Автономная некоммерческая образовательная организация
высшего образования Центросоюза Российской Федерации
«Сибирский университет потребительской кооперации»

Методические указания и задания
по выполнению практических и самостоятельных работ
по дисциплине

ОП.03 ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

по специальности
09.02.13 Интеграция решений с применением технологий искусственного интеллекта

(направленность программы: Применение искусственного интеллекта)

квалификация выпускника:
Специалист по работе с искусственным интеллектом

Новосибирск
2025

Методические указания и задания по выполнению практических и самостоятельных работ по дисциплине *«Теория вероятностей и математическая статистика»* для обучающихся по специальности *09.02.13* Интеграция решений с применением технологий искусственного интеллекта/[сост. В. В. Комиссаров доцент];– Новосибирск, 2025.

РЕЦЕНЗЕНТ: С.Л. Злобина, канд.физ-мат.наук, доцент кафедры статистики и математики

Методические указания и задания по выполнению практических и самостоятельных работ рекомендованы к использованию в учебном процессе на заседании кафедры статистики и математики, протокол от 28 мая 2025 г. № 9.

.

СОДЕРЖАНИЕ

ОБЩИЕ ПОЛОЖЕНИЯ

Данная работа предназначена для студентов, изучающих дисциплину «Теория вероятностей и математическая статистика».

Предлагаемые методические указания содержат краткие теоретические сведения и примеры решения типовых задач по дисциплине «Теория вероятностей и математическая статистика» Поскольку эти задачи не охватывают весь программный материал, то ответы на вопросы студент может найти в любом учебнике из «Списка рекомендуемой литературы», а на многие важные вопросы ответы имеются в этой методической работе.

ТЕМА И ИХ КРАТКОЕ СОДЕРЖАНИЕ

Основные понятия теории вероятностей

Испытание **–** это изначальное понятие**,** разъясняется как действие, наблюдение, опыт и прочее.

Событие – это результат испытания.

Пример 1: Некто подбросил монету, которая упала гербом вверх. Здесь испытание – подбрасывание монеты, а результат этого испытания – выпадение герба – это событие.

Пример 2: В результате подбрасывания игрального кубика выпало три очка на верхней грани. В этом случае, испытание – подбрасывание кубика, а выпадение трех очков – событие.

Заметим, что монета в примере 1 могла упасть не гербом, а решкой (цифрой вверх). Аналогично, в примере 2, подбрасывание кубика могло бы закончиться выпадением, например, двух или пяти очков. Событие, которое в результате испытания может произойти, а может и не произойти, называется ***случайным***.

Пусть в результате испытания могут появиться несколько событий. События называются ***несовместными,*** если появление одного из них исключает появление других.

Пример 3: Рассмотрим такое испытание, как сдача экзамена по математике одним из студентов. События, которые, например, могут произойти в результате этого испытания, есть следующие:

A – экзамен сдан на оценку «4»,

B – экзамен сдан на оценку «3»,

C – экзамен сдан на оценку выше, чем «3».

В этом случае, события А и В несовместны, так как получение оценки «3» делает невозможным получение оценки «4» за этот же экзамен. Наоборот, события А и С совместны, поскольку они могут произойти одновременно.

Пространством элементарных исходов (или ***событий***), соответствующих рассматриваемому испытанию, будем называть такое множество несовместных событий, одно из которых обязательно произойдет в результате испытания, так что любой интересующий нас результат испытания может быть однозначно описан с помощью элементов этого множества.

В примере с игральным кубиком пространство элементарных исходов образуют 6 событий: $E_1, E_2, E_3, E_4, E_5, E_6,$ которые заключаются в том, что количество выпавших очков составит соответственно 1, 2, 3, 4, 5 или 6 очков. Действительно, эти события несовместны, одно из них обязательно произойдет в результате подбрасывания кубика, и с их помощью можно описать любые другие события. Например, событие A – выпало четное число очков – означает, что появились события E_2 или E_4 или E_6, эти три элементарных исхода благоприятствуют наступлению события A.

Классическое определение вероятности

Элементарные исходы называются ***равновозможными,*** если ни у одного из них нет преимуществ перед другими, чтобы произойти в результате испытания.

Вероятностью события А называется число $P(A)$, равное отношению числа ***m*** благоприятствующих событию A **элементарных** исходов к общему числу ***n*** элементарных **равновозможных** исходов:

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

Рассмотрим пример с урновой схемой. Урна – это ёмкость с шарами.

Пример 4: Пусть в урне находится 20 одинаковых шаров, которые отличаются только цветом, например, 12 из них красные, а остальные – белые. Некто подошел к урне и наугад выбрал один шар. Найдем вероятность того, что этот шар – красный.

Пусть событие К – выбран красный шар. Всего элементарных исходов $n = 20$ (по количеству шаров), причём все эти исходы равновозможные. Событию К благоприятствует $m = 12$ исходов (по количеству красных шаров), поэтому

$$P(K) = \frac{m}{n} = \frac{12}{20} = 0{,}6.$$

Вероятность любого события может принимать значения только от 0 до 1 включительно, то есть

$$0 \le P(A) \le 1.$$

Достоверное событие – обязательно произойдет в результате испытания, то есть $m = n$, так как все исходы благоприятные:

$$P(A) = \frac{m}{n} = \frac{n}{n} = 1.$$

Невозможное событие – не может произойти в результате испытания, то есть $m = 0$, так как благоприятных исходов нет:

$$P(A) = \frac{m}{n} = \frac{0}{n} = 0.$$

Случайное событие – может произойти или не произойти в результате испытания: $0 < P(A) < 1$.

Вероятность является числовой мерой объективной возможности наступления события. Вероятность можно задать в процентах, например $P(A) = 0{,}8$ (80%).

Основные теоремы теории вероятностей

Условной вероятностью $P_A(B)$ называется вероятность события B, вычисленная в предположении, что событие A уже наступило.

Пример 5. В урне имеется 3 белых и 2 черных шара. Из урны дважды вынимают по одному шару, не возвращая их обратно. Найти вероятность появления белого шара при втором испытании (событие B), если при первом испытании был извлечен черный шар (событие A).

Решение: Изначально в урне было 5 шаров, из которых 3белых и 2 черных. После первого испытания в урне осталось 4 шара, из них 3 белых. Искомая условная вероятность $P_A(B) = 3/4$.

События A и B называются ***независимыми***, если появление одного из этих событий не изменяет вероятности наступления другого, то есть $P(A) = P_B(A)$ или $P(B) = P_A(B)$. В противном случае события называются ***зависимыми***.

Теорема сложения вероятностей **несовместных** событий:

$$P(A \text{ или } B) = P(A) + P(B)$$

Теорема умножения вероятностей **независимых** событий:

$$P(A \text{ и } B) = P(A) \cdot P(B)$$

Теорема умножения вероятностей **зависимых** событий:

$$P(A \text{ и } B) = P(A) \cdot P_A(B)$$

Теорема сложения вероятностей **совместных** событий:

$$P(A \text{ или } B) = P(A) + P(B) - P(A \text{ и } B)$$

Событие $\overline{A}$ (не ***А***) называется ***противоположным*** событию *А*, если оно наступает тогда и только тогда, когда не наступает событие A.

Пример 4: а) Событие A – изделие бракованное, тогда $\overline{A}$ – изделие без брака; б) *В* – студент сдал экзамен, тогда событие $\overline{B}$ – студент не сдал экзамен; с) *С* – хотя бы один лотерейный билет выиграл, тогда $\overline{C}$ – ни один билет не выиграл. Из приведенных примеров видно, что противоположное событие можно сформулировать путем простого логического отрицания.

Вероятность противоположного события находится по формуле

$$P(\overline{A}) = 1 - P(A)$$

Задача. В партии из 100 одинаковых по внешнему виду изделий смешаны 40 шт. первого сорта и 60 шт. второго сорта. Найти вероятность того, что взятые наугад два изделия окажутся:

а) оба первого сорта;

б) разных сортов;

в) хотя бы одно из них первого сорта.

Найти указанные вероятности, если изделия выбираются по схеме выборки 1) с возвращением; 2) без возвращения.

Решение. 1). Рассмотрим вначале случай, когда изделия выбирают по схеме выборки с возвращением. В этом случае первое изделие из партии выбирается случайным образом, определяется его сортность, затем оно возвращается в партию и может быть выбрано повторно. Второе изделие выбирается из той же партии, состоящей из ста изделий. Обозначим:

событие A – первое взятое изделие I сорта,

событие $\overline{A}$ – первое взятое изделие II сорта (не I сорта),

событие B - второе взятое изделие I сорта ,

событие $\overline{B}$ – второе взятое изделие II сорта (не I сорта).

Заметим, что в рассматриваемом случае события A и B независимые, так как вероятность события B не зависит от того, какого сорта было выбрано первое изделие.

а) P(оба изделия I сорта) = $P(A$ и $B) = P(A) \cdot P(B) = \frac{40}{100} \cdot \frac{40}{100} = 0{,}16.$

Здесь мы воспользовались теоремой умножения вероятностей независимых событий.

б) P (изделия разных сортов) = $P(A$ и $\overline{B}$ или $\overline{A}$ и $B)$ =

$= P(A) \cdot P(\overline{B}) + P(\overline{A}) \cdot P(B) = \frac{40}{100} \cdot \frac{60}{100} + \frac{60}{100} \cdot \frac{40}{100} = 0{,}4 \cdot 0{,}6 +$

$+ 0{,}6 \cdot 0{,}4 = 0{,}48.$

Здесь мы воспользовались теоремой сложения несовместных событий и теоремой умножения для независимых событий.

в) P (хотя бы одно изделие I сорта) = 1 – P (нет ни одного изделия I сорта) = 1 – P (оба изделия II сорта) = $1 - P(\overline{A}$ и $\overline{B}) = 1 - P(\overline{A}) \cdot P(\overline{B}) = 1 - \frac{60}{100} \cdot \frac{60}{100} = 0{,}64.$

Здесь мы воспользовались формулой для нахождения вероятности противоположного события.

2). Далее, рассмотрим случай, когда изделия выбирают по схеме выборки без возвращения. В этом случае первое изделие из партии выбирается случайным образом, определяется его сортность, но в партию оно **не** возвращается. Второе изделие выбирается из оставшихся изделий. Подчеркнем, что в этом случае события A и B являются зависимыми. Найдем вероятности событий.

а) P(оба изделия I сорта) = $P(A$ и $B) = P(A) \cdot P_A(B) = \frac{40}{100} \cdot \frac{39}{99} = \frac{78}{495}$.

Здесь мы воспользовались теоремой умножения вероятностей для зависимых событий. Поясним более подробно, как были найдены вероятности $P(A)$ и $P_A(B)$. При выборе первого изделия общее число исходов равно 100 (по числу изделий) и все они равновозможны. Благоприятствующих из них событию A – 40 (по числу изделий I сорта). По классическому определению вероятности $P(A) = m / n = 40 /100$. Найдем условную вероятность $P_A(B)$, то есть вероятность выбрать второе изделие I сорта при условии, что первое выбранное изделие было также I сорта. Посмотрим, как изменился состав партии после того, как выбрали первое изделие: изделий в партии осталось 100 – 1 = 99 (первое забрали), среди них изделий первого сорта осталось 40 – 1 = 39 (выбранное первое изделие было I сорта). Далее, выбираем второе изделие: общее число исходов этого испытания 99 (по числу оставшихся изделий). Благоприятствующих из них событию B – 39 (по числу оставшихся изделий I сорта). По классическому определению вероятности

$P_A(B) = m / n = 39 /99.$

б) P (изделия разных сортов) = $P(A$ и $\overline{B}$ или $\overline{A}$ и $B)$ =

$$= P(A) \cdot P_A(\overline{B}) + P(\overline{A}) \cdot P_{\overline{A}}(B) = \frac{40}{100} \cdot \frac{60}{99} + \frac{60}{100} \cdot \frac{40}{99} = \frac{16}{33}.$$

Здесь мы воспользовались теоремой сложения вероятностей несовместных событий и теоремой умножения вероятностей зависимых событий.

в) P (хотя бы одно изделие I сорта) = 1 – P (ни одного нет

изделия I сорта) = 1 – P (оба изделия II сорта) = $1 - P(\overline{A}$ и $\overline{B})$ =

$$1 - P(\overline{A}) \cdot P_{\overline{A}}(\overline{B}) = 1 - \frac{60}{100} \cdot \frac{59}{99} = \frac{106}{165}.$$

Формула полной вероятности

События $H_1, H_2, \ldots, H_n$ образуют **полную группу**, если они попарно несовместны и в результате испытания одно из них обязательно произойдет.

Для таких событий справедливо равенство:

$$P(H_1) + P(H_2) + \ldots + P(H_n) = 1.$$

Противоположные события A и $\overline{A}$ всегда образуют полную группу, поэтому

$$P(A) + P(\overline{A}) = 1 \qquad \text{или} \qquad P(\overline{A}) = 1 - P(A).$$

Пусть событие *A* наступает с одним из событий (гипотез) H_i, тогда вероятность этого события находится по формуле, называемой ***формулой полной вероятности***

$$P(A) = P(H_1) \cdot P_{H_1}(A) + P(H_2) \cdot P_{H_2}(A) + \ldots + P(H_n) \cdot P_{H_n}(A),$$

где события $H_1, H_2, \ldots, H_n$ образуют полную группу.

Задача. Два консервных завода поставляют в магазин мясные и овощные консервы, причем первый завод поставляет 75% всей продукции. Доля овощных консервов в продукции первого завода составляет 60%, а у второго 70%. Для контроля в магазине взято наугад одно изделие. Какова вероятность того, что это окажутся мясные консервы?

Решение. Обозначим:

событие *A* – взяты мясные консервы;

событие H_1 – консервы изготовлены I заводом;

событие H_2 – консервы изготовлены II заводом.

По условию задачи первый завод поставляет 75% продукции, тогда $P(H_1) = 0{,}75$, второй завод поставляет - 25%, следовательно $P(H_2) = 0{,}25$. Вероятность того, что консервы мясные, для первого завода составляет 40%, то есть $Ð_{f_1}(À) = 0{,}4$, для второго завода - 30%, то есть $Ð_{f_2}(À) = 0{,}3$. Учитывая, что событие *A* произойдет обязательно с одним из событий (гипотез) H_i, образующих полную группу, применяем формулу полной вероятности:

$$P(A) = P(H_1) \cdot P_{H_1}(A) + P(H_2) \cdot P_{H_2}(A) = 0.75 \cdot 0{,}4 + 0{,}25 \cdot 0{,}3 = 0{,}375.$$

Замечание. Задача решается аналогично, если количество заводов будет три или более. Соответственно увеличится число слагаемых в формуле полной вероятности.

Повторение независимых испытаний

Пусть известна вероятность появления события *A* в каждом испытании: *P(A)* = ***p***, тогда $P(\overline{A}) = 1 -$ ***p = q*** – вероятность не появления события *A*. Испытание повторяется ***n*** раз. Требуется найти вероятность того, что событие *A* наступит при этом ровно ***k*** раз.

Обозначим $\boldsymbol{P_n(k)}$ – вероятность того, что в ***n*** испытаниях событие *A* наступит ***k*** раз. Эта вероятность находится по формуле Бернулли:

$$P_n(k) = \frac{n!}{k!(n-k)!} \cdot p^k \cdot q^{n-k},$$

! - знак факториала, математической операции такой, что

$n! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n$. Например, $1! = 1$

$2! = 1 \cdot 2 = 2$

Внимание: **0!=1** $3! = 1 \cdot 2 \cdot 3 = 6$

$4! = 1 \cdot 2 \cdot 3 \cdot 4 = 24$

$5! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 120$ и т. д.

Наивероятнейшее число появлений события

Пусть в ***n*** повторных испытаниях событие *A* появляется ***k*** раз, где ***k*** может принимать значения: **0; 1; 2;** …; ***n*** (то есть $\boldsymbol{0 \le k \le n}$). Для каждого из этих значений ***k*** можно найти соответствующую ему вероятность по формуле Бернулли.

Значение ***k***, которому соответствует самая большая вероятность, называется ***наивероятнейшим*** числом появления события *A*.

Наивероятнейшее число k_0 находится как ***целое число*** из промежутка:

$$np - q \le k_0 \le np + p .$$

При этом $\boldsymbol{k_0}$ может принимать либо одно значение, либо два соседних целых значения с одинаковой вероятностью.

Вероятность $P_n(k_o)$, соответствующую значению $k = k_0$, находим по формуле Бернулли.

Появление события хотя бы один раз

Вероятность появления события «хотя бы один раз» в ***n*** испытаниях находится с помощью противоположного ему события «ни одного раза»:

P_n (событие наступит хотя бы один раз) = 1 – P_n (ни разу)

$$= 1 - P_n(0) = 1 - \frac{n!}{0! \cdot n!} \cdot p^0 \cdot q^{n-0} = 1 - q^n ,$$

при этом учтено, что $0! = 1$ и $p^0 = 1$.

Событие наступит «хотя бы один раз» означает, что оно наступит один или более раз, поэтому можно записать:

$$P_n(k \ge 1) = 1 - q^n .$$

Задача. Стрелок поражает цель с вероятностью 0,7. С какой вероятностью в серии из 5 выстрелов он поразит мишень:

а) ровно два раза;

б) более трех раз;

в) хотя бы один раз;

г) указать наивероятнейшее число попаданий и соответствующую ему вероятность.

Решение. По условию задачи: $\boldsymbol{p} = 0{,}7$; $\boldsymbol{n} = 5$; $\boldsymbol{k} = 2$; $\boldsymbol{m} = 3$;

вероятность промаха $\boldsymbol{q} = 1 - \boldsymbol{p} = 1 - 0{,}7 = 0{,}3$.

а) Вероятность попадания ровно два раза в серии из пяти выстрелов находим по формуле Бернулли:

$$P_5(2) = \frac{5!}{2!(5-2)!} \cdot p^2 \cdot q^3 = \frac{5!}{2!\cdot 3!} \cdot 0{,}7^2 \cdot 0{,}3^3 =$$

$$= \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5}{1 \cdot 2 \cdot 1 \cdot 2 \cdot 3} \cdot 0{,}49 \cdot 0{,}027 = 10 \cdot 0{,}01323 = 0{,}1323 .$$

б) Событие «стрелок поразит мишень более трех раз» запишем в виде: $m > 3$, тогда

$$P_5(m > 3) = P_5(4 \text{ или } 5) = P_5(4) + P_5(5) .$$

Здесь применена теорема сложения вероятностей несовместных событий. Используя формулу Бернулли, найдем:

$$P_5(4) = \frac{5!}{4!\cdot(5-4)!} \cdot p^4 \cdot q^1 = \frac{5!}{4!\cdot 1!} \cdot 0{,}7^4 \cdot 0{,}3 = 5 \cdot 0{,}2401 \cdot 0{,}3 = 0{,}36015;$$

$$P_5(5) = \frac{5!}{5!\cdot 0!} \cdot p^5 \cdot q^0 = 1 \cdot p^5 \cdot 1 = 0{,}7^5 = 0{,}16807;$$

$$P_5(m > 3) = 0{,}36015 + 0{,}16807 = 0{,}52822 .$$

в) Событию D – «стрелок поразит мишень хотя бы 1 раз», противоположно событие $\overline{D}$ – «не поразит ни разу», то есть стрелок промахнется все пять раз, следовательно, число попаданий $k = 0$:

$$P(D) = 1 - P(\bar{D}) = 1 - P_5(0) = 1 - \frac{5!}{0!\cdot 5!} \cdot p^0 \cdot q^5 = 1 - q^5 =$$

$$= 1 - 0{,}3^5 = 1 - 0{,}00243 = 0{,}99757 .$$

Здесь учтено, что $0! = 1$ и $p^0 = 1$.

г) Наивероятнейшее число попаданий k_o находим как ***целое*** число из промежутка:

$$np - q \le k_o \le np + p \ ;$$

$$5 \cdot 0{,}7 - 0{,}3 \le k_o \le 5 \cdot 0{,}7 + 0{,}7 \ ;$$

$$3{,}2 \le k_o \le 4{,}2 \ ;$$

$$k_o = 4 \ .$$

Соответствующую ему вероятность $P_5(4)$ вычислим по формуле Бернулли. В данной задаче она уже была найдена выше:

$$P_5(4) = \frac{5!}{4! \cdot 1!} \cdot 0{,}7^4 \cdot 0{,}3 = 0{,}36015.$$

Основные характеристики дискретной случайной величины

Случайной величиной называется переменная, принимающая свои возможные числовые значения с определенной вероятностью.

Например: X – балл, полученный на экзамене;

Y – число студентов, явившихся на лекцию;

Z – величина выигрыша в лотерее;

U – рост случайно выбранного человека и т.п.

Дискретная случайная величина X принимает отдельные числовые значения. Закон распределения дискретной случайной величины записывается в виде таблицы, где перечислены все значения случайной величины X и соответствующие им вероятности:

X	x_1	x_2	x_3	…	x_n
P(X)	p_1	p_2	p_3	…	p_n

Следует иметь в виду, что всегда $\sum_{i=1}^{n} p_i = p_1 + p_2 + ... + p_n = 1.$

Основные числовые характеристики закона распределения дискретной случайной величины:

1) ***Математическое ожидание*** (ожидаемое среднее значение случайной величины)

$$M(X) = \sum_{i=1}^{n} x_i p_i = x_1 p_1 + x_2 p_2 + \ldots x_n \ p_n = a.$$

2) ***Дисперсия*** (мера рассеяния значений случайной величины X от среднего значения ***a***):

$$D(X) = \sum_{i=1}^{n} (x_i - a\ )^2 p_i = (x_1 - a)^2 p_1 + (x_2 - a)^2 p_2 + \ldots + (x_n - a)^2 p_n$$

Второй способ вычисления дисперсии:

$$D(X) = M(X^2) - \left[M(X)\right]^2,$$

где $M(X)$ определено выше, а

$$M(X^2) = x_1^2 p_1 + x_2^2 p_2 + \ldots + x_n{}^2 p_n = \sum x_i^2 p_i \ .$$

3) ***Среднее квадратическое отклонение*** (характеристика рассеяния в единицах признака X):

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

Задача. В лотерее на каждые 100 билетов приходится 2 билета с выигрышем по 50 тыс. рублей, 5 билетов по 20 тыс. рублей, 10 билетов по 10 тыс. рублей, 20 билетов по 5 тыс. рублей и 25 билетов по 3 тыс. рублей. Остальные билеты не выигрывают. Составить закон распределения величины выигрыша для владельца одного билета и найти его основные характеристики.

Решение. Обозначим X тыс. рублей – величина выигрыша на один билет. Очевидно, что X – случайная дискретная величина. Составим закон распределения этой случайной величины, перечислив все ее возможные значения и найдя соответствующие им вероятности. Число выигрышных билетов из 100 составляет: 2+5+10+20+25=62, значит, число невыигрышных билетов: 100 – 62 = 38.

Располагая величины возможного выигрыша x_i в порядке возрастания, получим следующую таблицу:

x_i	0	3	5	10	20	50
p_i	0,38	0,25	0,20	0,10	0,05	0,02

где $p_1 = P(X=0) = \frac{38}{100} = 0{,}38$; $p_2 = P(X=3) = \frac{25}{100} = 0{,}25$ и т. д.

Отметим, что $\sum p_i = 0{,}38 + 0{,}25 + 0{,}20 + 0{,}10 + 0{,}05 + 0{,}02 = 1.$

1) Математическое ожидание случайной величины *X:*

$$M(X) = \sum x_i p_i = 0 \cdot 0{,}38 + 3 \cdot 0{,}25 + 5 \cdot 0{,}2 + 10 \cdot 0{,}1 + 20 \cdot 0{,}05 + 50 \cdot 0{,}02 = 4{,}75.$$

Таким образом, ожидаемый средний выигрыш на 1 билет составляет 4,75 тыс. рублей.

2) Дисперсию случайной величины найдем двумя способами:

1). $$D(X) = \sum_{i=1}^{6} \left[x_i - M(X)\right]^2 \cdot p_i =$$

$$= (0 - 4{,}75)^2 \cdot 0{,}38 + (3 - 4{,}75)^2 \cdot 0{,}25 + (5 - 4{,}75)^2 \cdot 0{,}2 +$$

$$+ (10 - 4{,}75)^2 \cdot 0{,}1 + (20 - 4{,}75)^2 \cdot 0{,}05 + (50 - 4{,}75)^2 \cdot 0{,}02 =$$

=8,57375+0,76525+0,0125+2,75625+11,628125+40,95125=64,6875.

2). $$D(X) = M(X^2) - \left[M(X)\right]^2.$$

$M(X^2) = \sum x_i^2 \cdot p_i = 0^2 \cdot 0{,}38 + 3^2 \cdot 0{,}25 + 5^2 \cdot 0{,}20 + 10^2 \cdot 0{,}1 + 20^2 \cdot 0{,}05 + \quad + 50^2 \cdot 0{,}02 =$
0+2,25+5+10+20+50 = 87,25.

Тогда:

$$D(X) = 87{,}25 - (4{,}75)^2 = 87{,}25 - 22{,}5625 = 64{,}6875.$$

Результаты вычислений дисперсии по обоим способам совпадают.

3) Среднее квадратичное отклонение:

$$\sigma(X) = \sqrt{D(X)} = \sqrt{64{,}6875} \approx 8{,}04285.$$

Таким образом, $\boldsymbol{\sigma}$ = 8,04285 тыс. рублей – характеристика разброса фактических значений выигрыша от найденного среднего значения $\boldsymbol{a}$ = 4,75 тыс. рублей. Это означает, что основные значения случайной величины выигрыша находятся в диапазоне $(4{,}75 \pm 8{,}04285)$ тыс. рублей, что соответствует таблице данных.

Математическая статистика

Математическая статистика базируется на теории вероятностей и является теоретической основой для всей статистики. Ее задачей является создание способов сбора и методов обработки статистической информации.

Обработка выборочных данных

Статистическое распределение выборки

Выборочный метод – один из основных методов математической статистики. Его сущность заключается в том, что изучение большой совокупности объектов относительно некоторого количественного признака X производится по сравнительно небольшому числу случайно отобранных объектов.

Генеральной совокупностью называется множество всех изучаемых объектов, из которых производится выборка.

***Выборочной совокупностью* (*выборкой*)** называется множество объектов, отобранных для изучения из генеральной совокупности. Выборка должна быть организована случайным образом, чтобы правильно представлять генеральную совокупность.

Объемом совокупности называется количество объектов в совокупности. Объем выборки $\boldsymbol{n}$, как правило, значительно меньше объема $\boldsymbol{N}$ генеральной совокупности: $\boldsymbol{n} << \boldsymbol{N}$.

Данные выборки записываются в виде таблицы, называемой статистическим распределением выборки:

x_i	x_1	x_2	…	x_k
n_i	n_1	n_2	…	n_k

В первой строке перечислены все наблюдаемые значения признака X в порядке их возрастания (или убывания). Они называются вариантами x_i (i = 1, 2, 3, …, k). Во второй строке указаны частоты n_i соответствующих вариант x_i , они показывают, сколько раз наблюдалось каждое значение признака X.

Очевидно, что сумма всех частот n_i равна объему выборки n:

$$n_1 + n_2 + \ldots + n_k = \sum_{i=1}^{k} n_i = n.$$

Основные числовые характеристики выборки

Средняя выборочная (среднее взвешенное значение признака в выборке):

$$\bar{x}_в = \frac{\sum_{i=1}^{к} x_i n_i}{n} = \frac{1}{n} \cdot (x_1 n_1 + x_2 n_2 + ... + x_k n_k).$$

Дисперсия выборочная. Характеризует разброс (рассеяние) значений вариант x_i от выборочного среднего значения $\overline{x_в}$ и измеряется в квадратных единицах признака *X*:

$$D_в = \frac{1}{n} \cdot \sum_{i=1}^{k} (x_i - \bar{x}_в)^2 n_i = \frac{1}{n} \left[(x_1 - \bar{x}_в)^2 n_1 + (x_2 - \bar{x}_в)^2 n_2 + ... + (x_к - \bar{x}_в)^2 n_k \right].$$

Для вычисления дисперсии используется также другая, часто более удобная формула:

$$D_в = \overline{x_в^2} - (\bar{x}_в)^2,$$

где

$$\overline{x_в} = \frac{1}{n} \cdot \sum_{i=1}^{k} x_i n_i; \qquad \overline{x_в^2} = \frac{1}{n} \cdot \sum_{i=1}^{k} x_i^2 n_i .$$

Среднее квадратическое отклонение выборки – это характеристика рассеяния значений признака x_i в выборке от среднего выборочного значения $\overline{x_в}$ в единицах признака *X:*

$$\sigma_в = \sqrt{D_в} .$$

Точечные оценки генеральных характеристик

С помощью найденных выборочных характеристик $x_в, D_в, \sigma_в$ оцениваются соответствующие генеральные характеристики:

$\bar{X}$ – генеральная средняя;

D – генеральная дисперсия;

σ – генеральное среднее выборочное отклонение.

Точечными называются оценки с помощью числа. Они имеют следующий вид:

$$\bar{x} \approx \overline{x_в}\ ; \quad D \approx \frac{n}{n-1} \cdot D_в = {S_в}^2, \quad \sigma \approx S_в = \sqrt{\frac{n}{n-1} D_в}\ ,$$

где $S_в^2$ - так называемая исправленная выборочная дисперсия.

Приведенные точечные оценки носят случайный характер, так как зависят от выборки. Эти оценки удовлетворяют следующим требованиям.

Несмещённость, означает отсутствие систематических ошибок, то есть, нет отклонений только в одну сторону от истинного значения.

Состоятельность, означает, что при увеличении объема выборки увеличивается вероятность того, что оценка будет более точной.

Эффективность, означает, что данная оценка имеют самый незначительный разброс по сравнению с другими возможными несмещёнными оценками.

Интервальные оценки генеральных характеристик

Точечные оценки генеральных характеристик являются приближенными, причём точность их приближения неизвестна. Эти оценки могут оказаться далекими от истинных значений характеристик генеральной совокупности: $\bar{x} = a$, *D,* σ. Поэтому для оценки генеральных характеристик используются также интервальные оценки, когда неизвестная характеристика заключена в некотором интервале с заданной надежностью (вероятностью) **γ**. Такой интервал называется доверительным. Значения надежности берутся, как правило, высокими: 0,9; 0,95; 0,99 или 0,999, что соответствует 90; 95; 99 или 99,9%.

Если количественный признак *X* в генеральной совокупности распределен по нормальному закону, причем среднее квадратическое отклонение **σ** этого распределения известно, то с вероятностью **γ** доверительный интервал, заданный выражением

$$(\overline{x_в} - t \cdot \sigma / \sqrt{n}; \quad \overline{x_в} + t \cdot \sigma / \sqrt{n}),$$

покрывает неизвестное математическое ожидание ***a***. Здесь вспомогательный параметр ***t*** находится из соотношения $2\Phi(t) = \gamma$ с помощью таблицы для интегральной функции Лапласа *Φ*(*t*) (см. приложение 2).

Графическое представление выборочных данных

Значения n_i называются абсолютными частотами, их сумма равна объему выборки: $\sum n_i = n$. Относительные частоты $w_i = \frac{n_i}{n}$ показывают

долю значений x_i в общем объеме выборки. Очевидно, что сумма всех относительных частот (долей) равна 1: $\sum w_i = 1$.

Графически дискретное статистическое распределение изображается в виде полигона частот, обычно относительных. ***Полигон*** частот представляет собой ломаную линию, соединяющую соседние точки с координатами (x_i ; w_i).

Статистическое распределение выборки часто носит интервальный характер. В этом случае указывают числовые частичные интервалы, куда попадают значения признака *X*, и n_i – количество значений, попавших в каждый частичный интервал.

Интервальное статистическое распределение изображается на графике в виде гистограммы относительных частот. ***Гистограмма*** – это ступенчатая фигура, состоящая из прямоугольников (рис.3). В основании каждого прямоугольника лежит частичный интервал, а высотой прямоугольника является относительная частота w_i, а чаще величина $\frac{w_i}{h_i}$, где h_i – длина частичного интервала. При таком построении площадь каждого частичного прямоугольника равна относительной частоте w_i, а сумма всех площадей, то есть площадь ступенчатой фигуры, равна единице, так как $\sum w_i = 1$.

Задача

В результате выборочного наблюдения за вкладами клиентов банка получено следующее распределение клиентов по величине вклада *X* т. р.:

X	До 100	100-200	200-300	300-400	400-500
n_i	10	18	20	32	28

где n_i – количество клиентов с величиной вклада в заданном интервале.

а) Изобразить данное распределение графически, построив гистограмму относительных частот.

б) Найти основные характеристики выборки: среднее значение, дисперсию и среднее квадратическое отклонение.

в) Оценить генеральные характеристики по найденным выборочным характеристикам.

г) С надежностью 95% указать доверительный интервал для генеральной средней, приняв гипотезу о нормальном распределении признака *X*, и считая, что генеральная дисперсия равна исправленной выборочной дисперсии.

Решение. Найдем объем выборки n:

$$n = \sum n_i = 10 + 18 + 20 + 32 + 28 = 108,$$

то есть для обследования выбрано 108 клиентов.

а) Вычислим относительные частоты для каждого частичного интервала:

$$w_1 = \frac{n_1}{n} = \frac{10}{108} = 0{,}093; \qquad w_2 = \frac{n_2}{n} = \frac{18}{108} = 0{,}167;$$

$$w_3 = \frac{n_3}{n} = \frac{20}{108} = 0{,}185; \qquad w_4 = \frac{n_4}{n} = \frac{32}{108} = 0{,}296;$$

$$w_5 = \frac{n_5}{n} = \frac{28}{108} = 0{,}259\,.$$

Рекомендуем все вычисления вести с точностью до 0,001.

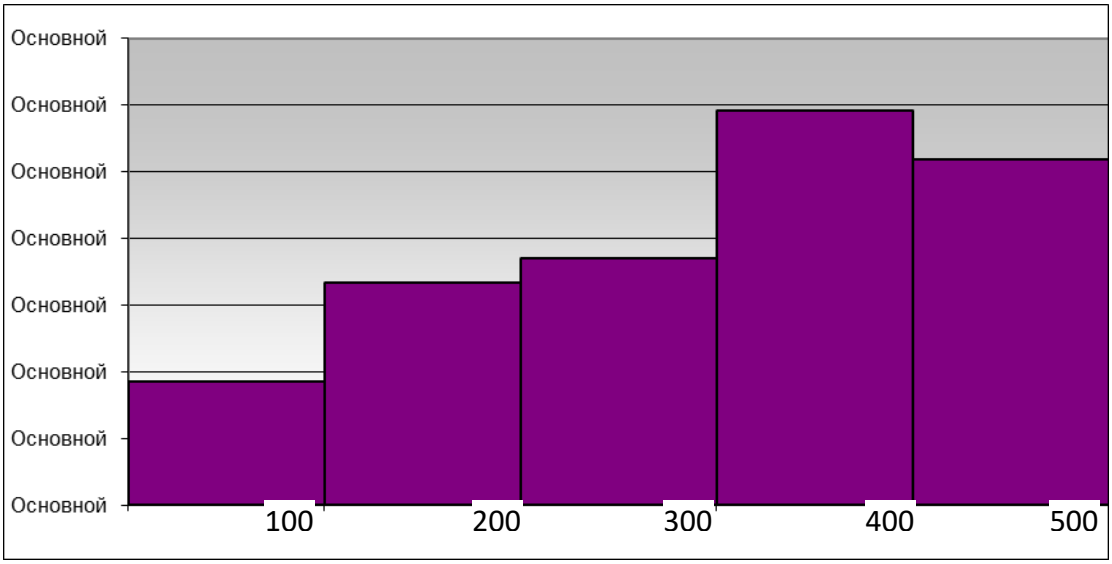
Контроль: $\sum w_i = 0{,}093 + 0{,}167 + 0{,}185 + 0{,}296 + 0{,}259 = 1$.

В итоге получено следующее интервальное распределение относительных частот признака *X:*

X	0-100	100-200	200-300	300-400	400-500
w_i	0,093	0,167	0,185	0,296	0,259

Шаг разбиения *h* это длина каждого частичного интервала: *h*=100. Строим гистограмму относительных частот (Рис. 3).

На графике по горизонтальной оси *OX* отложены частичные интервалы для признака *X*, а по вертикальной оси – значения относительных частот w_i.



б) Для нахождения характеристик выборки от заданного интервального распределения признака *X* перейдем к дискретному распределению, выбирая в качестве значений признака x_i *середины* частичных интервалов:

x_i	50	150	250	350	450
n_i	10	18	20	32	28

Найдем основные характеристики этого распределения.

Средняя выборочная (средняя величина вклада в т. р.):

$$\overline{x_в} = \frac{1}{n} \cdot \sum x_i n_i = \frac{1}{108}(50 \cdot 10 + 150 \cdot 18 + 250 \cdot 20 + 350 \cdot 32 + 450 \cdot 28) =$$

$$= \frac{1}{108}(500 + 2700 + 5000 + 11200 + 12600) = \frac{1}{108} \cdot 32000 \approx 296{,}296.$$

Выборочная дисперсия:

$$D_в = \frac{1}{n}\sum (x_i - \overline{x_в})^2 \cdot n_i = \frac{1}{108} \cdot \left[(50 - 296{,}296)^2 \cdot 10 + (150 - 296{,}296)^2 \cdot 18 + \right.$$

$$\left. + (250 - 296{,}296)^2 \cdot 20 + (350 - 296{,}296)^2 \cdot 32 + (450 - 296{,}296)^2 \cdot 28\right] =$$

$$= \frac{1}{108}(606617{,}196 + 385245{,}353 + 42866{,}392 + 92291{,}828 +$$

$$+ 661497{,}749) = \frac{1}{108} \cdot 1788518{,}518 = 16560{,}3566.$$

Второй способ вычисления дисперсии.

Найдем среднее квадратов значений признака:

$$\overline{x_в^2} = \frac{1}{n} \cdot \sum x_i^2 n_i = \frac{1}{108}(50^2 \cdot 10 + 150^2 \cdot 18 + 250^2 \cdot 20 + 350^2 \cdot 32 + 450^2 \cdot 28) =$$

$$= \frac{1}{108}(25000 + 405000 + 1250000 + 3920000 + 5670000) =$$

$$= \frac{1}{108} \cdot 11270000 \approx 104351{,}852\ ,$$ тогда

$$D_в = \overline{x_в^2} - \left(\overline{x_в}\right)^2 = 104351{,}852 - (296{,}296)^2 =$$

$$= 104351{,}8519 - 87791{,}4952 = 16560{,}3566.$$

Этот результат должен совпадать с результатом первого способа (иногда приближенно из-за округлений).

Выборочное среднее квадратическое отклонение:

$$\sigma_в = \sqrt{Dв} = \sqrt{16560{,}3566} \approx 128{,}687,$$

то есть, в среднем разброс вкладов составляет ± 128,687 тыс. рублей от среднего значения 296,296 тыс. рублей.

в) Оценим неизвестные генеральные характеристики.

Генеральное среднее значение: $\overline{x} \approx \overline{x_в} = 296{,}296$ т. р.

Генеральная дисперсия:

$$D \approx \frac{n}{n-1} D_в = \frac{108}{108-1} \cdot 16560{,}3566 \approx 16715{,}1263.$$

Генеральное среднее квадратическое отклонение:

$$\sigma = \sqrt{D} = \sqrt{16715{,}1263} \approx 129{,}287 \text{ т. р.}$$

г) Доверительный интервал для оценки генеральной средней ***a*** (среднего вклада) с надежностью γ находим по формуле:

$$a \in (\overline{x_в} - t \cdot \sigma / \sqrt{n};\quad \overline{x_в} + t \cdot \sigma / \sqrt{n}).$$

По условию задачи $n = 108;\ \overline{x_в} = 296{,}296;\ \sigma = 129{,}287,\ \gamma = 0{,}95$. Неизвестный параметр t находим из условия: $2\Phi(t) = \gamma$. Поскольку в данной задаче $\gamma = 0{,}95$, то есть $2\Phi(t) = 0{,}95$, то $\Phi(t) = 0{,}475$. По таблице Приложения 2, находим $t = 1{,}96$.

Вычислим по этим данным доверительный интервал:

$$\left(296{,}296 - 1{,}96 \cdot \frac{129{,}287}{\sqrt{108}};\ \ 296{,}296 + 1{,}96 \cdot \frac{129{,}287}{\sqrt{108}}\right),$$

$$(296{,}296 - 24{,}384;\ \ 296{,}296 + 24{,}384),$$

$$(271{,}912;\ \ 320{,}680).$$

Таким образом, с вероятностью 95% неизвестная генеральная средняя (математическое ожидание) находится в этом интервале:

$$\overline{x} = a \in (271{,}912;\ \ 320{,}680).$$

Длина полуинтервала $\delta = t \cdot \sigma / \sqrt{n} = 24{,}384$ характеризует точность оценки и называется предельной ошибкой оценки. Оценка тем точнее, чем меньше δ и, следовательно, доверительный интервал становится более узким.

Величина предельной ошибки δ зависит от n, σ и t. Очевидно, что с увеличением объема выборки n предельная ошибка δ уменьшается и, следовательно, точность оценки повышается. При увеличении рассеяния σ предельная ошибка δ увеличивается, то есть оценка делается менее точной. Увеличение надежности γ ведет к росту вспомогательного параметра t и расширению доверительного интервала (надежнее попасть в большой интервал). Это делает оценку менее точной. Таким образом, при повышении надежности оценки ухудшается ее точность.

***Ответы*:** $\overline{x_в} = 296{,}296$; $D_в = 16560{,}3566$; $\sigma_в = 128{,}683$.

$\bar{x} \approx 296{,}296$; $D \approx 16715{,}1263$; $\sigma \approx 129{,}287$.

$\bar{x} = a \in (271{,}912;\ 320{,}680)$ с надёжностью 95%.

Проверка статистических гипотез. χ^2 - критерий Пирсона

Назначения критерия

Критерий χ^2 применяется в двух целях;

1) для сопоставления *эмпирического* распределения признака с *теоретическим* - равномерным, нормальным или каким-то иным;
2) для сопоставления *двух, трех или более эмпирических* распределений одного и того же признака.

Описание критерия

Критерий χ^2 отвечает на вопрос о том, с одинаковой ли частотой встречаются разные значения признака в эмпирическом и теоретическом распределениях или в двух и более эмпирических распределениях.

При сопоставлении эмпирического распределения с теоретическим мы определяем степень расхождения между эмпирическими и теоретическими частотами.

Чем больше расхождение между двумя сопоставляемыми распределениями, *тем больше* эмпирическое *значение* χ^2.

Гипотезы **(**для сопоставления *эмпирического* распределения признака с *теоретическим***)**

H_0: Полученное эмпирическое распределение признака не отличается от теоретического (например, нормального) распределения.

H_1: Полученное эмпирическое распределение признака отличается от теоретического распределения.

Ограничения критерия

1. Объем выборки должен быть достаточно большим: $n \geq 30$. При $n < 30$ критерий χ^2 дает весьма приближенные значения. Точность критерия повышается при больших n.
2. Теоретическая частота для каждой ячейки таблицы не должна быть меньше 5: $n'_i \geq 5$. Это означает, что если число разрядов задано заранее и не может быть изменено, то мы не можем применять метод χ^2, не накопив определенного минимального числа наблюдений.
3. Выбранные разряды должны «вычерпывать» все распределение, то есть охватывать весь диапазон вариативности признаков. При этом группировка на разряды должна быть одинаковой во всех сопоставляемых распределениях.
4. Разряды должны быть неперекрещивающимися: если наблюдение отнесено к одному разряду, то оно уже не может быть отнесено ни к какому другому разряду.
5. Сумма наблюдений по разрядам всегда должна быть равна общему количеству наблюдений.

Алгоритм расчета критерия χ^2

1. Занести в таблицу данные статистического распределения: наименования разрядов и соответствующие им эмпирические частоты (n_i).
2. Рядом с каждой эмпирической частотой записать теоретическую частоту (n'_i).
3. Подсчитать разности между эмпирической и теоретической частотой по каждому разряду (строке) и записать их в третий столбец.
4. Определить число степеней свободы по формуле: $v=k-3$, где k – количество разрядов признака (при сравнении с нормальным распределением).
5. Возвести в квадрат полученные разности и занести их в четвертый столбец.
6. Разделить полученные квадраты разностей на теоретическую частоту и записать результаты в пятый столбец.
7. Просуммировать значения пятого столбца. Полученную сумму обозначить как $\chi^2_{эмп}$

$$\chi^2 = \sum_{j=1}^{k} \frac{(n_j - n'_j)^2}{n'_j}$$

8. Определить по таблице критических точек распределения χ^2 для данного числа степеней свободы v $\chi^2_т$. Если $\chi^2_{эмп}$ меньше критического значения, расхождения между распределениями статистически недо-

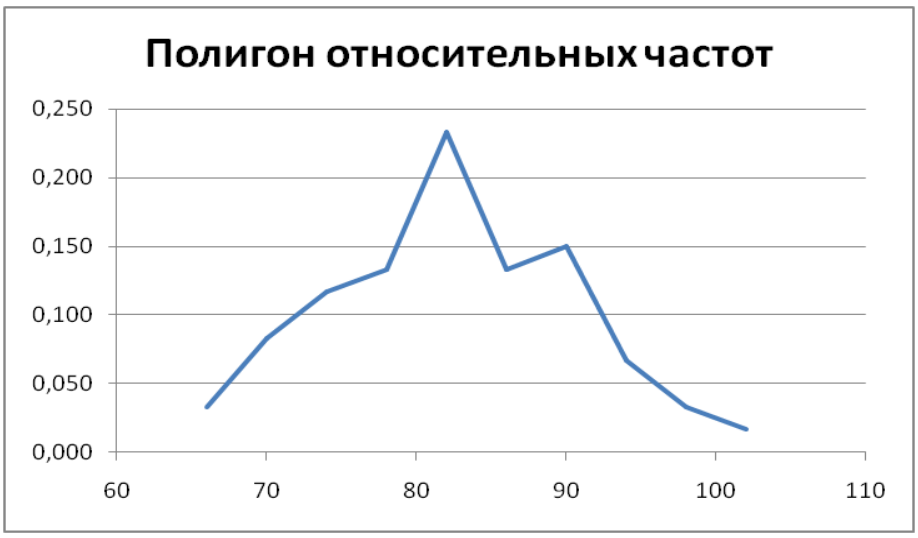
стоверны. Различия между двумя распределениями могут считаться достоверными, если $\chi^2_{эмп}$ достигает или превышает $\chi^2_{0.05}$- и тем более достоверными, если $\chi^2_{эмп}$ достигает или превышает $\chi^2_{0.01}$.

Критические значения $\chi^2_{теор}$ могут быть найдены с помощью стандартной функции Ms Excel: ХИ2ОБР(вероятность; число степеней свободы).

Пример нахождения точечных и интервальных оценок с помощью MS Excel

86	81	76	80	84	85			
95	77	85	95	89	83			
82	77	76	71	87	68			
89	64	81	90	72	97			
91	75	80	79	85	83			
78	94	87	103	70	87			
90	70	82	99	81	89			
84	79	78	74	81	75			
81	76	73	81	89	93			
89	85	83	92	84	72			
X_{min}=	64	64						
X_{max}=	103	104						
R=	39	40						
h=		4						
$X_л$	$X_{пр}$	X_c	n_i	□$_i$	$X_c n_i$	$X^2_c n_i$	$X^3_c n_i$	$X^4_c n_i$
64	68	66	2	0,033	132	8712	574992	37949472
68	72	70	5	0,083	350	24500	1715000	120050000
72	76	74	7	0,117	518	38332	2836568	209906032
76	80	78	8	0,133	624	48672	3796416	296120448
80	84	82	14	0,233	1148	94136	7719152	632970464
84	88	86	8	0,133	688	59168	5088448	437606528
88	92	90	9	0,150	810	72900	6561000	590490000
92	96	94	4	0,067	376	35344	3322336	312299584
96	100	98	2	0,033	196	19208	1882384	184473632

100	104	102	1	0,017	102	10404	1061208	108243216
			60		**4944**	**411376**	**3,5E+07**	**2,93E+09**



$X_{выб}$=	82,40	M_1=	82,4	m_3=	41,728		
$D_{выб}$=	66,51	M_2=	6856,27	m_4=	11176		
S^2=	67,63	M_3=	575958,4				
$\sigma_{выб}$=	8,16	M_4=	48835156,3				
S=	8,22						
A_s=	0,08						
E_s=	-0,47						
γ=	0,90						
t=	1,64		Доверительный интервал				
δ=	1,75	(	80,65	84,15	)		
γ=	0,95						
t=	1,96						
δ=	0,50						
N=	1039,25						

Минимальный объем выборки, обеспечивающий заданную точность: 1040

Замечание. Для нахождения абсолютных частот в Ms Excel можно воспользоваться функцией =ЧАСТОТА(A1:F10;B18:B27), в которой первый аргумент – блок в котором располагаются исходные данные, второй аргумент – правые границы частичных интервалов. Функция матричная, для корректного её использования необходимо воспользоваться следующим алгоритмом: а) выделяется блок, в котором должны располагаться результаты; б) вводится формула; в) завершается операция нажатием [Ctrl]+[Shift]+[Enter]. Найти решение уравнения $\Phi(t) = \gamma/2$ относительно переменной t можно найти по таблице или воспользовавшись функцией
=НОРМОБР(0,5+НАДЕЖНОСТЬ/2;0;1), где НАДЕЖНОСТЬ – значение γ.

Проверка статистических гипотез. χ^2 - критерий Пирсона

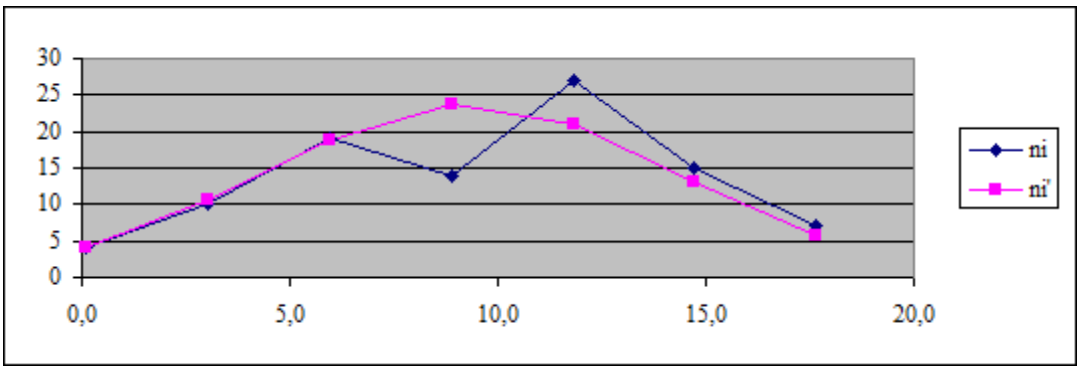
=НОРМРАСП(F32;0;1;0)

x_i	x_{i+1}	x_c	n_i	u_i	$\varphi(u_i)$	n_i'	$((n'-n)^2)/n'$
-1,34	1,58	0,12	4	-1,873	0,069	4,129	0,004
1,58	4,50	3,04	10	-1,278	0,176	10,541	0,028
4,50	7,42	5,96	19	-0,683	0,316	18,887	0,001
7,42	10,35	8,88	14	-0,088	0,397	23,758	4,008
10,35	13,27	11,81	27	0,507	0,351	20,978	1,729
13,27	16,19	14,73	15	1,101	0,218	13,004	0,307
16,19	19,11	17,65	7	1,696	0,095	5,658	0,318
19,11	22,03	20,57	4	2,291	0,029	1,728	2,986
			100			$\chi^2_э$=	**9,379**

Гипотезы

H_0: Полученное эмпирическое распределение признака не отличается от нормального распределения с параметрами a σ.

H_1: Полученное эмпирическое распределение признака отличается от нормального распределения.



Число степеней свободы ν=k-1-s	ν=	5	=ХИ2ОБР(F42;F41)
Уровень значимости	α=	0,01	$\chi^2_т$= 15,086
	α=	0,05	$\chi^2_т$= 11,0705

$\chi^2_э < \chi^2_т$ – следовательно нет оснований отвергнуть гипотезу H_0.

Элементы теории корреляции

Виды зависимостей

Пусть каждый из рассматриваемых объектов характеризуется двумя признаками *X* и *Y*. Между этими признаками *X* и *Y* могут существовать различные виды зависимостей.

Функциональная зависимость – это «жёсткая» (детерминированная) зависимость, когда каждому значению признака *X* соответствует *единственное* значение признака *Y*. Зависимость задается в виде функции $y = f(x)$.

Статистическая зависимость – в этом случае каждому значению признака *X* соответствует статистическое распределение признака *Y*. Эта зависимость задается в виде корреляционной таблицы.

Корреляционная зависимость – это частный случай статистической зависимости, когда каждому значению *x* признака *X* соответствует среднее значение $\overline{y}_x$ признака *Y* и связь между ними достаточно хорошо описывается функцией $\overline{y}_x = f(x)$, которая называется функцией регрессии *Y* по *X*.

Аналогично, если каждому значению признака *Y* соответствует среднее значение $\overline{x}_y = \varphi(y)$, то последняя функция называется функцией регрессии *X* по *Y*.

Корреляционная зависимость между признаками, это зависимость между средними значениями этих признаков.

Корреляционная зависимость между признаками может проявляться с разной степенью силы.

Две основные задачи теории корреляции:

1) оценить силу (тесноту) связи между признаками *X* и *Y*;
2) найти вид (форму) этой связи в виде уравнения регрессии.

Наиболее простой и употребляемый вид связи – линейная связь. Она задается уравнением линейной регрессии $\overline{y}_x = a \cdot x + b$ и изображается на графике в виде прямой регрессии.

Оценка тесноты линейной связи

Оценка тесноты линейной связи между признаками *X* и *Y* производится с помощью коэффициента линейной корреляции ***r***:

$$r = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sigma_x \cdot \sigma_y}.$$

Коэффициент r может принимать значения от –1 до +1 включительно:

$$-1 \le \boldsymbol{r} \le 1 \quad \text{или} \quad |\boldsymbol{r}| \le 1.$$

Знак r указывает направление связи: прямая или обратная. Абсолютная величина $|\boldsymbol{r}|$ указывает на силу (тесноту) связи и устанавливается по приведённой ниже шкале.

Шкала Чаддока

Значение $\lvert\mathbf{r}\rvert$	0-0,1	0,1-0,3	0,3-0,5	0,5-0,7	0,7-0,9	0,9-0,99	1
Теснота линейной связи	Связи нет	Слабая	Умеренная	Заметная	Высокая	Очень высокая	Функциональная

При $\boldsymbol{r} > 0$ связь ***прямая***, то есть с ростом $\boldsymbol{x}$ растет $\boldsymbol{y}$.
При $\boldsymbol{r} < 0$ связь ***обратная***, то есть с ростом $\boldsymbol{x}$ убывает $\boldsymbol{y}$.

Уравнение линейной регрессии

Уравнение регрессии показывает, как средние значения одного признака ***Y*** зависят от значений другого признака ***X***. Часто эта зависимость является линейной:

$$\overline{y}_x = ax + b\,.$$

Параметры ***a*** и ***b*** в уравнении линейной регрессии находятся по методу наименьших квадратов, который приводит к следующим формулам для их вычисления:

$$a = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sigma_x^{\,2}};$$

$$b = \overline{y} - a \cdot \overline{x}.$$

Задача

С целью анализа взаимного влияния прибыли предприятия и его издержек выборочно были проведены наблюдения за этими показателями в течение ряда месяцев: X – величина месячной прибыли в т. р., Y – месячные издержки в процентах к объему продаж. Результаты выборки представлены в виде таблицы:

X	50	56	60	62	65	72	78
Y	20,4	18,1	15,2	10,6	8,8	8,7	8,0

По данным выборки:

а) оценить тесноту линейной связи между признаками X и Y;

б) найти зависимость между признаками в виде уравнения линейной регрессии $\overline{y}_x = a \cdot x + b$;

в) построить графически наблюдаемые выборочные значения признаков и прямую регрессии.

г) Используя уравнение линейной регрессии, спрогнозировать величину месячных издержек в процентах к объему продаж, если величина месячной прибыли будет составлять 82 т. р.

Решение. По условию имеется $\boldsymbol{n}$ = 7 наблюдений для соответственных значений признаков X и Y.

Найдем средние значения признаков $\overline{x}$ и $\overline{y}$, а также их средние квадратические отклонения σ_x и σ_y по тем же формулам, что и в предыдущей задаче, но с учетом того, что каждое значение признака встречается только один раз, то есть все $\boldsymbol{n_i}$ = 1.

Вычисления будем вести с точностью до 0,001.

$$\bar{x} = \frac{\sum x_i n_i}{n} = \frac{\sum x_i}{n} = \frac{1}{7}(50+56+60+62+65+72+78) = \frac{443}{7} \approx 63{,}286;$$

$$\bar{y} = \frac{\sum y_i n_i}{n} = \frac{\sum y_i}{n} = \frac{1}{7}(20{,}4+18{,}1+15{,}2+10{,}6+8{,}8+8{,}7+8{,}0) =$$

$$= \frac{89{,}80}{7} \approx 12{,}829;$$

$$\overline{xy} = \frac{\sum x_i y_i}{n} = \frac{1}{7}(50\cdot 20{,}4+56\cdot 18{,}1+60\cdot 15{,}2+62\cdot 10{,}6+65\cdot 8{,}8+$$

$$+72\cdot 8{,}7+78\cdot 8{,}0) = \frac{1}{7}\cdot 5425{,}2 \approx 775{,}029;$$

$$\overline{x^2} = \frac{\sum x_i^2}{n} = \frac{1}{7}(50^2+56^2+60^2+62^2+65^2+72^2+78^2) \approx 4081{,}857;$$

$$\overline{y^2} = \frac{\sum y_i^2}{n} = \frac{1}{7}(20{,}4^2+18{,}1^2+15{,}2^2+10{,}6^2+8{,}8^2+8{,}7^2+8{,}0^2) =$$

$$= \frac{1304{,}3}{7} \approx 186{,}329;$$

$$\sigma_x = \sqrt{\overline{x^2} - \bar{x}^2} = \sqrt{4081{,}857 - 63{,}286^2} = \sqrt{76{,}739} \approx 8{,}760;$$

$$\sigma_y = \sqrt{\overline{y^2} - \bar{y}^2} = \sqrt{186{,}329 - 12{,}829^2} = \sqrt{21{,}746} \approx 4{,}663.$$

а) Оценим тесноту линейной связи по коэффициенту линейной корреляции ***r***:

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y} = \frac{775{,}029 - 63{,}286 \cdot 12{,}829}{8{,}760 \cdot 4{,}663} \approx -0{,}9025 \approx -0{,}9.$$

Так как $\boldsymbol{r} < 0$, то связь обратная, то есть с ростом значений признака X значения признака Y убывают.

Так как $|\boldsymbol{r}| = |-0{,}9| = 0{,}9$, то по шкале Чаддока, приведенной выше, определяем, что линейная связь очень высокая.

б) Найдем уравнение линейной регрессии. Его параметры:

$$a = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x^{\,2}} = \frac{775{,}029 - 63{,}286 \cdot 12{,}829}{76{,}739} \approx -0{,}48;$$

$$b = \bar{y} - a \cdot \bar{x} = 12{,}829 - (-0{,}48) \cdot 63{,}286 \approx 43{,}193.$$

В результате получим, что среднее значение издержек $\bar{y}_x$ связано с величиной прибыли x уравнением:

$$\bar{y}_x = -0{,}48x + 43{,}193.$$

в) Изобразим графически данные значения (x_i; y_i) в виде точек на плоскости xOy (Рис. 4).

Прямую регрессии $y = -0{,}48x + 43{,}193$ строим по двум точкам: $x = 0$; $y = -0{,}48 \cdot 0 + 43{,}193 \approx 43$.

$x = 80$; $y = -0{,}48 \cdot 80 + 43{,}193 = 43{,}193 - 38{,}4 \approx 4{,}8$.

Получены точки (0; 43) и (80; 4,8).

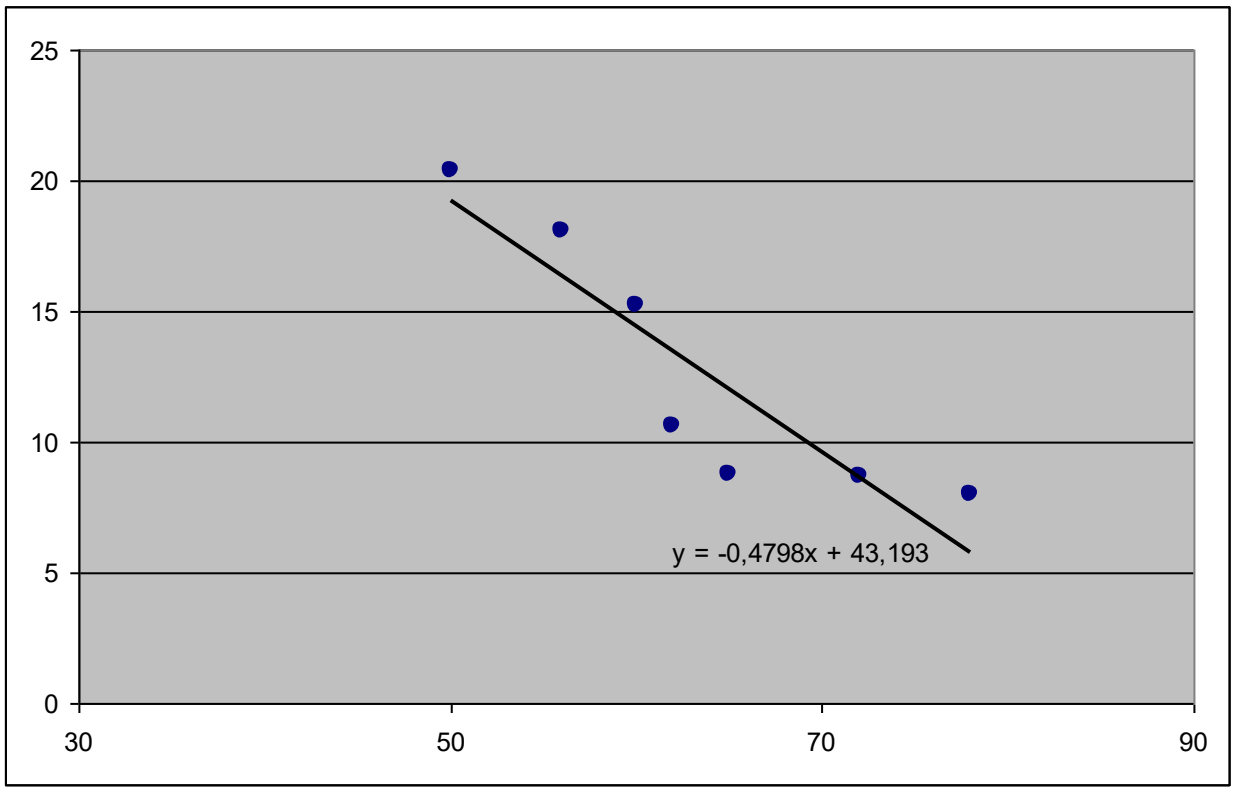


Рис. 4

На графике прямая регрессии убывает и проходит через точку $A(\bar{x};\bar{y})$, то есть (63,3; 12,8). Прямая регрессии наилучшим образом приближена ко всем данным точкам, которые расположены вблизи прямой по обе стороны от нее.

г) Используя найденную зависимость, спрогнозируем величину месячных издержек, если месячная прибыль составит 82 тыс. руб.:

$y = -0{,}48 \cdot 82 + 43{,}193 \approx 3{,}8$, то есть 3,8% к объёму продаж.

Ответ. Корреляционная зависимость между признаками X и Y очень высокая, ее можно описать линейным уравнением регрессии:

$$\boldsymbol{y} = -0{,}48\boldsymbol{x} + 43{,}193.$$

Прогнозируемые издержки составят 3,8% к объёму продаж.

Список рекомендуемой литературы
Основная учебная литература

1. Теория вероятностей и математическая статистика : учебник / Е.С. Кочетков, С.О. Смерчинская, В.В. Соколов. – 2-е изд., перераб. и доп. – М. : ФОРУМ : ИНФРА-М, 2017. – 240 с. – (Среднее профессиональное образование). - Режим доступа: http://znanium.com/catalog/product/760157.

2. Теория вероятностей и математическая статистика : учебник / Е.С. Кочетков, С.О. Смерчинская, В.В. Соколов. – 2-е изд., испр. и перераб. – М. : ФОРУМ : ИНФРА-М, 2018. – 240 с. – (Среднее профессиональное образование). - Режим доступа: http://znanium.com/catalog/product/944923.

3. Теория вероятностей и математическая статистика : учебник / Е.А. Коган, А.А. Юрченко. – Москва : ИНФРА-М, 2019. – 250 с. – (Высшее образование: Бакалавриат). –www.dx.doi.org/10.12737/textbook_5cde54d3671a96.35212605. - Текст : электронный. - URL: http://znanium.com/catalog/product/971766.

Дополнительная учебная литература

4. КОМИССАРОВ ВАЛЕНТИН ВЛАДИСЛАВОВИЧ. Теория вероятностей и математическая статистика : методические указания / КОМИССАРОВ ВАЛЕНТИН ВЛАДИСЛАВОВИЧ ; АНОО ВО Центросоюза РФ СибУПК. – Новосибирск, 2019. – 40с.– электронный ресурс.