МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Тульский государственный университет»
Кафедра информационной безопасности

УТВЕРЖДАЮ
Зав. кафедрой ИБ
______________ А.А. Сычугов
«___» ______________ 20___ г.

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

к курсовой работе по дисциплине

«Теория вероятностей и математическая статистика»

на тему

«Случайные величины. Распределение Стьюдента (t-распределение)»

Автор работы ____________________ студент гр. 230711 Павлова В.С.
(дата, подпись) (фамилия и инициалы)

Руководитель работы ____________________ ______________ ______________
(дата, подпись) (должность) (фамилия и инициалы)

Работа защищена ____________________ с оценкой ______________
(дата)

Члены комиссии ____________________ ______________ ______________
(дата, подпись) (должность) (фамилия и инициалы)

____________________ ______________ ______________
(дата, подпись) (должность) (фамилия и инициалы)

____________________ ______________ ______________
(дата, подпись) (должность) (фамилия и инициалы)

Тула 2023

ЗАДАНИЕ

на курсовую работу по дисциплине

«Теория вероятностей и математическая статистика»

студента гр. 230711 Павловой Виктории Сергеевны

Тема курсовой работы

«Случайные величины. Распределение Стьюдента (t-распределение)»

Исходные данные

Задание получил ______________________ ______________

(ФИО) (подпись)

Задание выдал ______________________ ______________

(ФИО) (подпись)

Дата выдачи задания 21.02.2023 г.

График выполнения КР 21.02-28.02 – Получение и ознакомление с заданием

01.03-22.03 – Изучение литературы и других исходных материалов

23.03-03.05 – Изучение теории, раскрывающей тему курсовой работы

04.05-17.05 – Реализация практической части курсовой работы

18.05-24.05 – Анализ результатов

25.05-07.06 – Оформление пояснительной записки и сдача на проверку

28.06.2023 – Защита курсовой работы

Рекомендации и особые отметки ______________________

«__» ________20__г

Содержание

Введение

Математическая дисциплина, которая изучает случайные явления и вероятности их возникновения, называется теорией вероятностей. Она имеет широкий спектр применений, начиная от криптографии и финансовой аналитики и заканчивая разработками в сфере искусственного интеллекта. История теории вероятностей уходит в далекое прошлое, и её основоположником считается математик Блез Паскаль. В 1654 году он начал свои исследования вероятностей, решая задачу о разделе выигрыша в азартной игре.

За несколько сотен лет теория вероятностей развилась в самостоятельную научную дисциплину, строго формализованную и включающую в себя различные математические методы, теории и анализ случайных процессов. Идеи о понятии распределения вероятностей возникли ещё в 17-ом веке у Пьера де Ферма и Блеза Паскаля, но основополагающую работу в этой области сделал Абрахам де Муавр в 18-ом веке, опубликовав 1733 году свою работу под названием «Доклад о шансах», в которой впервые применил анализ и статистику для описания случайных событий и введения понятия распределения вероятностей. Он показал, как можно использовать биномиальное распределение для приближения вероятности великого числа независимых случайных испытаний.

Окончательно теория вероятностей как математическая наука оформилась в 30-х годах 20-го века, когда Андреем Николаевичем Колмогоровым была предложено аксиоматическое определение вероятности. Колмогоров также формализовал понятие случайной величины и установил строгие математические основы для изучения вероятностных явлений. С помощью теории вероятностей стало возможным описывать случайные величины и их распределения.

Одно из наиболее изучаемых распределений случайных величин в теории вероятностей – нормальное распределение, также известное как распределение

Гаусса, названное в честь математика Карла Фридриха Гаусса, который изучил его основные свойства в 18-ом веке, например то, что нормальное распределение подходит для моделирования большого количества случайных процессов. В частном случае, когда известна только выборка, но неизвестна генеральная совокупность, такую модификацию нормального распределения называют распределением Стьюдента или t-распределением.

Темой данной курсовой работы является распределение Стьюдента, тесно связанное с нормальным распределением. Целью работы является изучение свойств этого распределения и решение практических задач, с ним связанных. Актуальность исследования распределение Стьюдента обоснована тем, что оно более устойчиво к выбросам и отклонениям от среднего значения в сравнении с нормальным, а значит может более эффективно использоваться в качестве инструмента для моделирования и анализа случайных процессов в теории вероятностей и математической статистике.

1 Случайная величина и её характеристики

1.1 Одномерная случайная величина

Понятие случайной величины играет важную роль в теории вероятностей и является ключевым инструментом, позволяющим формально описывать случайные явления и их вероятностные свойства. ***Одномерная случайная величина*** — это скалярная функция, определенная на пространстве элементарных событий, сопоставляющая каждому элементарному событию число или, иначе говоря, это величина, значение которой зависит от того, каким элементарным событием закончился случайный опыт [1]. Случайные величины обозначаются заглавными буквами латинского алфавита, например X, Y, Z. Среди общепринятых обозначений можно выделить также N, которое часто используется для обозначения случайной величины, имеющей нормальное (гауссово) распределение, и T, которое используется для обозначения распределения Стьюдента (t-распределение).

Под ***элементарным исходом*** или элементарным событием ω понимается любой простейший, то есть неразделимый в контексте рассматриваемого опыта, исход. Множество всех возможных значений или результатов случайного эксперимента называют ***пространством элементарных исходов*** и обозначают символом Ω. Иначе говоря, множество исходов опыта образует пространство элементарных исходов, если верно следующее [2]:

- в результате опыта один из исходов обязательно происходит;
- появление одного из исходов исключает появление всех прочих;
- в рамках данного опыта элементарных исход неразделим.

Более строго данную связь можно выразить следующим образом: $\omega \in \Omega$, а множество $\Omega = \{\omega_1, \omega_2, \dots, \omega_n, \dots\} = \{\omega_i,\ i = 1,2, \dots, n, \dots\}$. Чтобы задать некоторую случайную величину X, необходимо поставить в соответствие каждому элементарному исходу ω_i числовое значение x_i, которое случайная величина примет, если в результате опыта произойдет именно этот исход.

Например, если каждому элементарному исходу ω_i, $i = 1, \dots, n$, ставится в соответствие число i, то получим случайную величину, определяемую следующим образом (таблица 1):

Таблица 1 – Соответствие между элементарными исходами ω и значениями X

ω	ω_1	ω_2	$\dots$	ω_n
X	1	2	$\dots$	n

Таким образом, скалярную функцию $X(\omega)$, заданную на пространстве элементарных исходов $\Omega = \{\omega_1, \omega_2, \dots, \omega_n, \dots\}$, и называют случайной величиной, если для любого $x \in \mathrm{R}$ множество $\{\omega\colon \mathrm{X}(\omega) < x\}$ элементарных исходов, удовлетворяющих условию $\mathrm{X}(\omega) < x$, является событием.

Случайные величины могут быть разделены на ***дискретные***, то есть такие, которые принимают набор отдельных изолированных значений (конечное (счётное) или бесконечное число), или ***непрерывные***, набор значений которых целиком заполняет некоторый интервал [3].

Основной характеристикой случайных величин является ***закон распределения случайной величины***, также называемый распределением её вероятностей. Общим законом распределения и для дискретных, и для непрерывных случайных величин является ***функция распределения*** – функция $F(x)$ для случайной величины X, значение которой в точке x равно вероятности события $\{X < x\}$, то есть события, состоящего из тех и только тех элементарных исходов ω, для которых $X(\omega) < x$:

$$F(x) = P\,\{X < x\}$$

Типичный вид функции распределения приведён на рисунке 2.

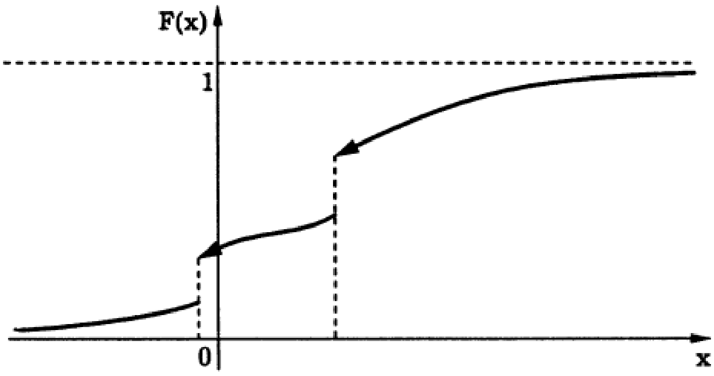


Рисунок 2 – Вид графика функции распределения

Из того факта, что функция распределения – это вероятность, следует, что она обладает следующими свойствами:

1) Функция распределения – это неубывающая функция: $F(x_1) < F(x_2)$ при $x_1 < x_2$;
2) Множество значений принадлежит отрезку [0, 1], откуда следует, что $F(-\infty) = 0, F(+\infty) = 1$;
3) Вероятность попадания в интервал (x_1, x_2) можно определить следующим образом: $P\{x_1 \leq X \leq x_2\} = F(x_2) - F(x_1)$;

Законом распределения дискретной случайной величины X называют таблицу, состоящую из двух строк, в одной из которых перечислены все возможные значения случайной величины, а во второй – вероятности p_i, что случайная величина X примет эти значения, то есть $p_i = P\{X = x_i\}$. При этом $\sum_{i=1}^{n} p_i = 1$, то есть сумма по строке вероятностей должна равняться единице [2]. Общий вид закона распределения дискретной случайной величины приведён в таблице 2.

Таблица 2 – Закон распределения дискретной случайной величины в общем виде

X	x_1	x_2	$\dots$	x_i	$\dots$	x_n
P	p_1	p_2	$\dots$	p_i	$\dots$	p_n

График функции распределения дискретной случайной величины представлен на рисунке 3. График представляет собой ступенчатую линию, где каждая ступень соответствует возможному значению случайной величины, а её высота (высота скачка в данной точке) определяет вероятность возникновения этого значения.

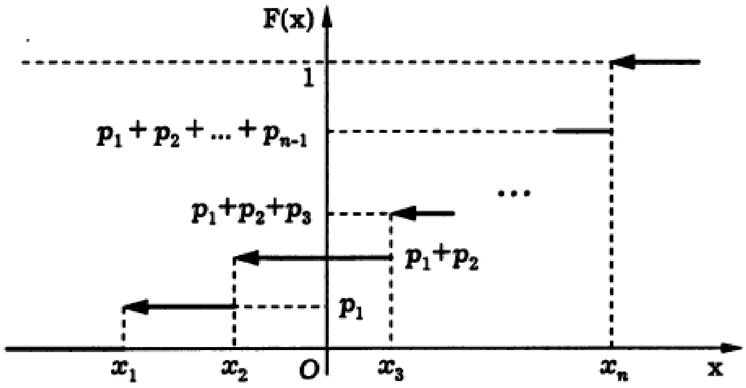


Рисунок 3 – График функции распределения дискретной случайной величины

Для непрерывной случайной величины X функцию распределения $F(x)$ можно выразить с помощью ***функции плотности распределения*** $f(x)$ следующим образом:

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

Из этого следует, что функцию $f(x)$ можно называть дифференциальным законом распределения непрерывной случайной величины, а $F(x)$ – интегральным, причём для этих законов верно следующее:

$$F(x) = \int_{-\infty}^{x} f(x)dx$$

Функция плотности обладает следующими свойствами:

1) $f(x) \geq 0$, поскольку $F(x)$ не убывает;
2) $\int_{-\infty}^{+\infty} f(x)dx = 1$;
3) Вероятность попадания в интервал (x_1, x_2) можно определить следующим образом $P\{x_1 \leq X < x_2\} = \int_{x_1,}^{x_2,} f(x)dx$;
4) $P\{X = x\} = 0$.

График функции плотности распределения приведён на рисунке 4. Согласно свойству 3, площадь криволинейной трапеции, обозначенной на рисунке штриховкой, равна вероятности попадания непрерывной случайной величины в полуинтервал $[x_1, x_2)$. Аналогичным образом в силу свойства 2 площадь под всей кривой распределения равна единице.

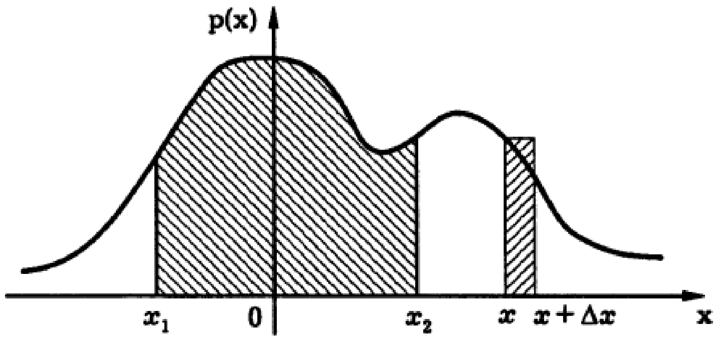


Рисунок 4 – График функции плотности распределения непрерывной случайной величины

1.2 Основные виды распределения случайных величин

Среди наиболее часто встречающихся на практике распределений можно выделить следующие [2]:

1. ***Биномиальное распределение*** дискретной случайной величины, задаваемое следующим образом: $P_n(i) = C_n^i p^i q^{n-i} = (p+q)^n$, где $i = 0, 1, \ldots n$. Его закон распределения приведён в таблице 3.

Таблица 3 – Ряд биномиального распределения

X	0	1	…	i	…	n
P	q^n	$C_n^1 p q^{n-1}$	…	$C_n^i p^i q^{n-i}$	…	p^n

2. ***Распределение Пуассона*** дискретной случайной величины, определяемое следующим образом: $P\{i, \lambda\} = \frac{\lambda^i}{i!} e^{-\lambda}$, где $i = 0, 1, \ldots n, \ldots$,

а $\lambda > 0$ – параметр распределения Пуассона. Его закон распределения приведён в таблице 4.

Таблица 4 – Ряд распределения Пуассона

X	0	1	2	…	n	…
P	$e^{-\lambda}$	$\lambda e^{-\lambda}$	$\frac{\lambda^2}{2}e^{-\lambda}$	…	$\frac{\lambda^n}{n!}e^{-\lambda}$	…

3. ***Равномерное распределение*** непрерывной случайной величины, плотность распределения которой определяется выражением:

$$f(x) = \begin{cases} \frac{1}{b-a}, x \in [a,b] \\ 0, \quad x \notin [a,b] \end{cases}$$

Исходя из того факта, что $F(x) = \int_{-\infty}^{x} f(x)dx$, закон распределения равномерного распределения записывается следующим образом:

$$F(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & x \in [a,b] \\ 1, & x > b \end{cases}$$

Графики обеих функций представлены на рисунке 5.

А) Б)

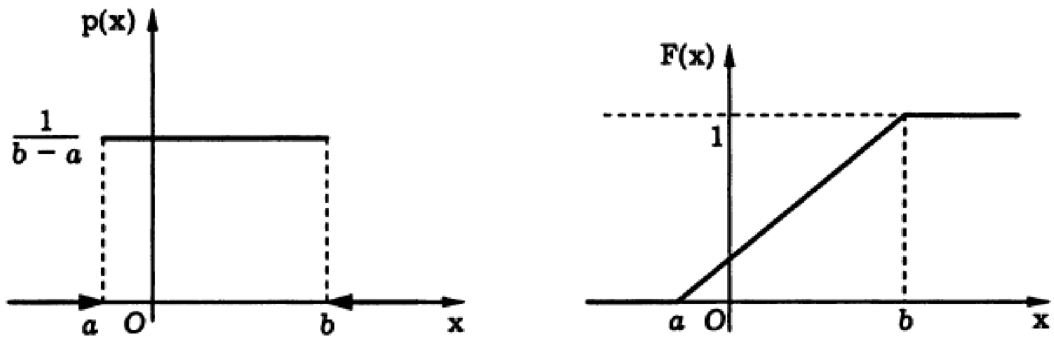


Рисунок 5 – График функции: а) плотности равномерного распределения; б) закона равномерного распределения

4. ***Экспоненциальное распределение*** непрерывной случайной величины, функция плотности распределения которого имеет параметр λ и определяется как:

$$f(x) = \begin{cases} 0, & x < 0 \\ \lambda e^{-\lambda x}, & x \geq 0 \end{cases}$$

Случайная величина, распределённая по экспоненциальному (показательному) закону, имеет функцию распределения вида:

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

Графики данных функций представлены на рисунке 6.

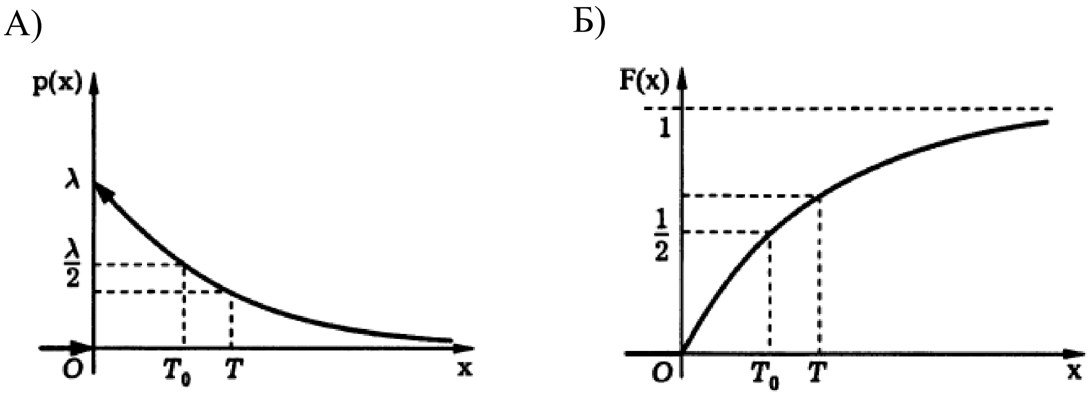


Рисунок 6 – График функции: а) плотности показательного распределения; б) закона показательного распределения

5. ***Нормальное распределение*** непрерывной случайной величины, функция плотности распределения которого имеет вид:

$$\varphi_{m,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}, (-\infty < m < +\infty, \sigma > 0)$$

Как показано на рисунке 7, данное распределение, в частности, вид его графика, зависит от двух параметров – m и σ. Его закон распределения записывается следующим образом:

$$\Phi_{m,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(x-m)^2}{2\sigma^2}} dx$$

А) Б)

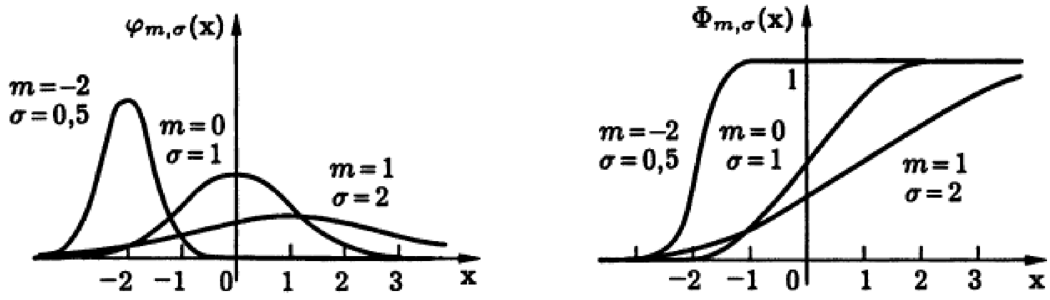


Рисунок 7 – График функции: а) плотности распределения Гаусса; б) закона распределения Гаусса

1.3 Распределение Стьюдента

Распределение Стьюдента (или t-распределение) используется для моделирования случайных величин, когда имеется ограниченное количество наблюдений и некоторая неопределенность в данных. Распределение Стьюдента получило свое название в честь псевдонима, используемого Уильямом Сили Госсетом, который в начале 20-го века внёс значительный вклад в решение проблемы, связанной с оценкой среднего значения на основе ограниченного количества наблюдений [4]. Вопрос, которым занимался Стьюдент, был связан с построением так называемых ***доверительных интервалов*** – диапазонов значений, в которых с определённой вероятностью содержится истинное значение параметра генеральной совокупности (например, среднего значения). Ранее для этого использовалось распределение Гаусса, однако при малом объеме выборки ($n \leq 30$ наблюдений) и неизвестном стандартном отклонении генеральной совокупности, использование нормального распределения становилось неточным – с этим и столкнулся Госсет, а после разработал распределение, основанное на t-статистике, которое позволяло делать более точные оценки и строить доверительные интервалы при использовании малых выборок и неизвестного стандартного отклонения генеральной совокупности. Он опубликовал свои исследования под псевдонимом «Стьюдент», откуда и произошло название распределения.

Распределением Стьюдента (t-распределением) называют распределение, функция плотности вероятности которого имеет вид:

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu} * \Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{\nu}{2}\right)}(1 + \frac{X^2}{\nu})^{-\frac{\nu+1}{2}},$$

где X – случайная величина, ν – число степеней свободы и $\Gamma(f)$ – гамма-функция.

Под гамма-функцией понимается гамма-функцией Эйлера – специальная функция в математике, введённая Леонардом Эйлером для обобщения понятия факториала на комплексные и действительные числа. Она имеет вид: [5]

$$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1}e^{-x}dx$$

Гамма-функция определена для всех комплексных чисел с положительной вещественной частью, то есть для $\alpha > 0$. Одним из интересных свойств гамма-функции является возможность применения к ней так называемой формулы понижения [5]:

$$\Gamma(\alpha + 1) \ = \ \alpha\Gamma(\alpha)$$

Это позволяет вычислять значения гамма-функции для разных аргументов, используя уже известные значения. Эйлерова гамма-функция также обладает симметрией, поэтому к ней применима формула отражения:

$$\Gamma(\mathrm{z}) \ * \ \Gamma(1 - \mathrm{z}) \ = \ \pi \ / \ \sin(\pi\mathrm{z})$$

Формула отражения позволяет вычислять значения гамма-функции для отрицательных аргументов или аргументов, удаленных от положительных целых чисел. В контексте теории вероятностей гамма-функция играет важную роль в определении и вычислении функций распределения и плотности распределения некоторых случайных величин, в частности, распределения Стьюдента. Гамма-функция входит в функцию плотности распределения Стьюдента через отношение $\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu}*\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{\nu}{2}\right)}$. Это отношение возникает из применения гамма-функции для нормировки t-статистики, которая определяется как отношение выборочного среднего и стандартной ошибки. Используя гамма-функцию, мы нормируем t-статистику, чтобы получить функцию плотности распределения Стьюдента.

Как видно из формулы функции распределения, t-распределение имеет один параметр, называемый ***степенью свободы*** распределения Стьюдента. Степень свободы влияет на ширину и форму распределения и является мерой количества независимых наблюдений, используемых для оценки распределения. Чем больше степень свободы, тем ближе t-распределение приближается к нормальному. При этом, при меньшей степени свободы, распределение имеет более плоскую форму.

График функции t-распределения, то есть ***кривая распределения Стьюдента*** имеет колоколообразную форму, напоминающую нормальное распределение, но с более тяжёлыми «хвостами». При $n > 30$ распределение стремится к нормальному закону распределения. На рисунке 8 показан график функции t-распределения при значениях $\nu_1 = 1, \nu_2 = 5, \nu_3 > 20$.

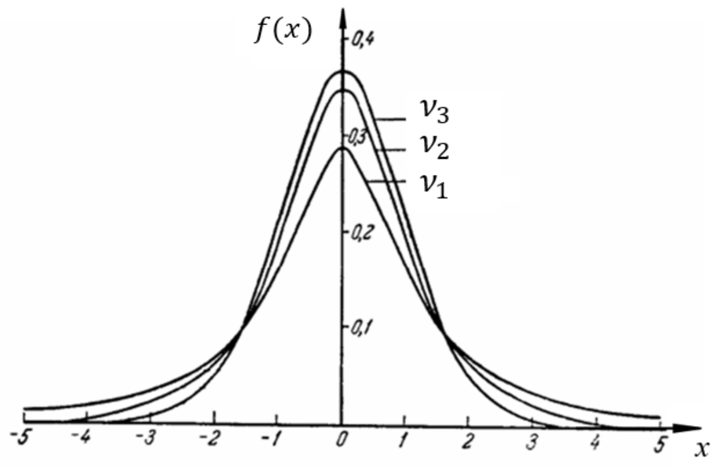


Рисунок 8 – График функции плотности t-распределения с числом степеней свободы $\nu_1 = 1, \nu_2 = 5$ и $\nu_3 > 20$

Функция распределения t-распределения в общем случае имеет вид:

$$F(x) = \int_{-\infty}^{x} f(x)dx$$

Она не имеет аналитического решения в данном случае, поэтому для вычисления её значений часто используются таблицы или специальные программы. Её график схож с графиком функции распределения нормального распределения, как показано на рисунке 9:

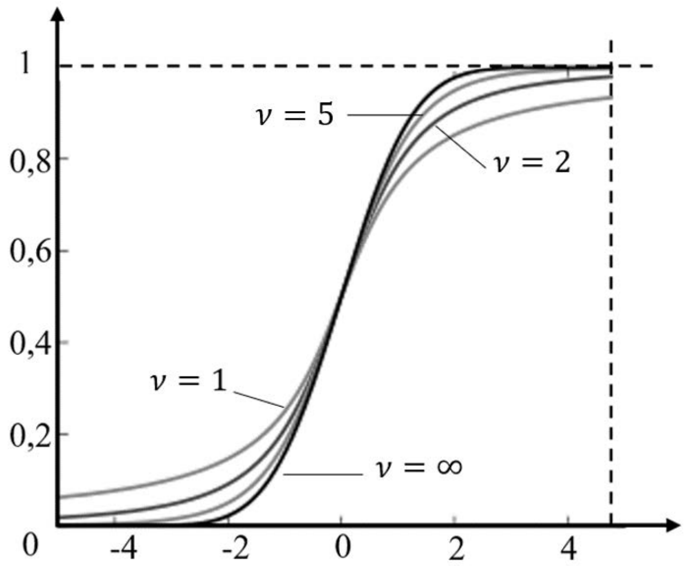


Рисунок 9 – График функции распределения t-распределения

Числовые характеристики распределения Стьюдента зависят от его степени свободы ν:

1. ***Математическое ожидание*** t-распределения определено и равно нулю при любом $\nu > 2$:

$$M = 0, \qquad \nu > 2$$

2. ***Дисперсия*** t-распределения определяется следующей формулой:

$$D = \frac{\nu}{\nu - 2}, \qquad \nu > 2$$

3. ***Среднеквадратичное отклонение*** t-распределения равно квадратному корню из дисперсии, то есть:

$$\sigma = \sqrt{\frac{\nu}{\nu - 2}}\,, \qquad \nu > 2$$

4. ***Центральный момент*** t-распределения n-го порядка определяется следующим образом:

$$\mu_n = \begin{cases} 0, & n = 2 * k + 1, k \in Z \\ (\frac{\nu}{\nu - 2})^{\frac{n}{2}} * \frac{\nu - 2}{(\nu - 4)!!}, > 4, & n = 2 * k, k \in Z \end{cases}$$

5. ***Начальный момент*** для t-распределения с ν степенями свободы не имеет аналитической формулы для $n \geq \nu$. Это связано с тем, что функция плотности вероятности t-распределения не интегрируется аналитически для степеней выше ν, а данная характеристика вычисляется численными методами или аппроксимациями.
6. ***Мода*** равна значению $f(x)$ при $x = 0$.
7. ***Медиана*** t-распределения также равна нулю.

Для анализа точности оценок статистических параметров на основе выборочных данных можно использовать формулу для ***вероятности попадания случайной величины X в доверительный интервал*** $(x_{\frac{\alpha}{2}}, x_{1-\frac{\alpha}{2}})$, которая выглядит следующим образом:

$$P\left(x_{\frac{\alpha}{2}} < X < \ x_{1-\frac{\alpha}{2}}\right) = 1 - \alpha,$$

где α – вероятность отклонения результатов эксперимента от нулевой гипотезы при условии, что она верна. Для нахождения значения ***уровня значимости*** α необходимо знать уровень доверия для построения доверительного интервала. Уровень доверия обычно задается в процентах и представляет собой вероятность того, что истинное значение параметра лежит внутри построенного доверительного интервала. Например, для уровня доверия 95% соответствующее значение α будет равно 0.05, так как это означает, что необходимо рассмотреть только 5% случаев, когда истинное значение параметра лежит за пределами доверительного интервала, а в остальных 95% случаев это значение лежит внутри этого интервала. Формула для нахождения уровня значимости может быть выражена следующим образом:

$$\alpha = \frac{1 - \text{уровень доверия (\%)}}{100}.$$

Если среднее значение выборки попадает в заданный доверительный интервал (то есть интервал, вероятность попадания среднего значения выборки в который определена заранее), то мы не можем отвергнуть нулевую гипотезу о том, что различия между двумя выборками статистически не значимы на уровне значимости 5%. Для того чтобы отклонить нулевую гипотезу и доказать статистическую значимость различий между двумя выборками, среднее значение выборки должно выходить за пределы заданного доверительного интервала [6].

Одним из важных понятий, связанных с t-распределением, является ***t-статистика*** – мера различия между двумя выборками данных. Она часто используется в статистических тестах гипотез, чтобы определить, является ли различие между выборками значимым или случайным. Если t-статистика больше критического значения, то можно сделать вывод о статистически значимом различии между выборками. Она может быть определена как отношение оценки параметра модели к её стандартной ошибке. В этом случае формула будет иметь вид:

$$t = \frac{b - \beta}{SE(b)},$$

где b – оценка параметра модели (числовое значение, которое определяет взаимосвязь между независимыми и зависимыми переменными в рамках выбранной модели), β – истинное значение параметра модели, а $SE(b)$ – стандартная ошибка оценки параметра модели.

Существует некоторое значение, которое разделяет вероятностное распределение Стьюдента на две области и называется оно ***квантилем***. Выражаясь более формально, t-квантиль уровня α – это такое число t_α, для которого вероятность $P(T < t_\alpha) = \alpha$, где T представляет распределение Стьюдента с n степенями свободы [7]. Квантиль порядка p также записывают как $Q(p)$.

В контексте задач систематизации и обработки результатов наблюдений массовых случайных явлений для выявления существующих закономерностей необходимо рассмотреть такое понятие, как ***совокупность всех подлежащих изучению объектов*** или возможных результатов наблюдений, проводимых в неизменных условиях над одним объектом, называется генеральной совокупностью. Выборочной совокупностью или ***выборкой*** называется совокупность объектов, отобранных случайным образом из генеральной совокупности. Число объектов в совокупности называется ее объемом. Считается, что объем генеральной выборки бесконечен. Конкретные значения выборки, полученные в результате наблюдений, называют реализацией выборки и обозначают $x_1, x_2, \dots, x_n$.

2 Применение t-распределения при решении практических задач

Среди задач, в которых для решения используется t-распределение, можно выделить следующие наиболее практически значимые:

- ***Оценка среднего значения*** генеральной совокупности при неизвестной дисперсии: если известен только размер выборки и выборочное среднее значение, то можно использовать распределение Стьюдента для оценки значимости различий между выборочным и генеральным средним.
- ***Расчет доверительного интервала для среднего значения*** генеральной совокупности: распределение Стьюдента также позволяет вычислить доверительный интервал для среднего значения генеральной совокупности с известной точностью.
- ***Оценка статистической значимости*** различий между двумя выборками: если известны средние значения и стандартные отклонения двух выборок, то можно использовать распределение Стьюдента для проверки гипотезы о равенстве этих средних значений.
- ***Оценка коэффициента корреляции***: если известны выборочные коэффициенты корреляции и размер выборки, то можно использовать распределение Стьюдента для проверки гипотезы о значимости корреляции.
- ***Оценка параметров линейной регрессии***: распределение Стьюдента используется для проверки гипотезы о значимости коэффициентов линейной регрессии при небольшом размере выборки.

Рассмотрим далее несколько конкретных примеров задач, решение которых связанно с t-распределением и t-статистикой Стьюдента.

2.1 Оценка среднего значения генеральной совокупности

Задача. Оценить средний рост студентов в университете среди случайным образом выбранных 10 студентов. Данные выборки X = [170, 165, 172, 168, 175, 172, 169, 171, 173, 170] [8]

Решение. Среднее значение выборки вычислим по формуле:

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} =$$

$$= \frac{170 + 165 + 172 + 168 + 175 + 172 + 169 + 171 + 173 + 170}{10} = 170.5.$$

Таким образом, средний рост студентов в выборке равен 170.5 см. Следом вычислим стандартное отклонение выборки:

$$s = \sqrt{\sum \frac{(x_i - \bar{X})^2}{n-1}} \approx 2.4314.$$

Следующим шагом будет вычисление значения стандартной ошибки среднего по формуле:

$$SE = \frac{s}{\sqrt{n}} = \frac{2.4314}{\sqrt{10}} \approx 0.7687.$$

Получим, что значение стандартной ошибки среднего составляет примерно 0.7687 см. При надежности 95% уровень значимости $\alpha = 0.05$ и число степеней свободы $\nu = n - 1 = 9.$ Из таблиц получаем значение t-критерия, равное $2{,}262.$ Теперь нам необходимо определить доверительный интервал CI, который показывает, с какой вероятностью истинное среднее значение генеральной совокупности находится в заданном интервале. Для вычисления доверительного интервала нам необходимо использовать формулу:

$$CI = \bar{X} \pm t * SE = 170.5 \pm 2$$

Мы получили, что доверительный интервал составляет $170.5 \pm 2.306 * 0.7687$, что равно $(169.24, 173.76)$. Таким образом, с 95% вероятностью мы можем утверждать, что истинное среднее значение роста студентов в университете находится в этом интервале.

Ответ: оценка среднего роста студентов в университете среди случайно выбранных 10 студентов со 95% вероятностью лежит в интервале $(169.24, 173.76)$.

2.2 Использование t-распределения для вычисления вероятности

Задача. Найти вероятность того, что случайная величина X с десятью степенями свободы лежит в пределах $(1{,}5 < X < 2)$ [9].

Решение. Известно, что $\nu = 10$. Для решения задачи необходимо сначала перейти от случайной величины X к соответствующей ей случайной величине t_{10} с распределением Стьюдента. По таблицам t-распределения найдём вероятности соответствующих интервалов $(1{,}372 < t_{10} < 1{,}812)$ и $(1{,}812 < t_{10} < 2{,}228)$.

$$P_1 = P(1{,}372 < t_{10} < 1{,}812) = 0{,}05$$
$$P_2 = P(1{,}812 < t_{10} < 2{,}228) = 0{,}025$$

Для распределения Стьюдента со степенями свободы ν существует также таблица значений, которая позволяет найти квантили $Q(p)$ порядка p для определенных значений ν и уровня значимости α. Исходя из них в данном случае уровень значимости $\alpha = 0{,}05$. Теперь необходимо найти значения t-статистики для интервалов, соответствующих квантилям $Q(0{,}025)$ и $Q(0{,}975)$. Находим, что первый из них $Q(0{,}025)$ соответствует t-статистике, равной 1,812, а $Q(0{,}975)$ соответствует t-статистике, равной 2,228.

Наконец, для вычисления искомой вероятности необходимо привести интервал $(1{,}5 < X < 2)$ к виду $(\alpha < t_{10} < \beta)$, где α и β соответствуют найденным значениям t-статистики. Для этого используется формула линейной интерполяции и искомую вероятность можно вычислить путем взвешенного среднего вероятностей попадания в каждый из интервалов при условии, что случайная величина находится в интервале $(1{,}5 < X < 2)$:

$$P(1{,}5 < X < 2) = 0{,}05 * \left(\frac{1{,}812 - 1{,}5}{1{,}812 - 1{,}372}\right) + 0{,}025 * \left(\frac{2 - 1{,}812}{2{,}228 - 1{,}812}\right) \approx 0{,}047.$$

Ответ: 0,047.

2.3 Задача об оценке статистической значимости (анализ регрессии)

Задача. Были случайным образом выбраны 8 учеников, для которых получены следующие выборки:

время на домашнюю работу (в часах) X = [2, 4, 3, 1, 5, 6, 7, 8];

результаты экзамена (в баллах) Y = [60, 70, 65, 55, 75, 80, 85, 90].

Необходимо оценить зависимость между количеством времени, проведенным на дому за учебой, и результатами экзамена по математике [10].

Решение. Для оценки параметров линейной регрессии используем распределение Стьюдента. Тогда искомая модель будет иметь вид: $y = b_0 + b_1 * x$, где b_0 и b_1 – неизвестные коэффициенты, которые необходимо найти.

Сперва найдем по известным формулам для каждой переменной такие характеристики, как выборочное среднее и стандартное отклонение:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{2 + 4 + 3 + 1 + 5 + 6 + 7 + 8}{8} = 4.125$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{60 + 70 + 65 + 55 + 75 + 80 + 85 + 90}{8} = 73.75$$

$$s_x = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n - 1}} = 2.439$$

$$s_y = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n - 1}} = 11.44552$$

Затем рассчитаем выборочный коэффициент корреляции r между x и y:

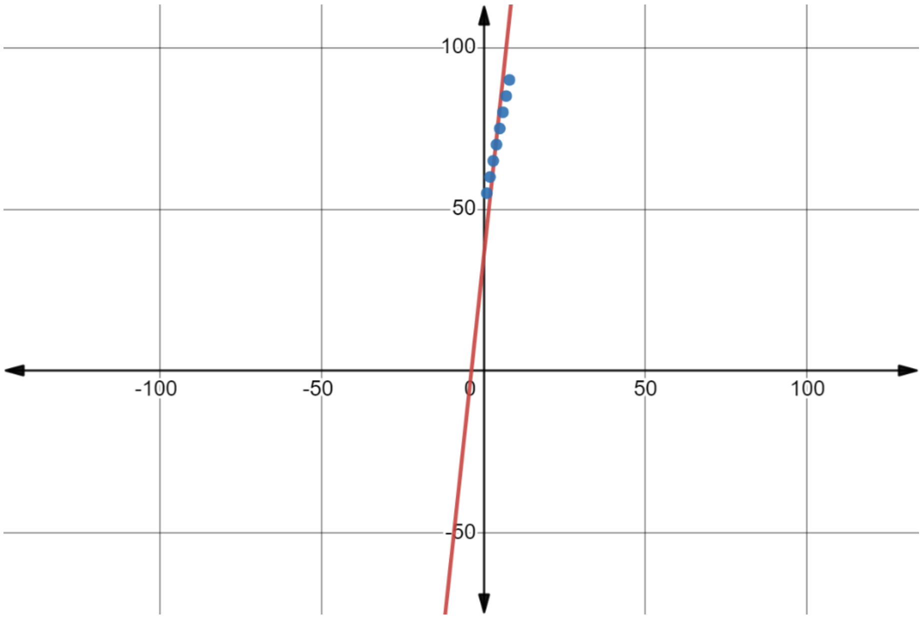
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x * s_y} = 0.86195$$

Теперь мы можем получить оценки коэффициентов b_0 и b_1:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \approx 9.28458$$

$$b_0 = \bar{y} - b_1 * \bar{x} = 35.00724$$

Таким образом, наша модель линейной регрессии будет иметь вид $y = 35.00724 + 9.28458 * x$. Графически это выглядит следующим образом:



Проверим значимость коэффициента наклона b_1 с помощью t-теста. Для этого нужно рассчитать значение t-статистики по формуле:

$$t = \frac{b_1 - 0}{SE(b_1)} = \frac{b_1\sqrt{\sum(x_i - \bar{x})^2}}{SE_{res}} = 5.309$$

Значение t-статистики соответствует уровню значимости $p = 0.0013$ при числе степеней свободы $n - 2 = 6$. Это означает, что полученный коэффициент наклона является статистически значимым.

Ответ: $b_1 = 9.28458$ является статистически значимым.

2.4 Проверка гипотезы о среднем значении показателя в генеральной совокупности

Задача. Дана выборка из 10 наблюдений. Необходимо проверить гипотезу о том, что среднее значение показателя в генеральной совокупности равно 50. Известно, что стандартное отклонение в выборке составляет 5, а выборочное среднее равно 48 [8].

Решение. Тогда для проверки гипотезы используется t-критерий Стьюдента. Формула для расчета значения t-статистики имеет вид:

$$t = \frac{(\bar{x} - \mu)\sqrt{n}}{s} = \frac{(48 - 50)\sqrt{10}}{5} \approx -1.41$$

Далее нам необходимо найти критическое значение t-статистики для заданного уровня значимости и числа степеней свободы. В данном случае число степеней свободы равно $\nu = n - 1 = 9$. Положим, что уровень значимости α = 0.05. Тогда критическое значение t-статистики можно найти с помощью таблицы распределения Стьюдента $t_{\text{крит}} = \pm 2.262$.

$$|t| = |-1{,}41| < 2{,}262.$$

Следовательно, мы не можем отвергнуть гипотезу о равенстве среднего значения 50 на уровне значимости α = 0.05.

Ответ: гипотеза верна.

2.5 Задача о проверке гипотезы о значимости коэффициента регрессии

Задача. Проверить гипотезу о значимости коэффициента регрессии для независимой переменной x при уровне значимости $\alpha = 0{,}05$ для следующих данных [8]: $n = 10$, x $= [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]$, y $= [1, 3, 2, 5, 7, 8, 8, 9, 10, 12]$.

Решение. Мы хотим проверить гипотезу о значимости коэффициента регрессии для независимой переменной, то есть проверить, является ли коэффициент значимым или нет. Пусть у нас есть гипотеза $H_0: \beta_1 = 0$ (нулевая гипотеза), где β_1 – коэффициент регрессии для независимой переменной x. Альтернативная ей гипотеза $H_а: \beta_1 \neq 0$. Таким образом задача сводится к проверке, является ли коэффициент регрессии β_1 значимым. Для проверки можно воспользоваться t-статистикой Стьюдента:

$$t = \frac{b_1 - 0}{\frac{s}{\sqrt{n-2}}}$$

Где b_1 – оценка коэффициента регрессии для x, s – среднеквадратическое отклонение резидуальных ошибок, n – число наблюдений.

Найдем оценку коэффициента регрессии b_1 и среднеквадратическое отклонение s следующим образом:

$$b_1 = \sum_{i=1}^{10} \frac{(x_i - \bar{x})(y_i - \bar{y})}{(x_i - \bar{x})^2} \approx 1.1697$$

$$s = \sqrt{\frac{\sum_{i=1}^{10}(x_i - \bar{x})^2}{n-1}} \approx 1{,}7845$$

И, наконец, определим t-статистику Стьюдента:

$$t = \frac{1.1697}{\frac{1{,}7845}{\sqrt{8}}} = 4{,}238$$

Согласно табличным данным, на уровне значимости $\alpha = 0{,}05$ и степеней свободы $n - 2 = 8$ критическое значение t-статистики равно $2{,}306$. Так как вычисленное значение $4{,}238 > 2{,}306$, то есть оно больше критического, то мы можем отвергнуть нулевую гипотезу о том, что коэффициент регрессии равен нулю на уровне значимости $\alpha = 0{,}05$. То есть, можно сделать вывод о том, что найденный коэффициент регрессии b_1 статистически значимо отличается от нуля и имеет влияние на зависимую переменную y.

Ответ: существует статистически значимая связь между зависимой переменной y и независимой переменной x.

Заключение

В ходе выполнения данной курсовой работы был рассмотрен такой раздел теории вероятностей, как случайные величины и их распределения, в частности, t-распределение Стьюдента – изучены его основные свойства и общие характеристики. Также были рассмотрены основные виды практических статистических и вероятностных задач, связанных с t-распределением, а также приведено решение нескольких примеров таких задач.

Можно заключить, что благодаря работе Уильяма Госсета и его вкладу в разработку распределения Стьюдента, статистики и исследователи получили важный инструмент, позволяющий делать более точные оценки и строить доверительные интервалы при работе с ограниченными выборками данных. Это имеет важное значение в различных областях, где доступ к большим объемам данных ограничен или невозможен.

Список использованных источников

1. Феллер В. Введение в теорию вероятностей и ее приложения. В 2-х томах. Т. 1: Пер. с англ. – М.: Мир, 1984. – 528 с.
2. Печинкин А.В., Тескин О.И., Г.М. Цветкова и др.; Под ред. Зарубина В.С., Крищенко А.П. Теория вероятностей: Учеб. для вузов. - 3-е изд., испр. изд. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2004. - 456 с.
3. Теория вероятностей и математическая статистика: учеб. пособие / Е. А. Трофимова, Н. В. Кисляк, Д. В. Гилёв; [под общ. ред. Е. А. Трофимовой ; М-во образования и науки Рос. Федерации, Урал. федер. ун-т. – Екатеринбург: Изд-во Урал. ун-та, 2018. – 160 с.
4. Гнеденко Б.В. Курс теории вероятностей: Учебник. – Изд. 8-е., испр. и доп. изд. – М.: Едиториал УРСС, 2005. – 448 с.
5. Зорич В. А. Математический анализ. Часть II. — Изд. 9-е, испр. — М.: МЦНМО, 2019. — xii+676 с. Библ.: 57 назв. Илл.: 41.
6. Понятие t-распределения в тестах на статистическую значимость // RadioProg URL: https://radioprog.ru/post/930 (дата обращения: 21.05.2023).
7. Фишер Р. Applications of «Student's» distribution (англ.) // metron. – 1925. – Vol. 5. – P. 90 – 104.
8. Леман, Э.Л. Проверка статистических гипотез / перевод с англ. Ю.В. Прохорова. – 2-е изд., испр. – М.: Наука, 1979. – 408 с.
9. Поллард. Д. Справочник по вычислительным методам статистики / Дж. Поллард; Перевод с англ. В. С. Занадворова. – М.: Финансы и статистика, 1982. – 344 с.
10. Кобзарь А.И. Прикладная математическая статистика [Текст]: для инженеров и науч. работников / А. И. Кобзарь. – Изд. 2-е, испр. – Москва : Физматлит, 2012. – 813 с.