Министерство образования и науки Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
«Томский государственный архитектурно-строительный университет»

ОСНОВЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Методические указания
для самостоятельной работы студентов

Составитель И.А. Иконникова

Томск 2014

Основы математической статистики/Составитель И.А. Иконникова. – Томск: Изд-во Том. гос. архит.-строит. ун-та, 2014. – 41 с.

Рецензент О.В. Иванова
Редактор О.А. Сергеева

Методические указания к самостоятельной работе по дисциплине Б2.Б.1 – «Математика» при изучении темы «Математическая статистика» студентами второго курса заочной формы обучения всех направлений всех специальностей и профилей подготовки специалистов и бакалавров.

Печатаются по решению методического семинара кафедры высшей математики, протокол № 4 от 7 февраля 2014 г.

срок действия с 01.09.2014 до 01.09.2019

Оригинал-макет подготовлен И.А. Иконниковой.

Подписано в печать 17.04.2014 г.
Формат 60×84. Бумага офсет. Гарнитура Таймс.
Уч.-изд. л. 2,37. Тираж 30 экз. Заказ № .

Изд-во ТГАСУ, 634003, г. Томск, пл. Соляная, 2.
Отпечатано с оригинал-макета в ООП ТГАСУ.
634003, г. Томск, ул. Партизанская, 15.

ОГЛАВЛЕНИЕ

1. ПРЕДИСЛОВИЕ

Предлагаемые методические указания предназначены для самостоятельной работы студентов заочного факультета в процессе выполнения контрольной работы по теме «Математическая статистика». Математическое содержание данного раздела направлено на формирование у студента следующих общекультурных (ОК) и профессиональных компетенций (ПК):

(ОК-1): владение культурой мышления, способностью к обобщению, анализу, восприятию информации, постановке цели и выбору путей её достижения.

(ОК-6): стремление к саморазвитию, повышению своей квалификации и мастерства.

(ОК-9): способность к целенаправленному применению базовых знаний в области математических, естественных, гуманитарных и экономических наук в профессиональной деятельности.

(ОК-15): владение методами количественного анализа и моделирования, теоретического и экспериментального исследования.

(ПК-1): способность использовать законы и методы математики, естественных, гуманитарных и экономических наук при решении профессиональных задач.

(ПК-32): способность выбирать математические модели организационных систем, анализировать их адекватность, проводить адаптацию моделей к конкретным задачам.

В результате освоения материала студент должен:

Знать:	понятие случайной величины, её математическое описание, специфические свойства.
Уметь:	использовать статистические методы в обработке экспериментальных данных.
Владеть:	методами теории вероятностей и математической статистики.

2. ВВЕДЕНИЕ В ПРЕДМЕТ

Несмотря на большое разнообразие конкретных форм инженерной деятельности, центральное место в ней занимают процессы обработки данных, таких как сведения о ходе технологического процесса, результаты контроля выпускаемой продукции и так далее. Важнейшую роль при этом играет умение инженера выбрать соответствующий задаче математический аппарат и эффективно его использовать на практике.

В этой связи важно отметить, что все *реально* наблюдаемые явления и их показатели являются случайными по своей природе. Это значит, что ход развития таких явлений и их результат существенно зависит от множества факторов, часть которых не поддаётся учёту и контролю. Как результат - почти все наблюдаемые показатели обнаруживают характерный для случайных величин разброс значений даже в неизменных условиях испытания. Поэтому корректное изучение случайных явлений требует привлечения методов теории вероятностей и математической статистики.

Теория вероятностей и математическая статистика тесно взаимосвязаны, при этом каждая из этих дисциплин решает свойственные только ей задачи.

Теория вероятностей нацелена на разработку теоретико-вероятностных моделей, которые описывают свойства *гипотетического* (не конкретного) явления случайной природы. Объектом моделирования выступает закономерность, обнаруженная в ходе массовых наблюдений и имеющая обобщённый характер. Например, изучение природы ошибок измерения “натолкнуло” Гаусса на формулировку нормального закона распределения. Полученная таким образом теоретическая модель оказалась пригодной для описания закономерностей поведения широкого круга практически наблюдаемых случайных величин (например, время “жизни” прибора, курс акций, количество брака в партии изделий и так далее).

Многочисленные теоретические исследования, как природных явлений, так и технологических процессов привели к удивительному заключению: существует весьма ограниченный набор *модельных* законов распределения (нормальное, равномерное, показательное), которые наиболее часто встречаются в практике реальных стати-

стических наблюдений. Это так называемые “реалистические” модели распределений с чётким практическим смыслом. Ещё есть вспомогательные теоретические модели (распределения Стьюдента, Фишера, χ^2-распределение и ряд других), которые применяют в качестве технических инструментов в статистических методах (например, при проверке гипотез).

Модельные вероятностные пространства глубоко исследованы, а результаты в виде числовых характеристик оформлены стандартными таблицами, которые можно найти в приложениях почти всех учебников.

В противоположность теории вероятностей математическая статистика нацелена на изучение *реального*, а не гипотетического объекта. Его систематическое и целенаправленное наблюдение – стартовая задача исследователя. По результатам такой работы создаётся информационная база в виде *выборки*, которая служит основой для выявления статистически устойчивых закономерностей в “жизни” объекта наблюдений.

Основное предназначение математической статистики – обоснованный выбор среди множества возможных теоретико-вероятностных моделей той модели, которая наилучшим образом соответствует *реальным* статистическим данным, собранным исследователем на *конкретном* (а не гипотетическом) объекте изучения.

Важно учитывать, что статистические выводы всегда основаны на ограниченном, выборочном числе наблюдений и при увеличении числа наблюдений эти выводы могут оказаться иными. Поэтому для вынесения более определённого заключения о закономерностях изучаемого явления математическая статистика использует аппарат теории вероятностей (например, интервальное оценивание, статистическая проверка гипотез).

Таким образом, научный подход к исследованию природных явлений предполагает объединение теории (математической модели) и практики (статистических данных): мы используем теоретические модели для описания и объяснения наблюдаемых процессов и собираем статистические данные с целью обоснования и верификации моделей.

2.1. Необходимые сведения из теории вероятностей

Теория вероятностей – это математическая дисциплина, изучающая *закономерности массовых случайных явлений*. При этом теория вероятностей не может предсказать результатов отдельного опыта со случайными исходами, но она надежно предсказывает результат большого числа таких опытов.

Основные *объекты изучения* в теории вероятностей – *случайные события* и *случайные величины*. Следует понимать, что случайное событие – это качественная категория: событие либо происходит, либо нет. Например, получение предприятием прибыли. При анализе этого случайного события нас заинтересует размер прибыли, поэтому понятие случайного события дополняют понятием случайной величины.

Случайная величина, её описание

Случайной величиной называют величину, которая в результате наблюдения принимает то или иное значение, заранее не известное и зависящее от случайных обстоятельств. Примеры случайных величин: время “жизни” электрической лампочки, число попаданий в цель из серии выстрелов, процент брака в партии изделий и так далее.

Различают дискретные и непрерывные случайные величины.

Дискретной называется случайная величина, возможные значения которой образуют счетное множество (конечное или бесконечное).

Непрерывной называется случайная величина, возможные значения которой непрерывным образом заполняют некоторый конечный или бесконечный интервал числовой оси. Число значений непрерывной случайной величины всегда бесконечно. На практике, однако, встречаются дискретные случайные величины, число значений которых настолько велико, что их условно считают непрерывными (курсы валют, доход, ВВП).

От привычной для нас детерминированной величины случайная отличается тем, что каждому её значению соответствует определённая вероятность реализации.

Таким образом, для описания случайной величины необходимо установить соответствие между всеми возможными её значе-

ниями и их вероятностями. Такое соответствие называется з*аконом распределения* случайной величины. Его можно задавать в любой форме: табличной, аналитической, графической.

Пусть возможными значениями случайной величины X являются $x_1, x_2, \ldots, x_n$, вероятность реализации каждого из них обозначена как $p_1, p_2, \ldots, p_n$. Тогда закон распределения дискретной случайной величины X может быть записан в виде таблицы, называемой *рядом распределения* дискретной случайной величины:

X	x_1	x_2	…	x_n
P	p_1	p_2	…	p_n

Обычно $x_1 < x_2 < \ldots < x_n$. Обязательно $p_1 + p_2 + \ldots + p_n = 1$.

Универсальной формой закона распределения, пригодной как для дискретной, так и непрерывной случайной величины, является функция распределения:

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

то есть $F(x)$ есть вероятность того, что случайная величина X примет значение, меньшее, чем x. Иногда $F(x)$ называют *функцией накопленной вероятности* (рис. 1).

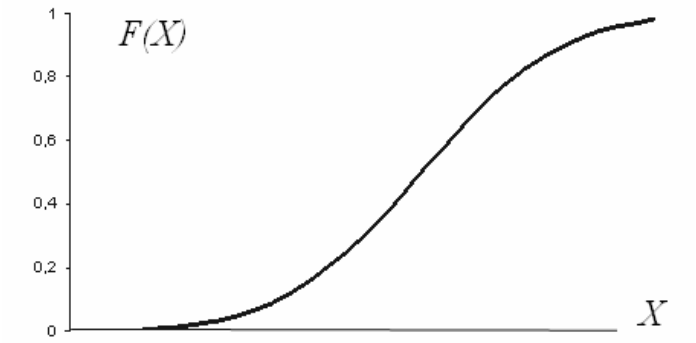

Рис. 1 Функция распределения непрерывной случайной величины

Прикладное назначение функции распределения связано с тем, что она определяет вероятность попадания случайной величины в заданный интервал значений:

$$P(\alpha \le X < \beta) = F(\beta) - F(\alpha).$$

Для непрерывных случайных величин наряду с функцией распределения используется еще одна форма задания закона распределения – плотность распределения.

Плотностью распределения $f(x)$ случайной величины X называется производная функции распределения $F(x)$:

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

Плотность распределения $f(x)$ характеризует вероятность попадания случайной величины в окрестность точки x. График плотности распределения $f(x)$ называют *кривой распределения*.

Числовые характеристики случайной величины

Закон распределения в виде функции распределения или функции плотности даёт полное описание случайной величины. Зачастую, однако, достаточно знать лишь некоторые характерные черты закона распределения, а не его полную форму. С этой целью в теории вероятностей используют числовые характеристики случайной величины, выражающие различные свойства закона распределения. Основными числовыми характеристиками являются математическое ожидание, дисперсия и среднее квадратическое отклонение.

Математическое ожидание $M(X)$ (или просто μ) – это некоторое среднее значение случайной величины, около которого группируются все её возможные значения.

Математическое ожидание для дискретной и непрерывной случайной величины, соответственно, вычисляется как

$$M(X) = \sum_{i=1}^{n} x_i p_i \ , \ M(X) = \int_{-\infty}^{+\infty} x f(x) dx.$$

Здесь x_i – отдельные значения (варианты) дискретной случайной величины X, p_i – вероятности, соответствующие значениям x_i , $f(x)$ – функция плотности распределения непрерывной случайной величины X.

Дисперсия $D(X)$ и *среднее квадратическое отклонение* $\sigma(X)$ – числовые характеристики случайной величины, которые отражают разброс её возможных значений относительно математического ожидания.

Для дискретной и непрерывной случайной величины, соответственно, дисперсия вычисляется как

$$D(X) = \sum_{i=1}^{n} (x_i - \mu)^2 p_i, \;\; D(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx.$$

Здесь μ - математическое ожидание для анализируемой случайной величины

Дисперсия имеет размерность квадрата случайной величины, это не всегда удобно. Поэтому в качестве характеристики рассеивания чаще используют среднее квадратическое отклонение, совпадающей по размерности со случайной величиной:

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

В заключение отметим, что подавляющее большинство используемых в статистических приложениях модельных законов распределения (биномиальный, нормальный, показательный…) могут быть однозначно восстановлены по одной-двум своим числовым характеристикам, чаще всего - по среднему значению и по дисперсии.

Законы распределений случайных величин.

Знание закона распределения случайной величины позволяет предсказать вероятность её попадания в интересующий нас интервал значений. Например, при анализе экономических показателей такие предсказания весьма желательны, так как позволяют выстраивать политику с учетом вероятности возникновения той или иной ситуации.

Прежде всего нас интересуют “реалистические” законы распределения вероятностей (нормальный, равномерный, показательный), которые имеют четкий практический смысл (доход семьи, вес единицы фасованного товара и так далее).

Нормальное распределение (распределение Гаусса) было открыто в начале 19 века Гауссом и Лапласом при изучении погрешности измерений.

Механизм порождения случайной величины в этом случае таков: непрерывная случайная величина X формируется под воздействием очень большого числа независимых случайных факторов, причем сила воздействия каждого отдельного фактора мала и не может доминировать среди остальных. При этом в результате воздействия случайного фактора F на величину X получается величина

$X + \Delta(F)$, где случайная "добавка" $\Delta(F)$ мала и равновероятна по знаку.

Случайная величина называется *нормально распределенной* или нормальной, если

$$f(x) = \frac{1}{2\pi\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{\frac{-(t-\mu)^2}{2\sigma^2}} dt .$$

Функция плотности $f(x)$ для нормального распределения имеет так называемую "колоколообразную" форму (рис. 2).

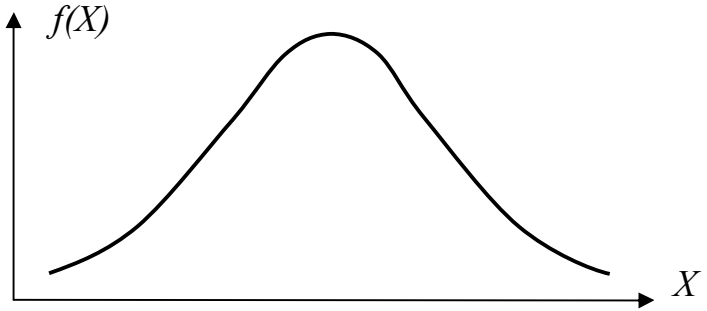

Рис. 2. Функции плотности $f(x)$ для нормального распределения

Нормальное распределение имеет два параметра: $\mu = M(X)$ и $\sigma = \sqrt{D(X)}$. В частности, при $\mu = 0$ и $\sigma = 1$ имеем *стандартное нормальное распределение*.

Полнота теоретических исследований, простота математических свойств и исключительная практическая значимость делают нормальный закон наиболее привлекательным и удобным в применении. Даже в случае отклонения исследуемых данных от нормального закона его можно использовать в качестве первого приближения – практика показала разумность такого подхода.

Кроме того, закон нормального распределения имеет большое теоретическое значение: с его помощью выведен целый ряд других важных распределений, таких как χ^2, Стьюдента, Фишера.

Равномерное распределение описывает непрерывные случайные величины, значения которых *равновероятны* внутри определенных границ (α, β). Например, ошибка округления числа равно-

мерно распределена на интервале от -5 до +5 единиц последнего (округляемого) разряда.

Формулы для плотности равномерного распределения *f(x)* и соответствующей ей функции распределения *F(x)* имеют вид:

$$f(x)=\begin{cases}\dfrac{1}{\beta-\alpha}, & \alpha<x<\beta\\ 0, & x\le\alpha,\quad x\ge\beta\end{cases},\quad F(x)=\begin{cases}0, & x\le\alpha,\\ \dfrac{x-\alpha}{\beta-\alpha}, & \alpha<x<\beta\\ 1, & x\ge\beta\end{cases}.$$

Функция плотности *f*(*x*) для равномерного распределения изображена на рис. 3.

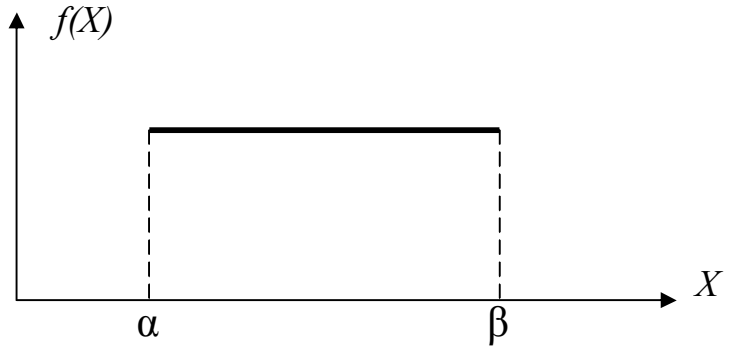

Рис. 3. Функция плотности *f*(*x*) для равномерного распределения

Основные числовые характеристики равномерного распределения таковы:

$$\mu=\frac{\alpha+\beta}{2},\quad \sigma=\frac{\beta-\alpha}{2\sqrt{3}}.$$

Равномерное распределение имеет два параметра – левая α и правая β границы промежутка распределения случайной величины.

Теоретическое значение равномерного распределения связано с использованием его в качестве “нулевого приближения” при описании некоторого нового распределения в отсутствии априорной информации о нем.

Экспоненциальное распределение моделирует процесс ожидания объекта, стоящего в очереди на различного рода обслуживание. В частности, это распределение моделирует “время жизни” некото-

рого механизма: чем меньше (ближе к нулю) время его эксплуатации, тем больше в среднем ожидаемое время его работы до поломки. Действительно, надёжнее покупать новый ($x = 0$) автомобиль, он позже сломается, чем подержанный.

Функция плотности и функция распределения в этом случае имеют вид:

$$f(x) = \begin{cases} 0, & x \le 0 \\ \lambda \cdot e^{-\lambda x}, & x > 0, \end{cases} \qquad F(x) = \begin{cases} 0, & x \le 0 \\ 1 - e^{-\lambda x}, & x > 0 \end{cases}.$$

Функция плотности $f(x)$ для экспоненциального распределения изображена на рис. 4.

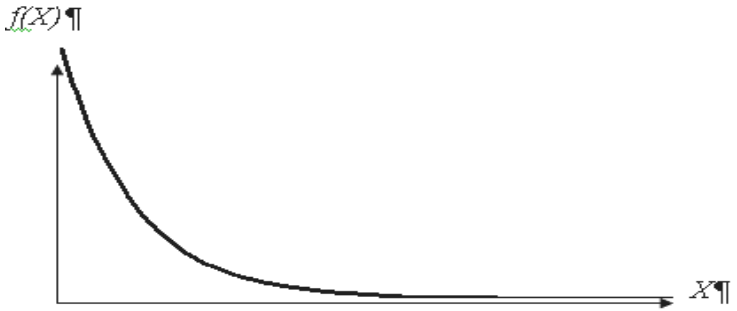

Рис. 4. Функция плотности $f(x)$ для экспоненциального распределения

Экспоненциальное распределение имеет один параметр λ, а основные числовые характеристики определяются соотношениями:

$$\mu = \frac{1}{\lambda}, \quad \sigma = \frac{1}{\lambda}.$$

В практических ситуациях, где используется экспоненциальный закон распределения, значение параметра λ либо известно, либо указан способ его вычисления.

Приведенные выше модельные законы распределения непрерывных случайных величин используются в практике статистических наблюдений для описания *реального* характера изменения как социально – экономических, так и технологических процессов.

2.2. Необходимые сведения из статистики

Аналитическая статистика предполагает использование выборочного метода, согласно которому реальный объект исследования X (*генеральная совокупность*) будет представлен при изучении набором его наблюдаемых значений $x_1, x_2, \ldots, x_n$ (*выборкой*). Действительно, оперативное обследование генеральной совокупности, численность которой бывает огромной, требует непомерных затрат (людских, материальных, финансовых). Поэтому для изучения свойств генеральной совокупности обследуют её часть – выборку, извлечённую *случайным образом* из генеральной совокупности. Случайный отбор предполагает равные возможности для элементов генеральной совокупности быть включёнными в выборку. Тогда (в силу закона больших чисел) выборка адекватно отображает структуру и свойства генеральной совокупности.

В практическом варианте под выборкой $x_1, x_2, \ldots, x_n$ понимают фактически полученные в данном конкретном эксперименте значения изучаемой случайной величины. Число элементов n в выборке называется ее *объемом.*

Вывод о свойствах генеральной совокупности, полученный на основании исследования выборки, называется *статистическим заключением.*

Таким образом, выборочный метод предполагает статистическую обработку полученных в эксперименте выборочных данных с целью теоретического описания свойств объекта изучения.

Способы представления статистических данных

Начальным этапом обработки выборочных данных должна быть систематизация беспорядочной массы чисел с целью предания ей удобной для последующего анализа формы.

При небольшом объеме выборки ($n < 20$) применяют довольно простой подход к систематизации данных – их упорядочивают, например, по неубыванию: $x_1 \le x_2 \le, \ldots, \le x_n$. В результате такой процедуры получают *ранжированный ряд*, анализ которого позволяет сделать содержательные статистические заключения. Так, в ранжированном ряду средний член определит центр группирования данных. Крайние в ранжированном ряду элементы ($x_1 = x_{min}$ и $x_n = x_{max}$)

дают сведения о диапазоне возможных значений случайной величины, разность (x_{max} - x_{min}) – о степени её случайного разброса. Разность между максимальным и минимальным значениями (x_{max} - x_{min}) называется *размахом выборки*.

При большом объёме выборки ($n > 50$ при изучении одномерной непрерывной величины или $n > 20$ для одномерной дискретной величины) ранжированный ряд становится громоздким и неудобным. Поэтому на начальном этапе выборку преобразуют к компактному виду. Наиболее эффективным способом обобщения и сжатия выборочных данных является построение *ряда распределения* (*вариационного ряда*). Формальной основой этой процедуры служит объединение в группы близких (или равных) по величине значений среди выборочных данных.

Группировку выборки в виде *частотного статистического* ряда выполняют для дискретной случайной величины, наблюдаемые значения которой содержат множественные повторы. А именно, из n наблюдаемых $x_1, x_2, \ldots, x_n$ различных значений (*вариант*) будет всего m ($m < n$). Тогда равные по величине значения разумно объединить в группы: для каждой k-ой варианты x_k получим группу численностью n_k, где $k = 1, 2, \ldots, m$.

При этом важно понимать, что численность одной группы может заметно отличаться от другой, то есть варианты x_k при группировке имеют разный “вес”.

Если значение x_k встретилось в выборке n_k раз, то целое число n_k называется *абсолютной частотой значения* x_k, а величина $\omega_k = n_k / n$ *относительной частотой значения* x_k. Тогда “вес” каждой варианты можно представить посредством соответствующей частоты (абсолютной или относительной).

В результате группировки выборочных данных по совпадающим значениям получим частотный статистический ряд:

X	x_1	x_2	…	x_m	
n_k	n_1	n_2	…	n_m	$\Sigma\, n_i = n$
$\omega_k = n_k/n$	ω_1	ω_2	…	ω_m	$\Sigma\, \omega_i = 1$

Статистический ряд изображается графически в виде *полигона частот*: по оси абсцисс откладывают значения вариант x_k, а по оси

ординат – частоты (абсолютные n_k или относительные ω_k). Полигон частот дает представление о *выборочной (эмпирической) функция плотности распределения* $f^*(x)$ для изучаемой случайной величины.

Также по статистическому ряду можно построить *эмпирическую функцию распределения* $F^*(x) = n_x / n$, где n_x – обозначение суммарного количества выборочных значений меньших, чем x; n – объем выборки.

В случае одномерной непрерывной случайной величины, которая представлена для изучения выборкой большого объёма ($n >$ 50), рекомендуется построение *равноинтервального статистического ряда*. Формальной основой этой процедуры служит объединение в группы близких по величине выборочных данных. Каждой группе соответствует свой диапазон значений (*интервал группировки*) и своя численность (*частота*)**.** Так как интервал группировки включает значения с незначительным отличием друг от друга, то их можно усреднить. Тогда среднее по интервалу значение выступает в качестве типичного элемента всей группы, численность которой отображает частота.

Следует отметить, что компактное представление выборочных данных в виде равноинтервального ряда будет адекватно отображать поведение изучаемой случайной величины X только при *слабой или умеренной вариации* значений $x_1, x_2,\ldots, x_n$, наблюдаемых для X в реальности.

Построение статистического ряда начинают с распределения выборочных значений $x_1, x_2,\ldots, x_n$ по m непересекающимся интервалам равной длины h (h – *шаг разбиения*). Предполагается, что выборка предварительно была упорядочена по возрастанию, то есть $x_1 = x_{min}$ и $x_n = x_{max}$. Тогда шаг разбиения h получим после деления размаха выборки, то есть ($x_{max} - x_{min}$) на m частей.

Число интервалов m изменяется в пределах 7 – 20 и существенно зависит от объёма выборки n. На практике при выборе m руководствуются оценкой: $m \approx log_2 n + 1$.

Далее подсчитывают n_k – количество выборочных значений, попадающих в каждый k-ый интервал. Тогда относительная частота для k-го интервала: $\omega_k = n_k / n$.

В результате из выборочных данных сформировано m групп (интервалов), каждая из которых содержит приблизительно равные

по величине элементы. “Вес” каждой группы (интервала) задаёт абсолютная частота n_k (или относительная частота ω_k). Это и есть интервальный статистический ряд, табличный вид которого:

$[x_k, x_{k+1})$	$[x_1 = x_{min}, x_1+h)$	$[x_1+h, x_1+2h)\ldots[x_1+(m\text{-}1)h, x_n = x_{max}]$
n_k	n_1	$n_2 \quad \ldots \quad n_m$
$\omega_k = n_k/n$	n_1/n	$n_2/n \quad \ldots \quad n_m/n$

Интервальный статистический ряд изображается графически в виде гистограммы: по оси абсцисс откладывают интервалы, на каждом из которых строится прямоугольник с высотой, равной ω_k. Гистограмма – это геометрическое изображение эмпирической функции плотности $f^*(x)$. Методологический смысл гистограммы состоит в том, что по её форме выдвигают гипотезу о виде распределения случайной величины X, руководствуясь при этом распределением её наблюдаемых значений $x_1, x_2, \ldots, x_n$ (выборки).

Эмпирическая функция распределения при интервальном представлении выборочных данных имеет вид: $F^*(x) = (n_1 + n_2 + \ldots + n_k)/n$, где k – порядковый номер интервала, правая граница которого накрывает x. Приведённая формула объясняет, почему эмпирическую функцию распределения $F^*(x)$ называют *накопленной относительной частотой*. График $F^*(x)$ называют *кумулятой* или *кумулятивной кривой*.

Итак, равноинтервальный ряд распределения служит для компактного представления выборочных данных и наиболее часто используется в практике статистических исследований. Сжатие исходной выборки достигается за счёт представления её посредством совокупности центров интервалов группировки и соответствующих им частот. Каждая частота отражает “вес” своего интервала относительно прочих интервалов.

В некоторых случаях, а именно, при значительной вариации изучаемого признака, рекомендуется группировка данных по интервалам с *изменяющейся* шириной. Типичным примером служит анализ уровня заработной платы в центральной России, когда Москва и Кострома присутствуют в пределах одной выборки.

Группировка выборочных данных значительно сокращает время и средства на их статистическую обработку. Именно поэтому она широко применяется на практике. Например, в процедуре голосования методом делегирования полномочий каждый объект статистической совокупности (город, штат, страна) голосует посредством своих представителей, число которых пропорционально размерам объекта. В этом примере каждый объект (город, штат, страна) – это интервал группировки, делегат – среднее значение по интервалу группировки. Численность делегации – это частота относительно интервала группировки. В итоге ограниченная численность делегатов "транслирует" широкое общественное мнение.

Заметим в заключение, что в результате группировки имеются некоторые потери информации: индивидуальность выборочных данных теряется после усреднения. При исследовании социальных явлений – это плохо, так как мнение каждого отдельного человека заменяют некоторым осреднённым по группе мнением «типичного представителя». Напротив, при обработке данных физического эксперимента усреднение измерений по интервалу кажется вполне целесообразным, так как оно в некоторой степени сглаживает неизбежные ошибки отдельных измерений.

Оценивание параметров генеральной совокупности

Статистическое заключение о генеральной совокупности подразумевает сведения о законе распределения исследуемой случайной величины и его параметрах. Реальные сведения такого характера получают на базе выборки ограниченного объёма, то есть в виде оценок для параметров генеральной совокупности.

Процесс нахождения оценок по определенному правилу (формуле) называют *оцениванием*. Различают два основных типа оценивания: для вида распределения и для параметров распределения.

В качестве оценки вида распределения (согласно закону больших чисел) можно взять выборочное распределение, рассчитанное, например, по интервальному статистическому ряду. Доказано, что выборочная (эмпирическая) функция распределения $F^*(x)$ и выборочная (эмпирическая) функция плотности $f^*(x)$ является несмещенными и состоятельными оценками своих теоретических

прототипов: функции распределения $F(X)$ и функции плотности распределения $f(X)$.

При изучении любой случайной величины X кроме определения её закона распределения желательно указать числовые характеристики: математическое ожидание, дисперсию, среднее квадратическое отклонение. В реальности для нахождения этих величин исследователь располагает только выборкой объёма n из наблюдаемых для X значений: $x_1, x_2, \ldots, x_n$.

Характеристики, вычисленные по выборке, условились называть *статистиками*, чтобы отличать их от аналогичных величин для генеральной совокупности - *параметров*.

Выборочное среднее (среднее арифметическое по выборке), *выборочная дисперсия, выборочное среднее квадратическое отклонение* определяются следующим образом:

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad D_n = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x}_n)^2, \quad \sigma_n = \sqrt{D_n}\ .$$

При задании выборки в виде статистического ряда для числовых характеристик имеем следующие формулы:

$$\bar{x}_n = \frac{1}{n}\sum_{k=1}^{m} n_k \cdot x_k, \quad D_n = \frac{1}{n}\sum_{k=1}^{m} n_k \cdot (x_k - \bar{x}_n)^2, \quad \sigma_n = \sqrt{D_n}\ .$$

Здесь m – число вариант в статистическом ряде, n – объём выборки.

При задании выборки в виде интервального ряда приведенные формулы будут содержать вместо x_k среднее значение по k-му интервалу: $x_k^* = (x_{k+1} + x_k)/2$, $k = 1,..,m\text{-}1$, m – число интервалов группировки.

Использование в приведённых формулах нижнего индекса n указывает на их зависимость от объёма выборки – разные по объёму выборки дают разные значения для числовых характеристик. В частности, для малых по объёму выборок формулу для выборочной дисперсии уточняют, заменив знаменатель n на n-1. В результате получают так называемую *исправленную* дисперсию. Отметим, что при $n > 30$ различие между выборочной и исправленной выборочной дисперсиями практически незначимо.

После формальных вычислений потребуется, очевидно, изучение свойств выборочных статистик, необходимое для заключения

о степени их соответствия параметрам генеральной совокупности. Соответствующие теоретические исследования были проведены и их результаты доказывают статистическую устойчивость основных выборочных статистик, то есть их сходимость к соответствующим генеральным параметрам.

Конкретно, перечислим основные статистики, которые служат оценками истинных параметров генеральной совокупности:

- эмпирическая функция распределения $F^*(x)$ – выборочная оценка $F(X) = P(X < x)$ теоретической функции распределения;
- эмпирическая функция плотности $f^*(x)$ - выборочная оценка $f(x)$ теоретической функции плотности распределения;
- эмпирическая относительная частота $\omega_k = n_k / n$ - выборочная оценка вероятности $P(X = x_k)$, то есть вероятности реализации k-го значения x_k случайной величины X;
- среднее арифметическое по выборке $\overline{x}_n$ - выборочная оценка математического ожидания $M(X)$ генеральной совокупности;
- исправленная дисперсия D_n – выборочная оценка генеральной дисперсии $D(X)$.

Методологическая связь теории вероятностей и статистики

Найденные по выборке статистики используют для формулировки заключения о генеральной совокупности, то есть о законе распределения исследуемого признака и его параметрах. При этом всегда существует риск ошибки ввиду неполноты информации. Действительно, результаты обследования части генеральной совокупности (то есть выборки) мы пытаемся обобщить на всю генеральную совокупность (реальный объект).

Отсюда возникает проблема количественной оценки степени риска в ситуации с неопределённостью. Используемый при этом инструмент хорошо известен – это метод проверки гипотез из области теории вероятностей.

Рассмотрим конкретную ситуацию. Пусть высказывается предположение о законе распределения исследуемого случайного показателя X (генеральной совокупности). На основе выборочных данных строится частотное распределение значений показателя

(например, в виде интервального ряда распределения). Возникает задача проверки гипотезы о том, что расхождение между предполагаемым (теоретическим) распределением и наблюдаемым (эмпирическим) распределением незначимо. Критерий, с помощью которого проверяется эта гипотеза, называется критерием согласия χ^2 или критерием Пирсона:

$$\chi^2 = \sum_{k=1}^{m} \frac{(n_k - l_k)^2}{l_k}.$$

Здесь n_k –наблюдаемая частота для k –го интервала группировки, l_k – отвечающая ей ожидаемая, теоретическая частота.

Статистика χ^2 сравнивает на согласие наблюдаемые и ожидаемые частоты, оценивая в итоге их суммарное расхождение по всем m интервалам группировки. Если это расхождение оказывается незначимым, то его можно считать случайным и гипотеза о виде распределения X принимается. При значимом, существенном расхождении ожидаемых и выборочных частот заключаем, что наблюдаемое распределение выборочных данных не согласуется с теоретическим распределением генеральной совокупности, откуда была взята выборка.

Отметим, что критерий согласия χ^2 наиболее часто используется на практике для проверки на согласованность наблюдаемого распределения с каким-либо конкретным распределением (равномерным, нормальным, показательным).

Резюме по вводной части

Рассмотренные выше базовые понятия теории вероятностей и математической статистики служат только “вехами ”, смысл которых – заострить внимание студентов на ключевых моментах в ходе изучения материала по серьёзным и объёмным учебникам. Приведённый в конце указаний библиографический список дает широкие возможности для выбора подходящего источника при изучении основного курса математической статистики в полном объёме.

3. ВОПРОСЫ ДЛЯ САМОПРОВЕРКИ

1. Что такое случайная величина? Приведите примеры.
2. Типы случайных величин. В чём разница между ними?

3. Что такое функция распределения?
4. Что такое функция плотности распределения?
5. Что такое закон распределения дискретной случайной величины?
6. Математическое ожидание. Его свойства, вероятностный смысл и вычисление.
7. Дисперсия, её свойства, вероятностный смысл и вычисление.
8. Среднее квадратическое отклонение, его вероятностный смысл.
9. Почему в качестве характеристики разброса значений случайной величины наряду с дисперсией используют и среднее квадратическое отклонение?
10. Как вычисляется вероятность попадания непрерывной случайной величины в заданный интервал значений?
11. Равномерное распределение: плотность и функция распределения, их графики, числовые характеристики.
12. Показательное распределение: плотность и функция распределения, их графики, числовые характеристики.
13. Нормальное распределение: плотность и функция распределения, их графики, числовые характеристики.
14. Почему нормальному закону распределения уделяется так много внимания?
15. Математическая статистика как наука.
16. Предмет математической статистики.
17. Методологическая взаимосвязь математической статистики и теории вероятностей.
18. Понятие генеральной и выборочной совокупности.
19. Сущность выборочного метода.
20. Требования, предъявляемые к выборке.
21. Что такое простой статистический ряд, вариационный ряд, варианта?
22. С какой целью и когда производится группировка выборочных данных?
23. Процедура равноинтервальной группировки выборочных данных.
24. Графическое представление сгруппированной выборки (полигон, гистограмма).

25. Вероятностные аналоги полигона и гистограммы.
26. С какой целью строится гистограмма, полигон?
27. Как можно выдвинуть предположение (гипотезу) о виде распределения?
28. Построение эмпирической функции распределения.
29. В чём состоит «близость» эмпирической и теоретической функций распределения.
30. Что такое параметры распределения? Примеры.
31. Постановка задачи оценки параметров распределения.
32. Что выступает в качестве точечных оценок параметров?
33. Выборочные числовые характеристики, формулы для их подсчёта.
34. Оценки параметров для равномерного, показательного, нормального распределений.
35. Что такое исправленное выборочное среднее квадратическое отклонение?
36. Как можно выдвинуть гипотезу о виде распределения?
37. Назначение критерия Пирсона.
38. На какой выборочной статистике основан критерий Пирсона?
39. Смысл статистики Пирсона.
40. Последовательность расчётов при проверке гипотезы по критерию Пирсона.

4. ПОСТАНОВКА ЗАДАЧИ И ИДЕЯ РЕШЕНИЯ

Выявление закона распределения случайной величины – одна из центральных задач теории вероятностей и математической статистики.

Во-первых, задача важна сама по себе, так как закон распределения даёт наиболее полную информацию об изучаемой случайной величине.

Во-вторых, исключительно важен частный случай этой задачи – проверка распределения на близость к нормальному закону. Получение ответа на этот вопрос является составной частью в решении ряда практически важных задач (например, при изучении статистических зависимостей между случайными переменными). Кро-

ме того, подавляющее большинство практически важных случайных показателей (прибыль предприятия, погрешность измерения…) подчиняются именно нормальному закону распределения.

В общем виде задача формулируется следующим образом.

Дано: выборка $x_1, x_2, \ldots, x_n$ из непрерывной генеральной совокупности.

Требуется: посредством статистической обработки выборки выдвинуть гипотезу о законе распределения изучаемой случайной величины. Оценить эту гипотезу на заданном уровне значимости.

Идея решения

1. Извлечь из выборочных данных $x_1, x_2, \ldots, x_n$ какую-нибудь характеристику, которая позволит выдвинуть гипотезу о виде распределения изучаемого показателя X.

Чаще всего используют выборочную функцию плотности, так как её графическое изображение (*гистограмма*) заметно отличается для разных распределений и потому может определить выбор соответствующей гипотезы.

2. Найти теоретический эталон для сравнения с ним выборочной, то есть практической характеристики из пункта 1.

Очевидно, что в качестве эталона следует рассматривать теоретическую функцию плотности для гипотетического (из пункта 1) вида распределения. Для обеспечения возможности последующего сравнения теоретической и выборочной функций плотности необходимо “свести” их в одну и ту же область изменения. А именно, если в качестве параметров теоретической функции плотности взять их выборочные оценки, то в результате получим “привязку” теоретической функции плотности к наблюдаемым в реальности данным.

3.Сравнить выборочную (из пункта 1) и отвечающую ей теоретическую (из пункта 2) функции. Если они соответствуют друг другу в пределах установленной погрешности (то есть на заданном уровне значимости), то выдвинутая в пункте 1 гипотеза о виде распределения случайного показателя X принимается, в противном случае гипотеза не принимается.

Изложенная выше идея составляет основу алгоритма решения поставленной задачи.

5. АЛГОРИТМ РЕШЕНИЯ

1. Группировка выборочных данных.

Для построения интервального статистического ряда распределим выборочные данные $x_1, x_2, \ldots, x_n$ по m непересекающимся интервалам равной длины.

1.1. Найдём x_{min} и $x_{max.}$

1.2. Вычислим размах выборки: $d = (x_{max} - x_{min})$.

1.3. Получим шаг разбиения $h = d / m$.

1.4. Найдём границы интервалов. Первый интервал: $[x_{\min}, x_{\min} + h = x_1]$, второй интервал будет: $(x_1, x_1 + h]$ и так далее. Здесь мы учли, что каждая внутренняя граница входит в два соседних интервала. Поэтому для определённости считаем, что каждый интервал содержит свою правую границу, то есть замкнут справа. В особом положении оказывается первый интервал, так как он замкнут и слева, и справа. Эта особенность будет учтена ниже, в пункте 4. Заметим, что замыкать интервалы можно было и слева – принципиально ничего не меняется, только особым (замкнутым с двух сторон) при этом станет последний интервал.

1.5. Определим центры интервалов: $x_k^* = (x_{k-1} + x_k)/2$, где x_{k-1} – левая, а x_k – правая границы k-ого интервала.

1.6. Для каждого интервала подсчитываем *абсолютные* частоты n_k –количество выборочных значений, попадающих в k-ый интервал.

1.7. Находим *относительные* частоты $W_k = n_k / n$, здесь объём выборки n.

1.8. Вычисляем *нормированные* частоты $H_k = W_k / h$, здесь h – длина интервала группировки.

По результатам выполнения перечисленных действий рекомендуется оформить таблицу с соответствующими графами.

Сжатие исходной выборки достигается за счёт представления её через совокупность значений центров интервалов и соответствующих им частот. Каждая частота отражает “вес” своего интервала по сравнению с любым другим.

2. Построение гистограммы. Формулировка гипотезы о виде распределения случайной величины.

Гистограмма – это графическое изображение интервалов группировки и соответствующих им частот. Вероятностный смысл гистограммы состоит в том, что по её виду можно "прикинуть" характер изменения функции плотности распределения. Поэтому именно гистограмма служит основанием для формулировки гипотезы о виде распределения исследуемой случайной величины. Кроме того, гистограмма позволяет выполнить быстрый визуальный анализ важных характеристик распределения: наибольшего и наименьшего значений, зон концентрации данных и так далее.

3. Числовые характеристики случайной величины.

В первую очередь это выборочное среднее и выборочная дисперсия, которые необходимы для определения параметров закона распределения для большинства практически важных случайных величин.

Отметим, что числовые характеристики случайной величины могут быть вычислены по выборке в двух вариантах: до и после её группировки. Понятно, что более точные значения даёт исходная, полная выборка. Её объём, однако, может оказаться весьма значительным (в реальности порядка тысячи и более), поэтому потребуется компьютер. Напротив, в "полевых" условиях либо в прикидочных расчетах можно опираться на сгруппированный аналог выборочных данных. Кстати, сравнение между собой величин, полученных в двух обозначенных подходах, даёт представление о погрешности вследствие группировки выборочных данных.

4. Построение теоретического закона распределения согласно выдвинутой в пункте 2 гипотезе.

Построение теоретической функции плотности обычно не вызывает затруднений: нужно взять из учебника соответствующую формулу для гипотетического закона распределения (показательного, равномерного, нормального) и подставить в неё числовые значения параметров распределения.

В качестве параметров теоретического закона распределения следует использовать их выборочные числовые оценки, найденные на предыдущем этапе. Тем самым достигается "при-

вязка" теоретического закона распределения к области выборочных, то есть реально наблюдённых значений случайной величины.

5. Формулировка статистического заключения.

Проверка гипотезы о виде распределения случайной величины может быть выполнена, например, с помощью критерия Пирсона χ^2. Статистика Пирсона фактически сопоставляет выборочную и теоретическую функции плотности распределения. Если отличие между ними не превышает заданной наперёд погрешности, то выдвинутая в пункте 2 гипотеза о виде распределения случайного показателя X принимается, в противном случае гипотеза не принимается.

6. ВАРИАНТЫ ИНДИВИДУАЛЬНЫХ ЗАДАНИЙ

В ходе освоения дисциплины студентами выполняется контрольная работа. Индивидуальные задания (табл. 1) представляют собой выборку из 80-ти наблюдаемых значений для непрерывной случайной величины. Каждая выборка занимает отдельный столбец, номер которого есть номер варианта.

Выбор варианта контрольной работы производится по последней цифре номера зачетной книжки. Так, если номер зачетной книжки заканчивается цифрой восемь, то Ваш вариант № 8, и так далее. Если номер зачетной книжки заканчивается нулём, то Ваш вариант № 10.

Таблица 1

1	2	3	4	5
7,705	10,515	6,693	11,604	6,339
8,947	7,971	8,038	9,018	5,407
9,14	10,315	7,424	11,013	7,61
9,923	8,484	9,322	10,573	6,266
5,703	7,245	11,056	10,283	8,13
9,166	5,751	6,311	8,944	6,505
9,551	7,026	9,082	7,605	5,802
4,600	8,334	9,629	8,762	6,763

1	2	3	4	5
6,946	6,753	10,799	8,057	6,474
8,882	9,924	8,858	7,980	6,57
7,921	6,137	9,168	9,444	7,614
5,533	5,822	9,861	12,557	5,925
9,097	7,754	10,855	7,322	7,276
5,770	9,422	10,565	12,886	7,22
5,704	6,117	6,795	12,006	7,586
4,532	5,217	10,234	12,915	5,452
4,998	9,002	6,875	8,754	7,648
6,064	8,193	10,282	12,501	5,327
4,255	10,468	7,846	8,447	6,686
7,482	10,602	8,072	8,876	7,104
6,389	10,995	10,394	11,322	6,106
5,425	10,624	6,903	7,274	7,868
6,044	10,919	7,313	12,281	5,323
6,415	10,804	11,983	8,756	8,967
4,065	8,406	8,433	7,491	7,955
6,876	7,822	9,958	10,334	6,413
4,428	5,757	6,979	12,200	6,332
7,757	8,176	7,245	9,008	7,152
4,821	10,325	10,92	11,517	6,12
9,619	7,061	11,008	9,370	6,32
9,795	5,894	11,508	7,860	6,034
9,757	6,373	10,947	8,737	6,184
7,633	5,493	8,801	11,454	7,12
6,774	6,955	10,404	11,106	7,575
7,159	6,841	11,931	10,308	6,373
4,135	5,448	9,941	11,259	5,766
8,614	6,520	6,354	8,536	7,004
4,657	5,940	11,882	8,999	7,379
6,936	6,675	6,152	7,883	7,099
6,574	5,969	10,219	8,870	6,16
7,726	5,227	9,985	10,693	5,857
4,585	8,042	11,992	12,396	6,309

1	2	3	4	5
9,710	6,064	7,185	10,602	7,55
6,931	7,514	7,992	11,271	6,887
6,388	8,414	10,139	11,909	5,642
4,312	7,880	6,432	7,964	5,214
6,690	9,276	7,495	8,645	6,319
7,004	8,150	7,588	10,152	5,36
4,419	9,371	7,106	12,839	7,306
6,937	5,586	10,386	11,819	7,808
5,736	8,034	11,031	7,941	6,482
9,343	6,219	10,447	9,399	6,578
6,415	8,815	7,893	11,313	6,843
5,721	5,292	6,618	10,903	7,257
9,59	5,800	9,614	8,920	6,852
4,372	7,140	7,317	11,275	4,716
6,574	10,832	11,677	8,139	7,332
9,003	8,460	6,808	11,915	7,087
7,460	5,111	6,873	8,440	6,84
7,415	7,305	8,100	7,561	7,467
9,985	6,103	6,128	8,269	7,179
8,001	9,766	9,711	9,899	5,469
9,142	10,183	8,723	11,937	5,685
9,244	10,796	11,791	10,908	6,957
5,412	10,407	6,491	12,195	7,942
7,849	10,188	10,981	8,855	7,618
9,801	8,388	10,277	10,995	6,841
8,728	5,772	8,698	8,997	6,424
9,593	5,195	7,819	9,331	7,983
9,344	10,456	8,330	12,837	7,993
7,388	6,761	7,502	11,502	6,233
5,299	9,962	7,084	12,433	6,163
6,735	6,935	11,473	11,127	6,604
7,289	6,285	11,444	7,303	5,307
9,735	6,765	10,679	11,489	7,327
6,113	9,075	9,127	11,385	6,512

1	2	3	4	5
6,782	5,996	8,672	11,300	6,751
9,322	9,395	8,915	8,412	7,511
5,453	10,330	7,081	9,983	8,73
4,164	6,250	10,593	10,795	5,665
6	**7**	**8**	**9**	**10**
0,132	-0,271	6,973	9,143	9,295
-0,825	0,551	9,994	6,838	9,012
0,055	-1,475	6,519	8,726	7,162
-1,009	-0,172	9,614	8,194	8,417
-1,914	1,444	7,292	9,887	4,108
-1,574	1,022	7,068	10,605	9,454
-0,068	2,647	9,121	8,796	3,305
1,238	-0,26	8,279	9,576	9,176
-0,474	-1,36	6,32	8,288	8,372
-1,734	2,48	8,035	7,181	7,525
0,238	-0,396	9,345	7,391	6,802
0,504	0,149	8,989	10,87	5,003
0,901	0,914	8,027	10,303	7,516
-0,475	0,144	9,521	7,249	4,910
1,522	-0,11	8,035	7,704	9,497
-0,424	0,954	8,919	7,91	6,532
-2,896	-1,087	8,008	10,538	6,268
1,773	2,613	8,876	9,694	9,494
1,563	-0,286	7,385	9,024	4,042
-0,48	-0,434	7,273	8,733	7,936
-2,021	1,005	7,656	9,791	5,721
-1,788	1,618	8,654	7,779	4,902
1,833	-1,945	8,236	6,607	8,750
1,145	0,724	6,189	9,742	7,444
1,43	-0,736	7,677	10,206	3,017
-1,516	1,301	8,387	8,451	6,567
-1,577	2,241	9,193	8,76	5,070
-0,36	-1,83	7,271	9,479	8,940
1,257	0,654	6,688	8,212	8,638

6	7	8	9	10
-3	0,003	8,316	9,529	8,557
1,3	-0,168	8,82	9,607	3,289
-2,005	1,967	8,738	8,609	9,738
1,622	2,58	9,488	8,897	6,133
1,817	-0,543	6,685	9,654	5,160
-1,3	1,612	8,771	9,797	9,702
0,808	-1,157	8,564	6,688	5,388
-0,129	2,603	8,296	9,497	6,209
-2,042	2,726	8,977	9,94	5,728
0,734	-0,687	8,856	9,225	9,664
0,871	2,346	7,106	7,939	5,482
-1,117	2,713	8,931	8,821	8,395
0,199	0,152	9,384	8,011	5,945
0,156	-0,03	8,579	9,607	3,071
1,918	0,683	6,312	9,299	6,398
0,674	2,166	9,315	9,38	5,528
1,107	1,38	6,635	8,429	8,073
1,343	-1,323	7,65	10,411	5,658
0,059	0,886	6,255	9,367	3,981
-1,732	-1,411	8,353	7,243	7,605
0,679	0,154	7,682	9,15	3,521
0,535	1,483	6,787	9,823	7,434
0,196	2,501	6,744	8,347	9,818
-1,567	-1,351	8,455	8,418	7,384
-0,134	1,73	11,177	10,454	6,379
-1,233	0,87	10,787	8,938	6,739
0,207	1,922	7,071	6,785	9,284
-1,61	-1,326	8,678	8,29	7,318
-0,723	-1,487	8,253	9,744	7,360
-2,08	0,598	7,223	8,539	9,923
1,257	-0,644	6,845	8,604	9,048
-1,339	-0,679	9,396	11,459	7,682
-1,31	1,987	7,744	7,266	8,801
-2,75	1,757	8,913	9,721	3,598

6	7	8	9	10
-2,711	-0,003	8,553	9,018	9,571
-0,312	-0,232	7,653	9,985	6,103
-0,473	0,479	7,61	9,816	3,133
-0,591	1,888	8,248	8,166	7,208
-0,226	-0,824	8,904	9,457	6,351
0,268	-0,984	6,92	9,084	7,265
1,555	-1,23	8,559	8,705	8,047
-1,777	0,35	8,616	7,469	8,432
-1,36	0,712	9,408	7,813	5,783
-0,772	0,554	8,005	9,366	7,924
0,986	2,695	7,097	8,391	9,407
0,474	1,559	6,885	8,563	5,438
1,108	1,779	8,059	9,343	4,975
-2,525	1,458	7,023	7,742	7,069
-2,945	-1,248	7,848	8,247	5,136
1,077	-1,85	7,514	9,744	5,213
-2,882	1,165	7,664	9,814	5,106

7. ПРИМЕР ВЫПОЛНЕНИЯ РАБОТЫ

Подобрать закон распределения для непрерывной случайной величины, которая представлена выборкой из 50 наблюдаемых значений (табл. 2).

Таблица 2

0,248	–1,487	–0,059	–0,078	0,194
1,701	0,072	0,318	0,899	–1,302
0,415	0,129	–1,088	0,318	–0,992
1,531	1,409	–0,267	–0,292	–0,056
–1,754	–0,934	0,512	0,085	–0,323
–1,382	–2,364	–1,085	0,367	0,529
–0,443	1,73	0,349	–0,882	0,757
–0,29	–0,449	–0,329	0,13	–0,882
–0,957	–0,095	–0,361	–0,54	0,229
–1,255	0,634	0,49	–1,088	1,028

Решаем задачу, следуя приведённому ранее алгоритму.

1. Выполним группировку выборочных данных. Для этого организуем таблицу со следующими графами (табл. 3):

В первой графе будем размещать номер интервала $k = 1, 2,\ldots, m$. Здесь m – общее число интервалов группировки. Для примера рекомендуем выбрать $m = 7$.

Далее, для того, чтобы найти границы интервалов группировки, выполним следующие действия:

– Находим $x_{\min} = -2{,}364$; $x_{\max} = 1{,}730$. Диапазон изменения выборочных данных (*размах выборки*): $d = x_{\max} - x_{\min} = 4{,}094$. Тогда длина каждого интервала группировки $h = d/m = 0{,}585$.

– Вычисляем границы интервалов. Первый интервал: $[x_{\min}, x_{\min} + h = x_1] = [-2{,}364; -1{,}779]$; второй интервал будет: $(x_1, x_1 + h] = (-1{,}779; -1{,}194]$ и так далее. Заносим полученные значения во вторую и третью графы таблицы.

Четвёртая графа - для центров интервалов: $x_k^* = (x_{k-1} + x_k)/2$. Здесь x_{k-1} – левая, а x_k – правая границы интервала.

Затем подсчитываем *абсолютные* частоты n_k – число выборочных данных, попадающих в k-ый интервал.

Находим *относительные* частоты $W_k = n_k / n$, n –объём выборки.

Наконец вычисляем *нормированные* частоты $H_k = W_k / h$, h – длина интервала группировки.

Выполнив перечисленные действия для нашего примера, получим следующую таблицу результатов (табл. 3).

Таблица 3

№ инт.	x_{k-1}	x_k	x_k^*	n_k	W_k	H_k
1	–2,364	–1,779	–2,072	1	0,02	0,034
2	–1,779	–1,194	–1,487	5	0,1	0,171
3	–1,194	–0,609	–0,902	8	0,16	0,274
4	–0,609	–0,025	–0,317	13	0,26	0,445
5	–0,025	0,56	0,268	15	0,3	0,513
6	0,56	1,145	0,853	4	0,08	0,137
7	1,145	1,73	1,438	4	0,08	0,137

Таким образом, выборка приведена в компактную форму: пятьдесят исходных значений $x_1, x_2, \ldots, x_{50}$ представлены всего 14 числами, а именно, для каждого k-го интервала ($k = 1,\ldots, 7$) это его центр x_k^* и нормированная частота H_k (если все интервалы равной длины, то вместо H_k можно использовать W_k).

В результате группировки данных упрощаются дальнейшие вычисления, особенно при больших объемах выборки.

2. Используя данные табл. 3, построим гистограмму. Для этого на оси OX отложим все 7 интервалов и на каждом из них построим прямоугольник с высотой n_k. (рис. 5).

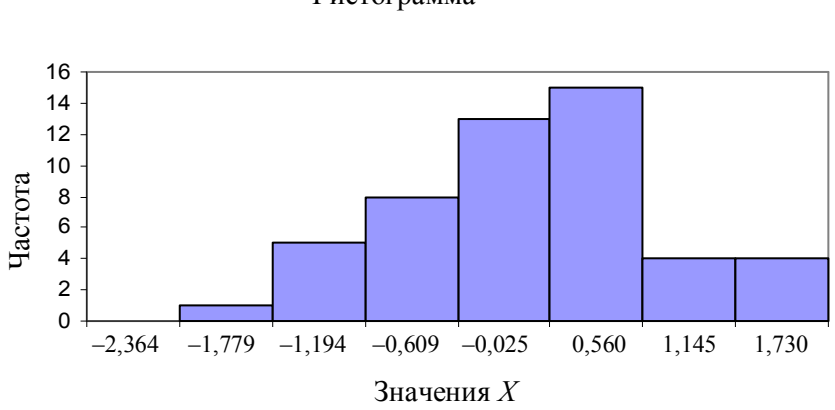

Рис. 5. Гистограмма по данным табл. 3

Сравним нашу гистограмму (рис. 5) с типовыми вариантами для показательного, равномерного и нормального законов распределения (рис. 6).

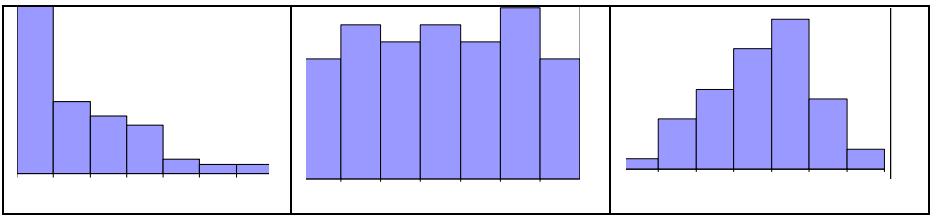

Рис. 6. Типовые варианты гистограмм

Гистограмма на рис. 5 имеет типичную *колоколообразную* форму, свойственную нормальному закону распределения.

Итак, по результатам сравнительного анализа выдвигаем гипотезу: ***представленная выборкой случайная величина подчиняется нормальному закону распределения.***

Заметим, что полученный нами вывод может быть не столь очевидным. В этом случае рекомендуем действовать методом исключения: 1) проверяем наличие явного "всплеска" наблюдаемых частот в начале гистограммы (первый вариант на рис. 6, отвечающий показательному распределению), если ответ "нет", тогда 2) "прикидываем" колоколообразную форму (то есть нормальный закон), если опять ответ отрицательный, то 3) предполагаем равномерный закон распределения.

3. Найдём числовые характеристики случайной величины по выборке и по сгруппированным данным.

Выборочное среднее:

$$\mu = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{50}\sum_{i=1}^{50} x_i = -0{,}139.$$

Среднее значение по сгруппированным данным:

$$\bar{x}_{gr} = \frac{1}{n}\sum_{k=1}^{m} x_k^* \cdot n_k = \frac{1}{50}\sum_{k=1}^{7} x_k^* \cdot n_k =$$

$$= (-2{,}072 - 1{,}4875 - 0{,}9028 - 0{,}31713 + 0{,}26815 +$$
$$+0{,}8534 + 1{,}4384)/50 = -0{,}153.$$

Относительная погрешность в вычислении среднего за счет замены выборки вариационным рядом:

$$\delta = \left|1 - \frac{\bar{x}_{gr}}{\bar{x}}\right| 100\ \% \approx 10\ \%.$$

Дисперсия по выборке и по сгруппированным данным, соответственно:

$$D = \frac{1}{50}\sum_{i=1}^{50}(x_i - \bar{x})^2 = 0{,}778,\ D_{gr} = \frac{1}{50}\sum_{k=1}^{7}(x_k^* - \bar{x})^2 n_k = 0{,}685.$$

Относительная погрешность в вычислении дисперсии при замене выборки вариационным рядом составляет около 12 %.

Исправленное среднее квадратическое отклонение:

$$\sigma = \sqrt{\frac{n}{n-1}D} = 0{,}891;\ \sigma_{gr} = \sqrt{\frac{n}{n-1}D_{gr}} = 0{,}836.$$

Относительно более точные варианты расчета выборочного среднего μ = -0,139 и среднего квадратического отклонения σ = 0,891 используем далее для оценки параметров теоретического закона распределения.

Действительно, большинство теоретических законов распределения могут быть однозначно восстановлены по одному или двум своим параметрам. А именно, показательное распределение характеризуется одним параметром:

$$\lambda = 1/\mu .$$

Равномерное распределение имеет два параметра:

$$a = \mu - \sqrt{3}\cdot\sigma,\ \ b = \mu + \sqrt{3}\cdot\sigma .$$

В нашем примере предполагается нормальное распределение, которое однозначно определяется двумя параметрами μ и σ. Их выборочные оценки были найдены выше и составляют μ = -0,139 и σ = 0,891.

4. Построение теоретического закона распределения согласно выдвинутой в пункте 2 гипотезе.

Приведём формулы для вычисления теоретической функции распределения *F(x)* для изучаемых в этом задании законов.

Для показательного распределения с параметром $\lambda = 1/\mu$:

$$F(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-\lambda x}, & x \ge 0 \end{cases}.$$

Для равномерного распределения с параметрами *a* и *b*:

$$F(x) = \begin{cases} 0, & x < a, \\ \dfrac{x-a}{b-a}, & a \le x \le b . \\ 1, & x > b \end{cases}$$

В нашем примере предполагается нормальное распределение, которое однозначно определяется двумя параметрами μ и σ. Тогда, используя для них найденные ранее (пункт 3) оценки, теоретическая

функция распределения в области наблюдаемых значений случайной величины будет иметь следующий вид:

$$F(x) = 0{,}5 + \Phi\left(\frac{x-\mu}{\sigma}\right).$$

$\Phi(z)$ – функция Лапласа. Известно, что она нечётная (то есть $\Phi(-z) = -\Phi(z)$), а для значений $z > 4$ имеем $\Phi(z) \approx 0{,}5$. Стандартную таблицу значений $\Phi(z)$ можно найти в любой книге по теории вероятностей и в приложении 1 данных указаний.

Обратите внимание, здесь для удобства введена новая переменная z, которая легко вычисляется по имеющимся данным:

$$z = \frac{x-\mu}{\sigma}.$$

Таким образом, *перед* использованием таблицы значений $\Phi(z)$ нужно предварительно "заготовить" её аргумент z во всех точках x из области изменения случайной величины, то есть от $x_{\min}$ до $x_{\max}$.

При вычислении $F(x)$ по сгруппированным данным (смотри табл. 3) в качестве значений аргумента x следует брать правую или левую границу каждого интервала. Выберем для определённости правую, то есть первый интервал будет представлять его правая граница $x_1 = x_{\min} + h$, второй интервал – $x_2 = x_{\min} + 2h$ и так далее вплоть до $x_7 = x_{\max}$. В результате обнаруживаем, что минимальное значение $x_{\min}$ выпадает из рассмотрения. Для восполнения потери введём дополнительный (фиктивный) нулевой интервал, который обеспечит вычисление $F(x)$ в точке $x_{\min}$. (замена правой границы на левую не решает проблему, так как в этом случае из рассмотрения выпадет $x_{\max}$).

С учётом сделанных замечаний расчет теоретической функции для нормального распределения оформлен в виде таблицы (табл. 4): каждый интервал (его номер содержит первая графа) представлен своей правой границей (вторая графа), соответствующий ей аргумент функции Лапласа z вычислен (при μ = -0,139 и σ = 0,891) и приведём в третьей графе. Далее по таблице из приложения 1 выписаны значения $\Phi(z)$ и, наконец, в последней графе

приведены значения теоретической функции распределения, найденные по формуле $F(x) = 0{,}5 + \Phi(z)$.

Таблица 4

№ инт.	Правая граница	z	$\Phi(z)$	$F(x)$
0	–2,364	-2,496	-0,494	0,007
1	–1,779	-1,840	-0,467	0,033
2	–1,194	-1,184	-0,381	0,119
3	–0,609	-0,528	-0,200	0,300
4	–0,025	0,129	0,052	0,552
5	0,56	0,785	0,282	0,782
6	1,145	1,441	0,425	0,925
7	1,73	2,097	0,482	0,982

Напомним, что найденная здесь теоретическая функция распределения $F(x)$ будет использована далее при оценке гипотезы о нормальном распределении случайной величины.

5. Проверку гипотезы о виде распределения случайной величины можно выполнить с помощью χ^2-критерия Пирсона:

$$\chi^2 = \sum_{k=1}^{m} \frac{(n_k - l_k)^2}{l_k}.$$

Здесь n_k – выборочная частота для k –го интервала, l_k – теоретическая частота для k –го интервала, m – общее число интервалов группировки (в нашем примере $m = 7$).

Таким образом, формально критерий Пирсона сопоставляет наблюдаемые и предполагаемые (гипотетические) частоты по всем интервалам группировки. Если суммарное отличие этих величин не превышает заданной погрешности, то выдвинутая в пункте 2 гипотеза о виде распределения случайной величины принимается, в противном случае гипотеза не принимается.

Преступая к вычислению наблюдаемого значения χ^2, отметим, что выборочные частоты (n_k) уже есть, их можно взять из табл. 3. А вот их теоретический аналог (l_k) подлежит определению. С этой целью, очевидно, следует использовать теоретический закон распределения.

Действительно, прикладной смысл функции распределения $F(x)$ состоит в том, что она определяет вероятность попадания случайной величины в заданный интервал значений: $P(a \le X \le b) = F(b) - F(a)$. Следовательно, теоретическая вероятность P_k попадания значений случайной величины в каждый k-ый интервал: $P_k = F_k - F_{k-1}$. Отсюда нужная нам теоретическая частота для k –го интервала: $l_k = P_k \cdot n$ (n – объём выборки).

Соответствующие расчеты оформлены в виде табл. 1.5. В последней графе приведены слагаемые статистики Пирсона для каждого из 7 интервалов.

Таблица 5

№ инт.	Теоретическая вероятность	Теоретическая частота	Статистика Пирсона
1	0,027	1,325	0,080
2	0,086	4,300	0,114
3	0,181	9,050	0,122
7	0,252	12,600	0,013
5	0,230	11,500	1,065
6	0,143	7,150	1,388
7	0,057	2,850	0,464
Суммарная величина статистики Пирсона			**3,245**
Критическое значение статистики Пирсона			**9,488**

Две последние строки содержат показатели, необходимые при формулировке статистического заключения. А именно, сумма значений из последнего столбца даёт *наблюдаемое значение статистики Пирсона*: $\chi^2 = 3{,}245$.

С другой стороны, по таблице критических значений для χ^2-распределения найдём предельную величину $\chi^2_*(\alpha, \nu)$. Здесь α - заданный по условию задачи уровень значимости, обычно берут $\alpha = 0{,}05$ (то есть при обобщении свойств выборки на генеральную совокупность вероятность ошибки не выше 5%). Вторая переменная $\nu = m - r - 1$ – число степеней свободы, здесь m – количество интервалов группировки (в нашем примере m = 7), величина r представляет число параметров в предполагаемом законе распре-

деления. Так, для показательного закона r =1, тогда [1 - 3]: $\chi^2_*(\alpha,\nu) = \chi^2(0{,}05;5) = 11{,}071$.

Для равномерного и нормального законов (то есть в нашем примере) имеем $r = 2$, тогда $\chi^2_*(\alpha,\nu) = \chi^2(0{,}05;4) = 9{,}488$.

***Статистическое заключение*:** так как наблюдаемое значение статистики Пирсона меньше критического, то гипотеза о нормальном распределении случайной величины, представленной выборкой из наблюдаемых значений, принимается на заданном уровне значимости**.**

8. БИБЛИОГРАФИЧЕСКИЙ СПИСОК

Основная литература

1. Гмурман, В.Е. Теория вероятностей и математическая статистика / В.Е Гмурман. – М. : Высшая школа, 2011. – 480 с.
2. Гмурман, В.Е. Руководство к решению задач по теории вероятностей и математической статистике / В.Е Гмурман. – М. : Высшая школа, 2011. – 400 с.
3. Кремер, Н.Ш. Теория вероятностей / Н.Ш. Кремер. – М. : Высшая школа, 2011. – 174 с.
4. Дорофеева, Н.С. Первичная обработка выборочных данных / Н.С. Дорофеева. - Томск : Офсетная лаборатория ТГАСУ, 2009. – 61 с.
5. Просветов, Г.И. Теория вероятностей и математическая статистика. Задачи и решения / Г.И. Просветов. – М. : Альфа Пресс, 2009. – 272 с.

Дополнительная литература

6. Вентцель, Е.С. Теория вероятностей / Е.С. Вентцель. – М. : Высшая школа, 2006. – 575 с.
7. Слободской, М.И. Теория вероятностей и математическая статистика / М.И. Слободской. – Томск : Офсетная лаборатория ТГАСУ, 2001. – 87 с.
8. Протасов, К.В. Статистический анализ экспериментальных данных / К.В. Протасов. – М. : Мир, 2005. – 142 с.

ПРИЛОЖЕНИЕ 1

Таблица значений функции Лапласа $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{\frac{-z^2}{2}} dz$

x	0	1	2	3	4	5	6	7	8	9
0,0	0,00000	00399	00798	01197	01595	01994	02392	02790	03188	03586
0,1	03983	04380	04776	05172	05567	05962	06356	06749	07142	07535
0,2	07926	08317	08706	09095	09483	09871	10257	10642	11026	11409
0,3	11791	12172	12552	12930	13307	13683	14058	14431	14803	15173
0,4	15542	15910	16276	16640	17003	17364	17724	18082	18439	18793
0,5	19146	19497	19847	20194	20540	20884	21226	21566	21904	22240
0,6	22575	22907	23237	23565	23891	24215	24537	24857	25175	25490
0,7	25804	26115	26424	26730	27035	27337	27637	27935	28230	28524
0,8	28814	29103	29389	29673	29955	30234	30511	30785	31057	31327
0,9	31594	31859	32121	32381	32639	32894	33147	33398	33646	33891
1,0	34134	34375	34614	34850	35083	35314	35543	35769	35993	36214
1,1	36433	36650	36864	37076	37286	37493	37698	37900	38100	38298
1,2	38493	38686	38877	39065	39251	39435	39617	39796	39973	40147
1,3	40320	40490	40658	40824	40988	41149	41309	41466	41621	41774
1,4	41924	42073	42220	42364	42507	42647	42786	42922	43056	43189
1,5	43319	43448	43574	43699	43822	43943	44062	44179	44295	44408
1,6	44520	44630	44738	44845	44950	45053	45154	45254	45352	45449
1,7	45543	45637	45728	45818	45907	45994	46080	46164	46246	46327
1,8	46407	46485	46562	46638	46712	46784	46856	46926	46995	47062
1,9	47128	47193	47257	47320	47381	47441	47500	47558	47615	47670
2,0	47725	47778	47831	47882	47932	47982	48030	48077	48124	48169
2,1	48214	48257	48300	48341	48382	48422	48461	48500	48537	48574
2,2	48610	48645	48679	48713	48745	48778	48809	48840	48870	48899
2,3	48928	48956	48983	49010	49036	49061	49086	49111	49134	49158
2,4	49180	49202	49224	49245	49266	49286	49305	49324	49343	49361
2,5	49379	49396	49413	49430	49446	49461	49477	49492	49506	49520
2,6	49534	49547	49560	49573	49585	49598	49609	49621	49632	49643
2,7	49653	49664	49674	49683	49693	49702	49711	49720	49728	49736
2,8	49744	49752	49760	49767	49774	49781	49788	49795	49801	49807
2,9	49813	49819	49825	49831	49836	49841	49846	49851	49856	49861
3,0	0,49865		3,1	49903	3,2	49931	3,3	49952	3,4	49966
3,5	49977		3,6	49984	3,7	49989	3,8	49993	3,9	49995
4,0	499968									
4,5	499997		–	–	–	–	–	–	–	–
5,0	49999997									