудовлетворительно»; 2) «зачтено», «не зачтено»; 3) 100-балльную/процентную систему и правило перевода оценок в пятибалльную систему (таблица 4).

Таблица 4 – Система оценок и критерии выставления оценки

Система оценок	**2**	**3**	**4**	**5**
	0–40 %	**41–60 %**	**61–80 %**	**81–100 %**
	«неудовлетворительно»	**«удовлетворительно»**	**«хорошо»**	**«отлично»**
Критерий	**«не зачтено»**	**«зачтено»**		
1 Системность и полнота знаний в отношении изучаемых объектов	Обладает частичными и разрозненными знаниями, которые не может научно-корректно связывать между собой (только некоторые из которых может связывать между собой)	Обладает минимальным набором знаний, необходимым для системного взгляда на изучаемый объект	Обладает набором знаний, достаточным для системного взгляда на изучаемый объект	Обладает полнотой знаний и системным взглядом на изучаемый объект
2 Работа с информацией	Не в состоянии находить необходимую информацию, либо в состоянии находить отдельные фрагменты информации в рамках поставленной задачи	Может найти необходимую информацию в рамках поставленной задачи	Может найти, интерпретировать и систематизировать необходимую информацию в рамках поставленной задачи	Может найти, систематизировать необходимую информацию, а также выявить новые, дополнительные источники информации в рамках поставленной задачи
3 Научное осмысление изучаемого явления, процесса, объекта	Не может делать научно-корректных выводов из имеющихся у него сведений, в состоянии проанализировать только некоторые из имеющихся у него сведений	В состоянии осуществлять научно-корректный анализ предоставленной информации	В состоянии осуществлять систематический и научно-корректный анализ предоставленной информации, вовлекает в исследование новые релевантные поставленной задаче данные	В состоянии осуществлять систематический и научно-корректный анализ предоставленной информации, вовлекает в исследование новые релевантные поставленной задаче данные, предлагает новые ракурсы поставленной задачи
4 Освоение стандартных алгоритмов решения профессиональных задач	В состоянии решать только фрагменты поставленной задачи в соответствии с заданным алгоритмом, не освоил предложенный алгоритм, допускает ошибки	В состоянии решать поставленные задачи в соответствии с заданным алгоритмом	В состоянии решать поставленные задачи в соответствии с заданным алгоритмом, понимает основы предложенного алгоритма	Не только владеет алгоритмом и понимает его основы, но и предлагает новые решения в рамках поставленной задачи

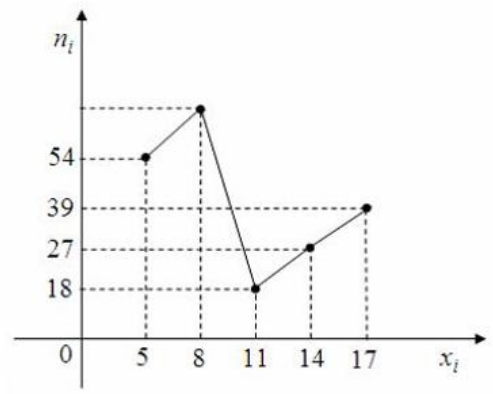
Оценивание тестовых заданий закрытого типа осуществляется по системе зачтено/ не зачтено («зачтено» – 41–100 % правильных ответов; «не зачтено» – менее 40 % правильных ответов) или пятибалльной системе (оценка «неудовлетворительно» – менее 40 % правильных ответов; оценка «удовлетворительно» – от 41 до 60 % правильных ответов; оценка «хорошо» – от 61 до 80% правильных ответов; оценка «отлично» – от 81 до 100 % правильных ответов).

Тестовые задания открытого типа оцениваются по системе «зачтено/ не зачтено». Оценивается верность ответа по существу вопроса, при этом не учитывается порядок слов в словосочетании, верность окончаний, падежи.

УК-4 способен осуществлять деловую коммуникацию в устной и письменной формах на государственном языке Российской Федерации и иностранном(ых) языке(ах).

Тестовые задания открытого типа:

1. Из генеральной совокупности извлечена выборка объема n =200, полигон частот которой имеет вид



Тогда частота варианты $x_2=8$ равна …

Ответ: 62

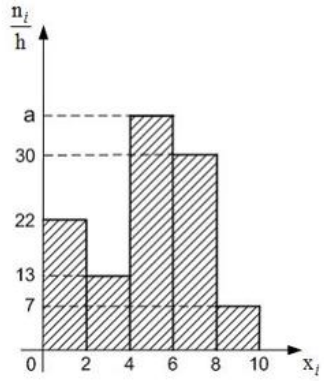
2. Выборка объема n=10 задана рядом частот

x_i	2	3	5	7
n_i	n_1	2	1	3

Тогда значение выборочной средней равно …

Ответ: 4

3. Из генеральной совокупности извлечена выборка объема n =220, гистограмма частот которой имеет вид



Тогда значение **а** равно …

Ответ: 38

4. Размах вариационного ряда 1,2,2,3,5,5,7,7,7,12 превышает его моду на …

Ответ: 4

5. Дан вариационный ряд выборки объема n = 10: –2, 0, 3, 3, 4, 5, 9, 11, 12, 15.

Выборочная медиана для этого ряда равна … (записать ответ в виде десятичной дроби)

Ответ: 4,5

6. Проведено четыре измерения (без систематических ошибок) некоторой случайной величины (в мм): 5, 6, 9, 12.

Несмещенная оценка математического ожидания равна …

Ответ: 8

7. В результате измерений некоторой физической величины одним прибором (без систематических ошибок) получены следующие результаты (в мм): 11, 13, 15. Тогда несмещенная оценка дисперсии измерений равна …

Ответ: 4

8. Точечная оценка математического ожидания нормально распределенного генерального признака равна 10, интервальная – (8;12). Точность оценки равна …

Ответ: 2

9. При проверке статистической гипотезы вероятность совершить ошибку первого рода называется …

Ответ: уровень значимости

10. Ошибка, состоящая в том, что будет отвергнута верная гипотеза, называется ошибкой …

Ответ: первого рода

11. p-value – это наименьшая величина уровня значимости, при которой нулевая гипотеза …

Ответ: отвергается

12. Величина, характеризующая существование линейной зависимости между двумя величинами, называется …

Ответ: коэффициент корреляции

Тестовые задания закрытого типа:

13. Выборка наблюдений, представленная в порядке возрастания, называется:
-: упорядоченным рядом;
+: вариационным рядом;
-: упорядоченной выборкой;
-: статистическим рядом.

14. С ростом значения надежности ширина доверительного интервала
+: увеличивается;
-: уменьшается;
-: не изменяется;
-: может как увеличиться, так и уменьшиться.

15. Критерий согласия – это критерий для проверки гипотезы

-: о равенстве параметров нескольких генеральных совокупностей;

-: о числовых значениях параметров генеральной совокупности;

-: об однородности выборок;

+: о законе распределения генеральной совокупности.

16. Какое значение НЕ может принимать парный коэффициент корреляции:
-: -0,973;
-: 0,005;
+: 1,111;
-: 0,721.

УК-6 способен управлять своим временем, выстраивать и реализовывать траекторию саморазвития на основе принципов образования в течение всей жизни

Тестовые задания открытого типа:

17. Задано статистическое распределение выборки

x_i	2	4	5	6
n_i	8	9	10	3

Выборочная средняя равна …

Ответ: 4

18. Задано статистическое распределение выборки

x_i	2	4	5	6
n_i	8	9	10	3

Медиана равна …

Ответ: 4

19. Задано статистическое распределение выборки

x_i	2	4	5	6
n_i	8	9	10	3

Мода равна …

Ответ: 5

20. Задано статистическое распределение выборки

x_i	2	4	5	6
n_i	8	9	10	3

Несмещенная оценка генеральной дисперсии равна … (записать ответ с точностью до сотых)

Ответ: 1,86

21. Проверяется гипотеза о нормальном распределении генеральной совокупности. При уровне значимости $\alpha = 0{,}05$ получены следующие результаты: $\chi^2_{\text{набл}} = 15{,}02$, $\chi^2_{\text{крит}} = 18{,}5$.
Вывод: выборочные данные …. с нормальным распределением

Ответ: согласуются

22. Минимальное количество групп необходимое для проведения однофакторного дисперсионного анализа (ANOVA) равно …

Ответ: 2

23. Количество степеней свободы ошибки в однофакторном дисперсионном анализе с 5 группами и 50 наблюдениями равно …

Ответ: 45

24. Для выборки 2; 7; 4; 9; 7; 8; 5; 12; 11; 8; 12; 14; 12; 6; 3;

исправленная дисперсия равна … (*укажите значение с точностью 2 знака после запятой)*

Ответ: 13,28

25. Отношение среднего квадратического отклонения к математическому ожиданию называется …

Ответ: коэффициент вариации

26. Оценка математического ожидания равна 5, выборочная дисперсия равна 625. Тогда выборочный коэффициент вариации равен ….

Ответ: 5

27. Если между двумя величинами отсутствует линейная зависимость, то для них линейный коэффициент корреляции равен …

Ответ: 0

28. Если между двумя величинами существует прямая функциональная линейная зависимость, то для них коэффициент корреляции равен …

Ответ: 1

Тестовые задания закрытого типа:

29. При каком значении линейного коэффициента корреляции связь между признаками можно считать тесной:

+: -0,975;

-: 0,657;

-: -0,111;

-: 0,421.

30. Пусть для гипотезы о равенстве двух средних H_0: $a_x = a_y$ альтернативная гипотеза H_1 имеет вид $a_x > a_y$. Значение $t_{набл}$=1,2, $t_{крит}$ = 2,015, α=0,05. Верным является решение:

-: различие между генеральными средними статистически значимо и не является случайным;

+: различие между генеральными средними статистически незначимо и объясняется случайными причинами;

-: генеральное среднее выборки *X* больше генеральной средней выборки *Y*;

-: гипотеза H_0 верна с вероятностью 0,95;

-: гипотеза H_0 отвергается и принимается H_1.

31. Основными свойствами точечных оценок являются

+: несмещенность;

+: состоятельность;

-: достоверность;

+: эффективность;

-: полнота.

32. При увеличении объема выборки доверительный интервал

-: не изменяется;

-: смещается на величину доверительной вероятности;

+: уменьшается;

-: увеличивается.

СПИСОК ЛИТЕРАТУРЫ

1. Кремер, Н. Ш. Теория вероятностей и математическая статистика: учебник и практикум для вузов / Н. Ш. Кремер. – 5-е изд., перераб. и доп. – Москва: Издательство Юрайт, 2024.

2. Айвазян, С. А. Прикладная статистика. Основы эконометрики: учебник для вузов: в 2 т. / С. А. Айвазян, В. С. Мхитарян. – Москва: ЮНИТИ-ДАНА, 2001. – Т. 1, 2

3. Буре, В. М. Методы прикладной статистики в R и Excel: учеб. пособие / В. М. Буре, Е. М. Парилина, А. А. Седаков. – Санкт-Петербург: Издательство «Лань», 2018.

Локальный электронный методический материал

Елена Юрьевна Скоробогатых

ПРИКЛАДНАЯ СТАТИСТИКА

Редактор С. Кондрашова
Корректор Т. Звада

Уч.-изд. л. 4,5. Печ. л. 4,1.

Издательство федерального государственного бюджетного
образовательного учреждения высшего образования
«Калининградский государственный технический университет».
236022, Калининград, Советский проспект, 1