Учебное пособие

Т. В. Ашихмина

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА В ПРИМЕРАХ И ЗАДАЧАХ

Киров
2010

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
ВЯТСКИЙ ГОСУДАРСТВЕННЫЙ ГУМАНИТАРНЫЙ УНИВЕРСИТЕТ

ФАКУЛЬТЕТ ИНФОРМАТИКИ

Т. В. Ашихмина

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА В ПРИМЕРАХ И ЗАДАЧАХ

Учебное пособие

Киров
2010

УДК 517.16
ББК 22.161
А17

Печатается по решению редакционно-издательского совета Вятского государственного гуманитарного университета

Рецензенты:

С. М. Окулов, доктор педагогических наук, профессор кафедры информатики и методики обучения информатике, декан факультета информатики Вятского государственного гуманитарного университета;

А. Н. Рапопорт, доктор физико-математических наук, профессор, заведующий кафедрой прикладной математики Вятского государственного университета.

А17 Ашихмина, Т. В. Теория вероятностей и математическая статистика в примерах и задачах: учебное пособие / Т. В. Ашихмина. – Киров: Изд-во ВятГГУ, 2010. – 108 с.

ISBN 978-5-93825-840-2

В учебном пособии освещены вопросы: случайные события и их вероятности, случайные величины и их характеристики и распределения, вариационные ряды, выборочное наблюдение, критерии согласия, дисперсионный, корреляционно-регрессионный и кластерный анализы.

Пособие включает основные понятия, примеры решения задач, вопросы и тесты для самоконтроля, задания для практического выполнения.

Учебное пособие предназначено для студентов высших учебных заведений, обучающихся по специальностям 010501.65 Прикладная математика и информатика, 080801.65 Прикладная информатика (по областям), 050202.65 Информатика, а также может быть использовано учащимися старших классов и учителями математики школ с углубленным изучением математики.

УДК 517.16
ББК 22.161

ISBN 978-5-93825-840-2

ОГЛАВЛЕНИЕ

Предисловие

Методы теории вероятностей широко применяются в различных отраслях естествознания и техники: в теории надежности, теории массового обслуживания, в теоретической физике, геодезии, астрономии, теории стрельбы, теории ошибок наблюдений, теории автоматического управления, общей теории связи и во многих других теоретических и прикладных науках. Теория вероятностей служит также для обоснования математической и прикладной статистики, которая в свою очередь используется при планировании и организации производства, при анализе технологических процессов, контроле качества продукции и для многих других целей.

Курс теории вероятностей и математической статистики помогает развитию теоретико-вероятностной интуиции, т. е. умению строить математические модели, правильно отражающие те или иные стороны реальных случайных явлений.

Пособие включает теоретические сведения, примеры решения задач, вопросы для самоконтроля и задания для практического выполнения и охватывает основные разделы курса «Теория вероятностей и математическая статистика».

Объем и глубина излагаемого материала варьируются в зависимости от специальности: более полно для специальностей 010501.65 Прикладная математика и информатика и 080801.65 Прикладная информатика (по областям), в сокращенном варианте – для специальности 050202.65 Информатика.

Глава 1. Случайные события и их вероятность

§ 1.1. Определения вероятностей

1.1.1. Основные понятия

Случайные события, пространство событий, алгебра событий. Классическое определение вероятности. Применение элементов комбинаторики к нахождению вероятности. Геометрические вероятности. Статистическое и аксиоматическое определение вероятности.

Классическое определение вероятности

Вероятностью события А называется отношение числа случаев, благоприятствующих этому событию, к общему числу случаев.

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

Комбинаторные формулы

Число перестановок из n элементов: $P_n = n!$.

Число размещений из n по k без повторений: $A_n^k = \frac{n!}{(n-k)!}$.

Число размещений из n по k с повторениями: $\widetilde{A}_n^k = n^k$.

Число сочетаний из n по k без повторений: $C_n^k = \frac{n!}{(n-k)! \cdot k!}$.

1.1.2. Примеры решения задач

1.1.2.1. Случайные события и действия над ними

Пример 1

Опыт – бросание игральной кости.

События: А – появление числа очков, кратного трем.

В – появление не более 4 очков.

С – появление 5 очков.

Является ли группа событий А, В и С полной? Можно ли эту группу назвать группой совместных событий? Есть ли среди этих событий равновозможные?

Решение

Событие А – появление 3 или 6 очков.

Событие В – появление 1, 2, 3 или 4 очков.

Событие С – появление 5 очков.

Группа является полной, так как каждый из элементарных исходов при данном опыте содержится хотя бы в одном из событий.

Группа событий является совместной, так как. совместными являются события А и В (исход «выпало 3 очка» является для этих событий общим благоприятным).

Равновозможных среди этих событий нет, так как p(A)=2/6=1/3, p(B)=4/6=2/3, p(C)=1/6.

Пример 2

Среди студентов, собравшихся на лекцию по теории вероятностей, выбирают наудачу одного. Пусть событие А заключается в том, что выбранный окажется юношей. Событие В в том, что он не курит, а событие С в том, что он живет в общежитии.

Задание:

a) Описать событие $AB\overline{C}$.

b) При каком условии будет иметь место тождество ABC = A?

c) Когда будет справедливо отношение $\overline{C} \subset B$?

d) Когда будет справедливо равенство $\overline{A} = B$?

Решение

a) Выбран юноша, который не живет в общежитии и не курит.

b) BC=U – все юноши живут в общежитии и не курят.

c) Все, не живущие в общежитии, не курят или курящие живут только в общежитии.

d) Все девушки не курят, а все юноши курят.

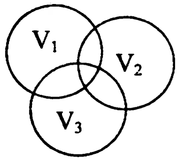
Пример 3

Опыт состоит в бросании точки в прямоугольник с изображенными областями V_1, V_2, V_3. События А, В и С означают соответственно попадание точки в области V_1, V_2, V_3. Изобразите возможную область попадания точки при следующих событиях:

a) A+B+C

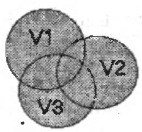
b) ABC

c) $\overline{A}+\overline{B}+C$

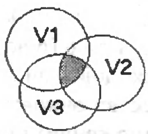


Решение

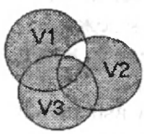
a)



b)

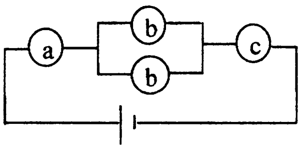


c)



Пример 4

Электрическая цепь составлена по следующей схеме. Выход из строя элемента a – это событие A, элементов b_k (k=1,2) – события B_k, элемента c – событие C. Запишите событие D – разрыв цепи и противоположное ему событие.



Решение

$D = A + B_1 \cdot B_2 + C$.

$\overline{D} = \overline{A} \cdot (\overline{B}_1 + \overline{B}_2) \cdot \overline{C}$.

Пример 5

Используя свойства операций, доказать равенство $A\overline{B} + B = A + B$.

Решение

1-й способ

$A \cdot \overline{B} + B = A + B$ (*умножим обе части равенства на* $\overline{B}$)

$(A \cdot \overline{B} + B) \cdot \overline{B} = (A + B) \cdot \overline{B}$

$A \cdot \overline{B} + B \cdot \overline{B} = A \cdot \overline{B} + B \cdot \overline{B}$

$A \cdot \overline{B} + V = A \cdot \overline{B} + V$

$A \cdot \overline{B} = A \cdot \overline{B}$ *получили верное тождество*

2-й способ

$$\overline{\overline{A \cdot \overline{B} + B}} = \overline{\overline{A \cdot \overline{B}} \cdot \overline{B}} = \overline{(\overline{A} + B) \cdot \overline{B}} = \overline{\overline{A} \cdot \overline{B} + B \cdot \overline{B}} = \overline{\overline{A + B}} = A + B.$$

1.1.2.2. Классическое определение вероятности

Пример 1

Игральная кость брошена три раза. Какова вероятность того, что при этом все выпавшие грани различны?

Решение

Число всех возможных исходов равно 6*6*6, так как при первом броске может выпасть любое количество очков: 1, или 2, или 3, или 4, или 5, или 6 (6 возможных исходов) и (умножение) при втором броске может выпасть любое количество очков: 1, или 2, или 3, или 4, или 5, или 6 (6 возможных исходов) и (умножение) при третьем броске может выпасть любое количество очков: 1, или 2, или 3, или 4, или 5, или 6 (6 возможных исходов).

Число всех благоприятных исходов равно 6*5*4, так как при первом броске может выпасть любое количество очков: 1, или 2, или 3, или 4, или 5, или 6 (6 возможных исходов) и (умножение) при втором броске может выпасть любое количество очков, кроме того, которое выпало при первом броске (5 возможных исходов) и (умножение) при третьем броске может выпасть любое количество очков, кроме тех, которые выпали при первом и при втором бросках (4 возможных исхода).

По классическому определению вероятность равна $\dfrac{6\cdot5\cdot4}{6\cdot6\cdot6}=\dfrac{10}{18}=\dfrac{5}{9}$.

Пример 2

В урне 6 белых и 4 черных шара. Из этой урны наудачу извлекли 5 шаров. Какова вероятность того, что два из них белые, а три – черные?

Решение

Число всех возможных исходов – это количество наборов из 10 элементов (всего шаров в урне 6+4=10) по 5 (5 шаров извлекли). Порядок элементов в данных наборах не важен (шары извлекаются не по одному и выкладываются в ряд, а сразу все пять), и повторений элементов в наборах быть не может (один шар дважды извлеченным быть не может). Поэтому количество таких наборов можно вычислить как число сочетаний без повторений из 10 по 5.

$$n=C_{10}^{5}=\frac{10!}{5!\cdot5!}=251.$$

Чтобы вычислить число благоприятных исходов, найдем число способов выбрать из всех белых шаров (а их 6) два и (умножение) из всех черных шаров (а их 4) три. Получаем $m=C_6^2\cdot C_4^3=\dfrac{6!}{4!\cdot2!}\cdot\dfrac{4!}{3!\cdot1!}=60$.

По классическому определению вероятность равна $\dfrac{m}{n}=\dfrac{60}{251}=\dfrac{5}{21}$.

Пример 3

10 человек случайным образом рассаживаются на 10-местную скамейку. Какова вероятность того, что два определенных лица окажутся рядом?

Решение

Число всех возможных исходов – это количество перестановок из 10 элементов, которое равно 10!. Чтобы вычислить число благоприятных исходов, найдем число способов посадить указанную пару рядом (таких способов 9: на 1-е и 2-е места, на 2-е и 3-е места, на 3-е и 4-е места, . . ., на 9-е и 10-е места), а затем (умножение) поменять местами людей из указанной пары между собой (2! способов) и (умножение) всеми способами поменять местами оставшихся 8 человек (8! способов, так как это число перестановок из 8 элементов). По классическому определению вероятность равна $\dfrac{m}{n}=\dfrac{9\cdot2!\cdot8!}{10!\cdot}=\dfrac{1}{5}$.

Пример 4

Из шести букв разрезной азбуки составлено слово «ананас». Ребенок, не умеющий читать, рассыпал эти буквы и затем собрал в произвольном порядке. Найти вероятность того, что у него снова получилось слово «ананас».

Решение

Число всех возможных исходов – это количество перестановок из 6 элементов (6 карточек с буквами), которое равно 6!. Не изменяя слова «ананас», можно переставить всеми возможными способами между собой буквы «а» (их 3, поэтому число перестановок равно 3!) и (умножение) переставить всеми возможными способами между собой буквы «н» (их 2, поэтому число перестановок равно 2!).

Поэтому число благоприятных исходов найдем как 2!*3!.

По классическому определению вероятность равна $\frac{m}{n} = \frac{2! \cdot 3!}{6!} = \frac{1}{60}$.

Пример 5

Полная колода карт (52) делится наугад на две равные части. Найдите вероятность следующих событий:

А – в каждой части окажется по два туза.

В – в одной из частей не будет ни одного туза.

С – в одной из частей будет ровно один туз.

Решение

Для всех событий число всех возможных исходов найдется как число наборов (неупорядоченных и без повторений) из 52 карт по 26 карт (52/2=26, т. е. число способов выбрать 26 карт из 52 в первую часть, а оставшиеся 26 карт будут считаться во второй части). Получаем n=C_{52}^{26}.

Для вычисления количества благоприятных исходов для события А необходимо подсчитать число способов выбрать два туза в первую группу (всего тузов в колоде 4, поэтому вычисляем C_4^2) и число способов выбрать 24 карты, не являющиеся тузами, в первую группу (всего карт, не являющихся тузами, в колоде 52-4=48, поэтому вычисляем C_{48}^{24}). Остальные не выбранные карты останутся во второй группе. Получаем p(A)=$\frac{m}{n} = \frac{C_4^2 \cdot C_{48}^{24}}{C_{52}^{26}}$.

Для вычисления количества благоприятных исходов для события В необходимо подсчитать число способов выбрать 26 карт, не являющихся тузами, в первую группу (всего карт, не являющихся тузами, в колоде 52-4=48, поэтому вычисляем C_{48}^{26}) и поменять местами 1-ю и 2-ю группы при каждом наборе (умножить на 2). Получаем p(A)=$\frac{m}{n} = \frac{2 \cdot C_{48}^{26}}{C_{52}^{26}}$.

Для вычисления количества благоприятных исходов для события С необходимо подсчитать число способов выбрать один туз в первую группу (всего тузов в колоде 4, поэтому вычисляем $C_4^1 = 4$) и число способов выбрать 25 карт, не являющихся тузами, в первую группу (всего карт, не являющихся тузами, в колоде 52-4=48, поэтому вычисляем C_{48}^{25}), после чего осталось только поменять местами 1-ю и 2-ю группы при каждом наборе. Получаем $p(A)=\frac{m}{n}=\frac{2\cdot C_4^1\cdot C_{48}^{25}}{C_{52}^{26}}$.

1.1.2.3. Геометрическое определение вероятностей

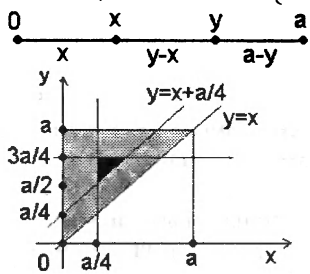
Пример 1

Стержень длины *a* наудачу разломан на 3 части. Найдите вероятность того, что длина каждой части окажется больше *a/4*.

Решение

Все случаи: $\begin{cases} 0 < x < a \\ 0 < y < a \\ x < y \end{cases}$.

Благоприятные случаи: $\begin{cases} x > a/4 \\ y - x > a/4 \\ a - y > a/4 \end{cases} \Leftrightarrow \begin{cases} x > a/4 \\ y < x + a/4 \\ y < 3a/4 \end{cases}$.



x и y – точки разломов

S(D)=a^2/2 (площадь серого большого треугольника)

S(d)=1/2*$(a/4)^2$=a^2/32

(площадь черного маленького треугольника)

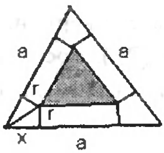
По геометрическому определению вероятность равна

$$\frac{S(d)}{S(D)} = \frac{a^2/32}{a^2/2} = \frac{1}{16}.$$

Пример 2

На паркет, составленный из правильных треугольников, со стороной *a* брошена монета радиуса *r (r<a)*. Найдите вероятность того, что монета не заденет границ ни одного из треугольников.

Решение



Число всех возможных исходов равно площади правильного треугольника со стороной a, т. е.

$$S(D)=\frac{\sqrt{3}\cdot a^2}{4}.$$

Число благоприятных исходов равно площади правильного треугольника со стороной $(a-2x)$.

$$x=\frac{r}{tg\,30^0}=\sqrt{3}\cdot r \quad\Rightarrow\quad S(d)=\frac{\sqrt{3}\cdot\left(a-2\sqrt{3}\cdot r\right)^2}{4}.$$

По геометрическому определению вероятность равна

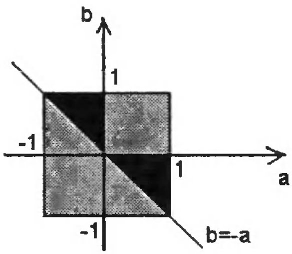
$$\frac{S(d)}{S(D)}=\frac{(a-2\sqrt{3}\cdot r)^2}{a^2}.$$

Пример 3

Найдите вероятность того, что сумма двух наудачу взятых действительных чисел из отрезка [-1,1] больше нуля, а их произведение отрицательно.

Решение

Все случаи: $\begin{cases}-1\le a\le 1\\ -1\le b\le 1\end{cases}.$



Благоприятные случаи: $\begin{cases}a+b>0\\ a\cdot b<0\end{cases}.$

Число всех возможных исходов равно площади квадрата, т. е. $S(D)=2^2=4$.

Число благоприятных исходов равно площади двух черных треугольников, т. е. $S(d)=1$.

По геометрическому определению вероятность равна $\dfrac{S(d)}{S(D)}=\dfrac{1}{4}$.

1.1.3. Задания и вопросы для самостоятельной работы

Задания

1. Образуют ли следующие события полную группу несовместных событий?

Опыт – два выстрела по мишени.

События: А – ни одного попадания;

В – одно попадание;

С – два попадания.

2. Являются ли равновозможными следующие события?

Опыт – вынимание одной карты из колоды

События: B1 – появление карты червонной масти;

B2 – появление карты бубновой масти;

B3 – появление карты трефовой масти.

3. Является ли следующая группа событий полной группой несовместных равновозможных событий?

Опыт – бросание двух монет.

События: C1 – появление двух гербов;

C2 – появление двух цифр;

C3 – появление одного герба и одной цифры.

4. Является ли следующая группа событий полной группой несовместных равновозможных событий?

Опыт – бросание двух монет.

События: C1 – появление двух гербов;

C2 – появление двух цифр.

C3 – появление хотя бы одного герба.

5. Является ли следующая группа событий полной группой несовместных равновозможных событий?

Опыт – бросание игральной кости.

События: C1 – появление менее двух очков;

C2 – появление трех или четырех очков;

C3 – появление более пяти очков.

6. Является ли следующая группа событий полной группой несовместных равновозможных событий?

Опыт – вынимание двух карт из колоды.

События: B1 – появление двух красных карт;

B2 – появление двух черных карт.

7. Является ли следующая группа событий полной группой равновозможных событий?

Опыт – два выстрела по мишени.

События: A – ни одного попадания;

B – одно попадание;

C – два попадания.

8. Являются ли несовместными и равновозможными следующие события?

Опыт – вынимание двух карт из колоды.

События: B1 – появление двух черных карт;

B2 – появление туза;

B3 – появление дамы.

9. Является ли следующая группа событий полной группой несовместных равновозможных событий?

Опыт – бросание игральной кости.

События: C1 – появление не более двух очков;

C2 – появление трех или четырех очков;

C3 – появление не менее пяти очков.

10. Являются ли несовместными и равновозможными следующие события?

Опыт – два выстрела по мишени.

События: A – хоты бы одно попадание;

B – хотя бы один промах.

11. Являются ли попарно несовместными и равновозможными следующие события?

Опыт – бросание двух монет.

События: C1 – появление двух гербов;

C2 – появление двух цифр;

C3 – появление одного герба и одной цифры.

12. Является ли следующая группа событий полной группой несовместных равновозможных событий?

Опыт – бросание монеты.

События: A1 – появление герба;

A2 – появление цифры.

13. Образуют ли полную группу совместных событий следующие события?

Опыт – два выстрела по мишени.

События: A – хотя бы одно попадание;

B – хотя бы один промах.

14. Являются ли равновозможными и несовместными следующие события?

Опыт – бросание игральной кости.

События: B1 – появление не менее трех очков;

B2 – появление не более четырех очков.

15. Является ли следующая группа событий полной группой несовместных равновозможных событий?

Опыт – два выстрела по мишени.

События: A1 – ни одного попадания;

A2 – одно попадание;

A3 – два попадания.

16. Мишень состоит из десяти кругов, ограниченных концентрическими окружностями с радиусами R_k ($k=1,2,\ldots,10$), причем $R_1<R_2<\ldots<R_{10}$. Событие A_k – попадание в круг радиуса R_k. Что означают события

$$B=\bigcup_{k=1}^{6}A_k, \qquad C=\bigcap_{k=1}^{10}A_k, \qquad D=A_5\cdot A_6, \qquad E=\overline{A}_1\cdot A_2?$$

17–21. Опыт состоит в бросании точки в прямоугольник с изображенными областями V_1, V_2, V_3. События A, B и C означают соответственно попадание точки в области V_1, V_2, V_3. Изобразите возможную область попадания точки при следующих событиях:

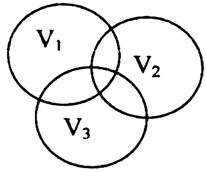
17. $\overline{A}+\overline{B}+\overline{C}$.

18. $A+B+\overline{C}$.

19. $AB+\overline{C}$.

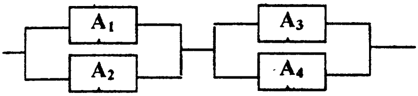
20. $AB\overline{C}$.

21. $A\overline{B}+C$.

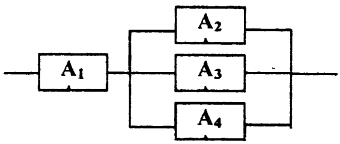


22–23. Электрическая цепь составлена по следующей схеме. Выход из строя элементов A_k (k = 1,2,3,4) – события A_k. Запишите событие D – разрыв цепи и противоположное ему событие.

22.



23.



24–26. Рабочий изготовил 4 детали. Пусть событие Ai (i=1,2,3,4) заключается в том, что i-я изготовленная им деталь имеет дефект. Записать события, заключающиеся в том, что:

24

a) Ни одна из деталей не имеет дефектов.

b) По крайней мере два изделия не имеют дефектов.

25

a) Хотя бы одна деталь имеет дефект.

b) Точно два изделия дефектны.

26

a) Только одна деталь имеет дефект.

b) Не более двух деталей имеют дефект.

27–30. Используя свойства операций, доказать равенства:

27. $A\cdot\overline{AB}+B=A+B$.

28. $\overline{\overline{A}+\overline{B}}=A\cdot B$.

29. $A\cdot\overline{B}+\overline{A}\cdot B=(A+B)\cdot\overline{A\cdot B}$.

30. $A\cdot\overline{B}+B\cdot\overline{C}+C\cdot\overline{A}+A\cdot B\cdot C=A+B+C$.

31. Ребенок играет с 10 буквами разрезной азбуки **А, А, А, Е, И, К, М, М, Т, Т**. Какова вероятность, что при случайном расположении букв в ряд он получит слово «**МАТЕМАТИКА**»?

32. Из колоды, насчитывающей 36 карт, наугад извлекаются 6 карт. Какова вероятность, что среди них окажется туз пик?

33. N книг произвольным образом расставляются на книжной полке. Какова вероятность, что две фиксированные книги окажутся стоящими рядом?

34. Из колоды, насчитывающей 36 карт, наугад извлекаются 6 карт. Какова вероятность, что среди них окажется ровно один туз?

35. У человека в кармане N ключей, из которых только один подходит к его двери. Ключи последовательно извлекаются (без возвращения) до тех

пор, пока не появится нужный ключ. Найти вероятность того, что нужный ключ появится при k-м извлечении.

36. Из колоды, насчитывающей 36 карт, наугад извлекаются 6 карт. Какова вероятность, что среди них окажутся ровно две бубновые карты?

37. В зрительном зале кинотеатра 500 мест. Какова вероятность, что при произвольном размещении в зале 490 зрителей пустыми останутся 10 первых мест второго ряда?

38. Группа, состоящая из 2n девушек и 2n юношей, делится произвольным образом на две равные по количеству подгруппы. Найти вероятность того, что в каждой подгруппе окажется поровну юношей и девушек.

39. В урне a белых и $в$ черных шаров. Из урны наудачу извлекают два шара. Какова вероятность того, что эти шары одного цвета?

40. Цифры 1, 2, 3, 4, 5 написаны на карточках. Карточки тщательно перемешиваются, а затем раскладываются в ряд. Какова вероятность того, что образовавшееся 5-значное число является четным?

41. На пяти карточках написаны цифры: 1, 2, 3, 4, 5. Две из них, одна за другой, вынимаются. Найти вероятность того, что число на второй карточке будет больше, чем на первой.

42. Имеется две урны: в первой a белых и b черных шаров, во второй – c белых и d черных. Из каждой урны вынимается по шару. Найдите вероятность того, что оба шара будут белыми.

43. Имеется две урны: в первой a белых и $в$ черных шаров, во второй – c белых и d черных. Из каждой урны вынимается по шару. Найдите вероятность того, что шары будут разного цвета.

44. В лифт семиэтажного дома на первом этаже вошли три человека. Каждый из них с одинаковой вероятностью выходит на любом из этажей, начиная со второго. Найти вероятность того, что все пассажиры выйдут одновременно (на одном и том же этаже).

45. В лифт семиэтажного дома на первом этаже вошли три человека. Каждый из них с одинаковой вероятностью выходит на любом из этажей, начиная со второго. Найти вероятность того, что все пассажиры выйдут на разных этажах.

46. В квадрат с вершинами в точках (0,0), (0,1), (1,1), (1,0) наудачу брошена точка (x,y). Найдите вероятность того, что координаты этой точки удовлетворяют неравенству $y<2x$.

47. Плоскость разграфлена параллельными прямыми, находящимися друг от друга на расстоянии 2a. На плоскость наудачу брошена монета радиуса r ($r<a$). Найдите вероятность того, что монета не пересечет ни одной из прямых.

48. На отрезке AB длиной 12 см наугад ставят точку M. Найдите вероятность того, что площадь квадрата, построенного на отрезке AM, будет заключена между 36 см^2 и 81 см^2.

49. Заданы две концентрические окружности с радиусами r и R ($r<R$). В области между окружностями взяли наугад точку и провели через нее ка-

сательные к меньшей окружности. Найдите вероятность того, что угол между касательными окажется меньше α.

50. На отрезке OA длины k наудачу поставлены две точки B и C (OB<OC). Найдите вероятность того, что длина отрезка BC меньше длины отрезка OB.

51. На отрезке длины k наудачу поставлены две точки. Найдите вероятность того, что из трех получившихся отрезков можно построить треугольник.

52. Из отрезка [–8; 2] выбраны случайным образом два числа. Найти вероятность того, что произведение этих чисел меньше 0, а сумма больше –2.

53. В сигнализатор поступают сигналы от двух устройств, причем поступление каждого из сигналов равновозможно в любой момент промежутка времени длительностью T. Моменты поступления сигналов независимы друг от друга. Сигнализатор срабатывает, если разность между моментами поступления сигналов меньше t (t<T). Найти вероятность того, что сигнализатор срабатывает за время T, если каждое из устройств пошлет по одному сигналу.

54. В одной из популярных в Америке игр игрок бросает монету с достаточно большого расстояния на поверхность стола, разграфленную на однодюймовые квадраты. Если монета (3/4 дюйма в диаметре) попадает полностью внутрь квадрата, то игрок получает награду, в противном случае он теряет свою монету. Каковы шансы выиграть при условии, что монета упала на стол.

55. На отрезок длины 20 см помещен меньший отрезок длины 10 см. Найти вероятность того, что отрезок, образовавшийся справа при отсечении меньшим отрезком от большего, длиннее, чем отрезок слева.

56. На отрезок [0,10] наудачу брошено 5 точек. Найти вероятность того, что две точки попадут в [0,2], одна – в [2,3] и две – в [3,10].

57. Из отрезка [0,1] наугад выбирается число. Какова вероятность, что в десятичной записи этого числа первая цифра после запятой будет двойкой?

58. В круг вписан квадрат. Какова вероятность того, что из 10 точек, брошенных наудачу в круг, четыре попадут в квадрат, три – в нижний сегмент и по одной – в оставшиеся три сегмента?

59. На бесконечную доску со стороной квадрата a бросается наудачу монета радиуса r ($2r<a$). Найти вероятность того, что монета будет иметь общие точки с 4 квадратами.

60. Дуэли в городе Осторожности редко кончаются печальным исходом. Дело в том, что каждый дуэлянт прибывает на место встречи в случайный момент времени между 5 и 6 часами утра и, прождав соперника 5 минут, удаляется. В случае же прибытия последнего в эти 5 минут дуэль состоится. Какая часть дуэлей действительно закончится поединком?

Вопросы

1. Случайные события, отношения между ними.
2. Случайные события, операции над ними.
3. Определения вероятности события. Классическое определение вероятности.

4. Определения вероятности события. Статистическое и геометрическое определения вероятности.

5. Геометрическое определение вероятности. Задача о встрече.

6. Геометрическое определение вероятности. Задача Бюффона.

§ 1.2. Вероятность суммы и произведения событий

1.2.1. Основные понятия

Независимость событий. Условная вероятность, свойства условной вероятности. Умножение и сложение вероятностей. Формула полной вероятности и формула Байеса.

Теоремы сложения и умножения

Вероятность суммы несовместных событий равна сумме их вероятностей:

$$P(A+B)=P(A)+P(B)$$

Вероятность суммы совместных событий равна сумме их вероятностей без вероятности их произведения:

$$P(A+B)=P(A)+P(B)-P(A\cdot B)$$

Вероятность произведения независимых событий равна произведению их вероятностей:

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

Вероятность произведения зависимых событий равна произведению вероятности одного из них на условную вероятность другого:

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

Формула полной вероятности:

$P(A)=P(H_1)\cdot P(A/H_1)+\ldots+P(H_n)\cdot P(A/H_n)$

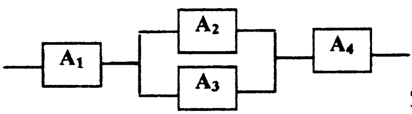
$H_1, \ldots, H_n$ – гипотезы, образующие полную группу.

Формула Байеса:

$$P(H_i/A)=\frac{P(H_i)\cdot P(A/H_i)}{P(A)}$$

1.2.2. Примеры решения задач

Пример 1



Электрическая цепь составлена из элементов Ai, i=1,2,3,4. Вероятность выхода из строя за данный период времени элемента A1 равна 0,3; A2 – 0,4; A3 – 0,1; A4 – 0,2. При выходе из строя любого элемента цепь в месте его включения разрывается. Предполагается, что элементы выходят или не выходят из строя независимо друг от друга. Найти вероятность того, что за рассматриваемый период по цепи будет проходить ток.

Решение

Введем события Ai – выход из строя i-го элемента, А – по цепи будет проходить ток. В условии задачи даны вероятности p(A1) = 0,3; p(A2) = 0,4; p(A3) = 0,1; p(A4) = 0,2. Отсюда вероятности противоположных событий (элемент цепи работает) равны: p($\bar{A}$1) = 0,7; p($\bar{A}$2) = 0,6; p($\bar{A}$1) = 0,9; p($\bar{A}$1) = 0,8. Найдем p(A).

$A = \overline{A1} \cdot (\overline{A2} + \overline{A3}) \cdot \overline{A4}$. При этом $\overline{A2}$ *и* $\overline{A3}$ – совместные события (поэтому вероятность их суммы равна сумме их вероятностей без вероятности произведения, $\overline{A2}$ *и* $\overline{A3}$ – независимые события, и вероятность их произведения равна произведению их вероятностей), $\overline{A1}$, $(\overline{A2} + \overline{A3})$, $\overline{A4}$ – независимые события (поэтому вероятность их произведения равна произведению их вероятностей).

$$p(A) = p(\overline{A1} \cdot (\overline{A2} + \overline{A3}) \cdot \overline{A4}) = p(\overline{A1}) \cdot p(\overline{A2} + \overline{A3}) \cdot p(\overline{A4}) =$$

$$= p(\overline{A1}) \cdot (p(\overline{A2}) + p(\overline{A3}) - p(\overline{A2} \cdot \overline{A3})) \cdot p(\overline{A4}) =$$

$$= p(\overline{A1}) \cdot (p(\overline{A2}) + p(\overline{A3}) - p(\overline{A2}) \cdot p(\overline{A3})) \cdot p(\overline{A4}) =$$

$$= 0{,}7 \cdot (0{,}6 + 0{,}9 - 0{,}6 \cdot 0{,}9) \cdot 0{,}8 = 0{,}5376$$

Пример 2

Из чисел 1, 2, ..., 49 наугад выбираются и фиксируются 6 чисел, считающиеся выигрышными. Некто, желающий выиграть, наугад называет свои 6 чисел из 49. Какова вероятность, что среди названных им чисел окажется не менее трех выигрышных?

Решение

Не менее трех выигрышных чисел означает, что их может быть либо 3, либо 4, либо 5, либо 6. Для нахождения вероятности заданного события необходимо найти вероятность суммы этих четырех событий. В данной задаче легче вычислить вероятность события, противоположного заданному в задаче событию (среди названных чисел окажется менее трех выигрышных, т. е. 0, 1 или 2 выигрышных числа), и вычислить его вероятность как вероятность суммы трех событий.

Введем события A0, A1, A2 – среди названных чисел окажется 0, 1 или 2 выигрышных числа, А – среди названных чисел окажется не менее трех выигрышных. Тогда вероятность заданного в задаче события найдем как p(A)=1-p($\bar{A}$).

$\bar{A} = A0 + A1 + A2$ и события A0, A1 и A2 – несовместны.

$$p(A0)=\frac{C_{43}^{6}}{C_{49}^{6}},\ p(A1)=\frac{C_{6}^{1}\cdot C_{43}^{5}}{C_{49}^{6}},\ p(A2)=\frac{C_{6}^{2}\cdot C_{43}^{4}}{C_{49}^{6}}.$$

$$p(A)=1-\left(\frac{C_{43}^{6}}{C_{49}^{6}}+\frac{C_{6}^{1}\cdot C_{43}^{5}}{C_{49}^{6}}+\frac{C_{6}^{2}\cdot C_{43}^{4}}{C_{49}^{6}}\right).$$

Пример 3

Стрелок А поражает мишень с вероятностью 0,6, стрелок В – с вероятностью 0,5, стрелок С – с вероятностью 0,4. Стрелки дали залп по мишени. Какова вероятность, что ровно две пули попали в цель?

Решение

Введем события А – стрелок А попал, В – стрелок В попал, С – стрелок С попал, Х – ровно две пули попали в цель.

По условию задачи p(A)=0,6; p(B)=0,5; p(C)=0,4.

Две пули попали в цель – значит, либо попали А и В, и не попал С, либо попали А и С, и не попал В, либо попали В и С, и не попал А.

Получаем $X=A\cdot B\cdot\overline{C}+A\cdot\overline{B}\cdot C+\overline{A}\cdot B\cdot C$ По условию задачи все события А, В, С и им противоположные независимые, а события $A\cdot B\cdot\overline{C}, A\cdot\overline{B}\cdot C, \overline{A}\cdot B\cdot C$ – несовместные.

$$p(X)=p(A\cdot B\cdot\overline{C}+A\cdot\overline{B}\cdot C+\overline{A}\cdot B\cdot C)=p(A\cdot B\cdot\overline{C})+p(A\cdot\overline{B}\cdot C)+p(\overline{A}\cdot B\cdot C)=$$
$$=p(A)\cdot p(B)\cdot p(\overline{C})+p(A)\cdot p(\overline{B})\cdot p(C)+p(\overline{A})\cdot p(B)\cdot p(C)=$$
$$=0{,}6\cdot 0{,}5\cdot 0{,}6+0{,}6\cdot 0{,}5\cdot 0{,}4+0{,}4\cdot 0{,}5\cdot 0{,}4=0{,}38$$

Пример 4

В первой урне 2 белых и 4 черных шара, а во второй – 3 белых и один черный шар. Из первой урны переложили во вторую два шара. Найти вероятность того, что шар, вынутый из второй урны после перекладывания, окажется белым.

Решение

Вероятность того, какой шар будет вынут из второй урны после перекладывания, зависит от того, какие два шара были переложены (это могут быть два белых шара – событие А, или два черных шара – событие В, или один белый и один черный шары – событие С). Поэтому для подсчета вероятности будем использовать формулу полной вероятности.

Пусть событие Х – шар, вынутый из второй урны после перекладывания, окажется белым. Тогда p(X) = p(A) · p(X/A) + p(B) · p(X/B) + p(C) · p(X/C).

$$p(A)=\frac{1}{C_{6}^{2}},\ p(B)=\frac{C_{4}^{2}}{C_{6}^{2}},\ p(C)=\frac{C_{4}^{1}\cdot C_{2}^{1}}{C_{6}^{2}},$$

$$p(X/A)=\frac{5}{6},\ p(X/B)=\frac{3}{6},\ p(X/C)=\frac{4}{6}.$$

$$p(X)=\frac{5}{6\cdot C_6^2}+\frac{3\cdot C_4^2}{6\cdot C_6^2}+\frac{4\cdot 4\cdot 2}{6\cdot C_6^2}=\frac{11}{18}.$$

Пример 5

Известно, что 5% всех мужчин и 0,25% всех женщин дальтоники. Наугад выбранное лицо страдает дальтонизмом. Какова вероятность, что это мужчина? (Считать, что мужчин и женщин одинаковое число.)

Решение

Введем событие А – наугад выбранное лицо страдает дальтонизмом – и две гипотезы до опыта: Н1 – выбрали мужчину, Н2 – выбрали женщину.

Тогда по формуле Байеса $p(H1/A)=\dfrac{p(H1)\cdot p(A/H1)}{p(A)}$.

По условию задачи p(H1)=p(H2)=0,5 (так как мужчин и женщин одинаковое число),

p(A/H1) = 0,05; p(A/H2) = 0,0025.

По формуле полной вероятности

p(A) = p(H1)p(A/H1) + p(H2)p(A/H2) = 0,5 · 0,05 + 0,5 · 0,0025 = 0,5 · 0,0525.

$$p(H1/A)=\frac{p(H1)\cdot p(A/H1)}{p(A)}=\frac{0{,}5\cdot 0{,}05}{0{,}5\cdot 0{,}0525}=\frac{20}{21}.$$

Пример 6

Группа студентов состоит из *a* отличников, *b* хорошо успевающих и *c* занимающихся слабо. Отличники на предстоящем экзамене могут получить только отличные оценки. Хорошо успевающие студенты могут получить с равной вероятностью хорошие и отличные оценки. Слабо занимающиеся могут получить с равной вероятностью хорошие, удовлетворительные и неудовлетворительные оценки. Для сдачи экзамена наугад вызывается один студент. Найти вероятность того, что он получит хорошую или отличную оценку.

Решение

Введем событие А – выбранный наугад студент получит хорошую или отличную оценку – и три гипотезы до опыта: Н1 – вызван студент отличник, Н2 – вызван хорошо успевающий студент, Н3 – вызван слабо занимающийся студент.

По формуле полной вероятности

p(A) = p(H1)p(A/H1) + p(H2)p(A/H2) + p(H3)p(A/H3).

$$p(H1)=\frac{a}{a+b+c},\ p(H2)=\frac{b}{a+b+c},\ p(H3)=\frac{c}{a+b+c}.$$

$$p(A/H1) = 1,\ p(A/H2) = 1,\ p(A/H3) = 1/3,\quad p(A) = \frac{3\cdot a + 3\cdot b + c}{3\cdot(a+b+c)}.$$

1.2.3. Задания и вопросы для самостоятельной работы

Задания

61. Из 20 сбербанков 10 расположены за чертой города. Для аудиторской проверки случайно выбраны 5 сбербанков. Какова вероятность того, что хотя бы два из них окажутся в черте города?

62. Найти вероятность того, что выбранное наудачу изделие является первосортным, если известно, что 4% всей продукции является браком, а 75% небракованных изделий удовлетворяют требованиям первого сорта.

63. Из группы туристов, отправляющихся за границу, 60% владеют английским языком, 40% – французским и 10% – обоими языками. Найти вероятность того, что наугад взятый турист будет нуждаться в переводчике (т. е. не знает ни английского, ни французского).

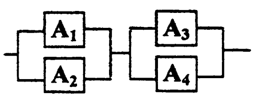
64. На полке стоят 10 книг, из них 3 – по теории вероятностей. Наугад выбираются 3 книги. Какова вероятность того, что среди них окажется хотя бы одна по теории вероятностей?

65. Среди поступающих в ремонт часов 40% нуждаются в общей чистке механизма. Какова вероятность того, что из пяти взятых наугад часов все нуждаются в чистке механизма?

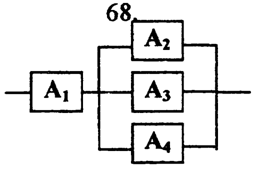
66. Какова вероятность того, что наугад выбранное трехзначное число окажется кратным хотя бы одному из чисел 4 или 6?

67–69. Электрическая цепь составлена из элементов A_i, $i=1,2,3,4$. Вероятность выхода из строя за данный период времени элемента A_1 равна 0,3; A_2 – 0,4; A_3 – 0,1; A_4 – 0,2. При выходе из строя любого элемента цепь в месте его включения разрывается. Предполагается, что элементы выходят или не выходят из строя независимо друг от друга. Найти вероятность того, что за рассматриваемый период по цепи будет проходить ток.

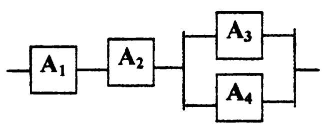
67.



68.



69.



70. На стеллаже библиотеки в случайном порядке расставлено 15 учебников, причем пять из них в переплете. Библиотекарь берет наудачу три учебника. Найти вероятность того, что хотя бы один из взятых учебников окажется в переплете (событие А).

71. Устройство состоит из трех элементов, работающих независимо. Вероятности безотказной работы (за время Т) первого, второго и третьего элементов соответственно равны 0,6; 0,7; 0,8. Найти вероятности того, что за

время Т безотказно будут работать: а) только один элемент; в) только два элемента; с) все три элемента.

72. Вероятности того, что нужная сборщику деталь находится в первом, втором, третьем, четвертом ящике, соответственно равны 0,6; 0,7; 0,8; 0,9. Найти вероятность того, что деталь содержится: а) не более чем в трех ящиках; в) не менее чем в двух ящиках.

73. Экзаменационный билет содержит 3 вопроса. Вероятность того, что студент ответит на первый вопрос билета, равна 0,9, на второй – 0,8, на третий – 0,5. Найти вероятность того, что студент сдаст экзамен, если для этого необходимо ответить: а) на все вопросы, в) хотя бы на два вопроса.

74. В двух урнах находятся шары, отличающиеся только цветом, причем в первой урне 5 белых, 11 черных и 8 красных шаров, а во второй – 10 белых, 8 черных и 6 красных. Из обеих урн наудачу извлекается по одному шару. Какова вероятность того, что оба шара одного цвета?

75. Одновременно три стрелка делают выстрел по одной и той же мишени. Какова вероятность того, что мишень будет поражена, если вероятность попадания первого стрелка равна 0,8; второго – 0,6; третьего – 0,5.

76. Читатель в поисках нужной книги обходит три библиотеки. Вероятность того, что книга имеется в очередной библиотеке, равна 0,3. Что вероятнее – найдет читатель книгу или нет?

77. Число грузовых автомашин, проезжающих по шоссе, на котором стоит бензоколонка, относится к числу легковых машин, проезжающих по тому же шоссе, как 3:2. Вероятность того, что будет заправляться грузовая машина, равна 0,1; для легковой машины эта вероятность равна 0,2. К бензоколонке подъехала для заправки машина. Найти вероятность того, что это грузовая машина.

78. Вероятность хотя бы одного попадания в мишень стрелком при трех выстрелах равна 0,784. Найти вероятность одного промаха при трех выстрелах.

79. По теории вероятностей и математической статистике имеется 30 экзаменационных билетов. Студент Павлов выучил только 20. Каким выгоднее ему зайти на экзамен, первым или вторым?

80. Группа студентов состоит из a отличников, b хорошо успевающих и c занимающихся слабо. Отличники на предстоящем экзамене могут получить только отличные оценки. Хорошо успевающие студенты могут получить с равной вероятностью хорошие и отличные оценки. Слабо занимающиеся могут получить с равной вероятностью хорошие, удовлетворительные и неудовлетворительные оценки. Для сдачи экзамена наугад вызываются три студента. Найти вероятность того, что они получат отметки: отлично, хорошо и удовлетворительно (в любом порядке).

81. Имеются три партии деталей по 20 деталей в каждой. Число стандартных деталей в 1, 2 и 3-й партиях соответственно равно 20, 15, 10. С какой вероятностью наудачу извлеченная деталь окажется стандартной?

82. Экономист считает, что вероятность роста стоимости акции компании в следующем году составит 0,75, если экономика страны будет на подъ-

еме, и 0,30, если экономика не будет успешно развиваться. По мнению экспертов, вероятность экономического подъема равна 0,6. Оценить вероятность того, что акции компании поднимутся в следующем году.

83. Инвестор вложил капитал в ценные бумаги двух финансовых фирм. При этом он надеется получить доход в течение обусловленного времени от первой фирмы с вероятностью 0,9; от второй – с вероятностью 1. Однако есть возможность банкротства фирм независимо друг от друга, которая оценивается для первой фирмы вероятностью 0,1; для второй – 0,02. В случае банкротства фирмы инвестор получает только вложенный капитал. Какова вероятность того, что инвестор получит прибыль?

84. База получает некоторую продукцию с трех заводов. В объемах 25, 35 и 40% соответственно. В продукции 1-го завода брак составляет 3%, 2-го завода – 1%, 3-го завода – 2%. Найти вероятность того, что потребитель получит с базы стандартное изделие.

85. Литье в болванках поступает из 2 заготовительных цехов: 70% – из первого, 30% – из второго. При этом материал 1-го цеха имеет 10% брака, а второго – 20%. Найти вероятность того, что одна взятая наудачу болванка не имеет дефектов.

86. В спартакиаде участвуют: из 1-й группы – 4 студента, из 2-й – 6, из 3-й – 5. Студент 1-й группы попадает в сборную института с вероятностью 0,9, для студента 2-й группы эта вероятность равна 0,7, для студента 3-й группы – 0,8. Найти вероятность того, что наудачу выбранный студент попал в сборную института.

87. В спартакиаде участвуют: из 1-й группы – 4 студента, из 2-й – 6, из 3-й – 5. Студент 1-й группы попадает в сборную института с вероятностью 0,9, для студента 2-й группы эта вероятность равна 0,7, для студента 3-й группы – 0,8. Найти вероятность того, что попавший в сборную института студент из 3-й группы.

88. В больницу поступают в среднем 50% больных с заболеванием K, 30% – с заболеванием L, 20% – с заболеванием M. Вероятность полного излечения болезни K равна 0,7; для болезней L и M эти вероятности равны соответственно 0,8 и 0,9. Больной, поступивший в больницу, был выписан здоровым. Найти вероятность того, что он страдал заболеванием K.

89. В вычислительной лаборатории имеются 6 клавишных автоматов и 4 полуавтомата. Вероятность того, что за время выполнения некоторого расчета автомат не выйдет из строя, равна 0,95; для полуавтомата эта вероятность равна 0,8. Студент производит расчет на наудачу выбранной машине. Найти вероятность того, что до окончания расчета машина не выйдет из строя.

90. Студент из 40 экзаменационных вопросов выучил только 30. Каким выгоднее ему зайти на экзамен, первым или вторым?

Вопросы

1. Вероятность суммы несовместных событий.
2. Зависимые и независимые события. Условные вероятности.
3. Вероятность произведения зависимых событий.

4. Вероятность произведения независимых событий.
5. Вероятность суммы совместных событий.
6. Формула полной вероятности.
7. Формула Байеса.

§ 1.3. Повторение n независимых опытов

1.3.1. Основные понятия

Распределение вероятностей Бернулли. Предельные теоремы Муавра – Лапласа, следствия из теорем.

Формула Бернулли

Вероятность того, что событие произойдет ровно m раз при n независимых испытаниях, если при одном испытании событие происходит с вероятностью p и не происходит с вероятностью q, равна $P_n(m) = C_n^m \cdot p^m \cdot q^{n-m}$.

Из **локальной предельной теоремы Муавра – Лапласа** следует, что при больших n вероятность $P_n(m) \approx \frac{1}{\sqrt{npq}} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(m-np)^2}{2npq}}$, где $\varphi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}$ для $x = \frac{m-np}{\sqrt{npq}}$ – функция, значения которой имеются в таблицах.

Из **интегральной предельной теоремы Муавра – Лапласа** следует, что при больших n вероятность $P_n(m_1 < m < m_2) \approx \Phi(x_2) - \Phi(x_1)$,где $\Phi(x) = \frac{1}{\sqrt{2\pi}} \cdot \int_0^x e^{-\frac{t^2}{2}} dt$ для $x_i = \frac{m_i - np}{\sqrt{npq}}$ $(i = 1{,}2)$ – функция Лапласа, значения которой имеются в таблицах.

Следствия из интегральной предельной теоремы Муавра – Лапласа:

$$P_n\left(\left|\frac{m}{n} - p\right| \le \varepsilon\right) = 2 \cdot \Phi\left(\varepsilon \cdot \sqrt{\frac{n}{p \cdot q}}\right).$$

$$P_n\left(|m - n \cdot p| \le \varepsilon\right) \approx 2 \cdot \Phi\left(\frac{\varepsilon}{\sqrt{n \cdot p \cdot q}}\right).$$

1.3.2. Примеры решения задач

1.3.2.1. Формула Бернулли

Пример

Что вероятнее, выиграть у равносильного противника три партии из четырех или пять из восьми?

Решение

Так как противники равносильные, то их вероятности выиграть одну партию одинаковы и равны 0,5. Воспользуемся формулой Бернулли, которая вычисляет вероятность наступления события ровно *m* раз при *n* независимых испытаниях: $P_n(m) = C_n^m \cdot p^m \cdot q^{n-m}$, где p – вероятность наступления события в одном испытании, q – вероятность, с которой событие в одном испытании не наступит. В данной задаче событие – это наступление выигрыша, испытания – это сыгранные партии, поэтому p=q=0,5.

Найдем вероятность выиграть три партии из четырех.

$$P_4(3) = C_4^3 \cdot 0{,}5^3 \cdot 0{,}5^1 = \frac{1}{4}.$$

Найдем вероятность выиграть пять партии из восьми.

$$P_8(5) = C_8^5 \cdot 0{,}5^5 \cdot 0{,}5^3 = \frac{7}{32}.$$

Получаем, что три партии из четырех выиграть более вероятно.

1.3.2.2. Локальная и интегральная теоремы Лапласа

Пример 1

Вероятность рождения мальчика равна 0,51. Найти вероятность того, что среди 100 новорожденных окажется 50 мальчиков.

Решение

Воспользуемся локальной теоремой Лапласа:

$$P_n(m) \approx \frac{1}{\sqrt{n \cdot p \cdot q}} \cdot \varphi(x), \quad x = \frac{m - np}{\sqrt{n \cdot p \cdot q}},$$

которая позволяет вычислить вероятность наступления события ровно *m* раз при *n* независимых испытаниях, где *p* – вероятность наступления события в одном испытании, *q* – вероятность, с которой событие в одном испытании не наступит.

В данной задаче событие – это рождение мальчика, испытания – это рождение ребенка. По условию задачи p=0,51; q=0,49; n=100; m=50.

$$x = \frac{m - n \cdot p}{\sqrt{n \cdot p \cdot q}} = \frac{50 - 100 \cdot 0{,}51}{\sqrt{100 \cdot 0{,}51 \cdot 0{,}49}} \approx -0{,}20004$$

Найдем значение функции по таблице, воспользовавшись ее свойством четности:

$$\varphi(x)=\varphi(-0{,}2)=\varphi(0{,}2)=0{,}3910.$$

$$P_{100}(50)\approx\frac{1}{\sqrt{100\cdot 0{,}51\cdot 0{,}49}}\cdot 0{,}3910\approx 0{,}0782.$$

Пример 2

Вероятность появления события в каждом из 100 независимых испытаний постоянна и равна p=0,8. Найти вероятность того, что событие появится не менее 75 раз и не более 90 раз.

Решение

Воспользуемся интегральной теоремой Лапласа:

$$P_n(m_1\le m\le m_2)\approx\Phi(x_2)-\Phi(x_1),\quad x_i=\frac{m_i-n\cdot p}{\sqrt{n\cdot p\cdot q}},$$

которая позволяет вычислить вероятность наступления события число раз m из отрезка от m_1 до m_2 при n независимых испытаниях, где p – вероятность наступления события в одном испытании, q – вероятность, с которой событие в одном испытании не наступит.

По условию задачи p=0,8; q=0,2; n=100; m_1=75; m_2=90.

$$x_1=\frac{m_1-n\cdot p}{\sqrt{n\cdot p\cdot q}}=\frac{75-100\cdot 0{,}8}{\sqrt{100\cdot 0{,}8\cdot 0{,}2}}\approx -1{,}25.$$

$$x_2=\frac{m_2-n\cdot p}{\sqrt{n\cdot p\cdot q}}=\frac{90-100\cdot 0{,}8}{\sqrt{100\cdot 0{,}8\cdot 0{,}2}}\approx 2{,}5.$$

Найдем значение функции Лапласа по таблице, воспользовавшись ее свойством нечетности.

$$\Phi(x_1)=\Phi(-1{,}25)=-\Phi(1{,}25)=-0{,}3944.$$

$$\Phi(x_2)=\Phi(2{,}5)=0{,}4938.$$

$$P_{100}(75\le m\le 90)\approx\Phi(2{,}5)-\Phi(-1{,}25)=0{,}4938+0{,}3944=0{,}8882.$$

Пример 3

Вероятность появления события в каждом из независимых испытаний равна 0,5. Найти число испытаний n, при котором с вероятностью 0,7698 можно ожидать, что относительная частота появления события отклонится от его вероятности по абсолютной величине не более чем на 0,02.

Решение

Воспользуемся следствием из интегральной теоремы Лапласа:

$$P_n\left(\left|\frac{m}{n}-p\right|\le\varepsilon\right)=2\cdot\Phi\left(\varepsilon\cdot\sqrt{\frac{n}{p\cdot q}}\right).$$

По условию задачи p=0,5; ε=0,02; $P_n\left(\left|\frac{m}{n}-p\right|\le 0{,}02\right)=0{,}7698$.

Найдем число испытаний n.

$$2\cdot\Phi\left(\varepsilon\cdot\sqrt{\frac{n}{p\cdot q}}\right)=0{,}7698.$$

$\Phi\left(0{,}02\cdot\frac{\sqrt{n}}{0{,}5}\right)=0{,}3849$. По таблице находим $0{,}02\cdot\frac{\sqrt{n}}{0{,}5}=1{,}2$. Отсюда $\sqrt{n}=30$, n=900.

Пример 4

Кандидата в высший орган власти поддерживают 80% населения. В каких пределах с вероятностью 0,95 находится число проголосовавших «за» на выборах, если число избирателей равно1200000?

Решение

Воспользуемся следствием из интегральной теоремы Лапласа:

$$P_n(|m-n\cdot p|\le\varepsilon)\approx 2\cdot\Phi\left(\frac{\varepsilon}{\sqrt{n\cdot p\cdot q}}\right).$$

В данной задаче опыт – это голосование одного избирателя, событие – голосование «за». Необходимо найти, в каких границах находится *m*.

$$|m-n\cdot p|\le\varepsilon \quad\Leftrightarrow\quad n\cdot p-\varepsilon\le m\le n\cdot p+\varepsilon.$$

По условию задачи n=1200000; p=0,8; q=0,2; $P_n(|m-n\cdot p|\le\varepsilon)=0{,}95$.

$$2\cdot\Phi\left(\frac{\varepsilon}{\sqrt{n\cdot p\cdot q}}\right)=0{,}95 \quad\Leftrightarrow\quad \Phi\left(\frac{\varepsilon}{\sqrt{n\cdot p\cdot q}}\right)=0{,}475.$$

По таблице находим $\frac{\varepsilon}{\sqrt{n\cdot p\cdot q}}=1{,}95$, получаем $\varepsilon=1{,}95\cdot\sqrt{1200000\cdot 0{,}8\cdot 0{,}2}\Leftrightarrow\varepsilon\approx 854$.

$$n\cdot p-\varepsilon\le m\le n\cdot p+\varepsilon.$$

$$1200000\cdot 0{,}8-854\le m\le 1200000\cdot 0{,}8+854.$$

$$959146\le m\le 960854.$$

1.3.3. Задания и вопросы для самостоятельной работы

Задания

91. Вероятность появления события в каждом из 100 независимых испытаний постоянна и равна p=0,8. Найти вероятность того, что событие появится не менее 75 раз.

92. В урне содержатся белые и черные шары в отношении 4:1. После извлечения шара регистрируется его цвет, и шар возвращается в урну. Чему равно наименьшее число извлечений n, при котором с вероятностью 0,95 можно ожидать, что абсолютная величина отклонения относительной частоты появления белого шара от его вероятности будет не более чем 0,01?

93. Вероятность появления события в каждом из независимых испытаний равна 0,8. Найти такое положительное число ε, чтобы с вероятностью 0,99 абсолютная величина отклонения относительной частоты появления события от его вероятности 0,8 не превысила ε.

94. Отдел технического контроля проверяет на стандартность 900 деталей. Вероятность того, что деталь стандартна, равна 0,9. Найти с вероятностью 0,95 границы, в которых будет заключено число m стандартных деталей среди проверенных.

95. Вероятность найти белый гриб среди прочих равна 0,25. Какова вероятность того, что среди 80 грибов белых будет 20?

96. В партии из 768 арбузов каждый арбуз оказывается неспелым с вероятностью 0,25. Найти вероятность того, что количество спелых арбузов будет находиться в пределах от 564 до 600.

97. Сколько нужно произвести бросаний монеты, чтобы с вероятностью 0,99 можно было утверждать, что частота выпадения герба отличается от 0,5 по модулю не более чем на 0,001?

98. Для поступления в университет необходимо успешно сдать вступительные экзамены. В среднем их выдерживают 25% абитуриентов. Предположим, что в приемную комиссию поступило 1800 заявлений. Чему равна вероятность того, что хотя бы 450 поступающих успешно сдадут экзамены?

99. Известно, что из 100 семей 80 имеют холодильник. Найти вероятность того, что из 400 семей 300 имеют холодильник.

100. В среднем 20% акций на аукционе продается по первоначально заявленной стоимости. Найти вероятность того, что из 10 пакетов акций в результате торгов будут проданы не менее двух пакетов акций.

101. Сборник содержит 350 задач с ответами. В каждом ответе может быть ошибка с вероятностью 0,01. Какова вероятность того, что ответы в 4 задачах даны с ошибками?

102. При эпидемии гриппа 40% населения заражены вирусом. В лаборатории числится 24 сотрудника. Какова вероятность того, что 10 из них будут носителями вируса?

103. На факультете 20% студентов выходцы из сельской местности. В каких пределах с вероятностью 0,9 находится число городских жителей среди студентов из группы в 28 человек?

104. Вероятность обнаружения упавшего на Землю метеорита, по подсчетам ученых, составляет 0,0001. Какова вероятность обнаружения двух и более метеоритов из 3000 упавших на Землю?

105. Сколько нужно произвести бросаний симметричной монеты, чтобы с вероятностью 0,9 отклонение частоты выпадения герба отличалось от 0,5 не более чем на 0,01?

Вопросы

1. Формула Бернулли.
2. Распределение вероятностей Бернулли.
3. Локальная теорема Муавра – Лапласа.

4. Интегральная теорема Муавра – Лапласа.
5. Применения интегральной теоремы Муавра – Лапласа.
6. Закон Пуассона.

§ 1.4. Варианты контрольной работы

Задания для самостоятельной работы, предлагаемые в главе 1, могут быть использованы для составления контрольной работы по данной теме. Все 105 задач можно разбить на 15 вариантов следующим образом.

Вариант	№ заданий
1	1, 16, 31, 46, 61, 76, 91
2	2, 17, 32, 47, 62, 77, 92
3	3, 18, 33, 48, 63, 78, 93
4	4, 19, 34, 49, 64, 79, 94
5	5, 20, 35, 50, 65, 80, 95
6	6, 21, 36, 51, 66, 81, 96
7	7, 22, 37, 52, 67, 82, 97
8	8, 23, 38, 53, 68, 83, 98
9	9, 24, 39, 54, 69, 84, 99
10	10, 25, 40, 55, 70, 85, 100
11	11, 26, 41, 56, 71, 86, 101
12	12, 27, 42, 57, 72, 87, 102
13	13, 28, 43, 58, 73, 88, 103
14	14, 29, 44, 59, 74, 89, 104
15	15, 30, 45, 60, 75, 90, 105

§ 1.5. Решение задач с применением программирования и EXCEL

Задание 1

Рассмотрим задачу: «С какой вероятностью при однократном подбрасывании игральной кости выпадет 5 очков?»

По классическому определению вероятность равна p=1/6=0,1(6).

Воспользуемся статистическим определением. Смоделируем подбрасывание игральной кости (т. е. сгенерируем случайное целое число от 1 до 6) n раз. Подсчитаем при этом m – количество выпадений числа 5. Найдем относительную частоту появления 5 очков как $w=m/n$.

Сравним абсолютные значения отклонений теоретической вероятности p от относительной частоты w при различных способах генерации случайных чисел.

1. Воспользуемся генератором случайных чисел в языке программирования Pascal.

```
. . .
Randomize; m:=0;
For i:=1 to n do
 Begin
    x:=1+Random(6);
    If x=5 then inc(m);
 End;
w:=m/n;
. . .
```

Например, для n=100000 возможное значение w=0,16704,
следовательно, |p-w|=0,00037.

2. Генератор случайных чисел как инструмент Анализа данных в электронных таблицах Excel.

В меню Сервис выберите команду Анализ данных. Если эта команда недоступна, загрузите пакет анализа. Для этого:

- В меню Сервис выберите команду Надстройки.
- В списке надстроек выберите Пакет анализа и нажмите кнопку ОК.

Выберите функцию Генерация случайных чисел в диалоговом окне Анализ данных и нажмите кнопку ОК.

Установите параметры анализа в соответствующем диалоговом окне. Число переменных = 1, число случайных чисел = 1000, распределение равномерное, параметры – между 1 и 6.

Для округления полученных случайных чисел используйте функцию ОКРУГЛ(<адрес ячейки>; 0), для вычисления числа *m* – функцию СЧЕТЕСЛИ(<диапазон>;5).

Например, для n=1000 возможное значение w=0,186,
следовательно, |p-w|=0,0193.

Задания для самостоятельной работы

1. Найти вероятность, с которой, дважды подбросив игральную кость, игрок получает сумму выпавших очков, равную 6.

2. Найти вероятность, с которой корни квадратного уравнения $ax^2+bx+c=0$ являются рациональными числами. Коэффициенты a, b, c равны количеству выпавших очков при троекратном подбрасывании игральной кости.

3. Три человека, сев в лифт на первом этаже семиэтажного дома, с равной вероятностью могут выйти на любом из этажей выше первого. Найти вероятность, с которой все пассажиры лифта выйдут на разных этажах.

Задание 2

«О победе футбольной команды». Группа из *n* фанатов выигрывающей футбольной команды на радостях бросает свои шляпы в воздух. Шляпы возвращаются в случайном порядке – по одной к каждому из болельщиков. С какой вероятностью *k* болельщиков получат назад свои собственные шляпы?

Формализовав задачу, мы приходим к следующему вопросу: «Сколько существует перестановок из пяти элементов, в которых k элементов стоят на своих местах?» Ответив на этот вопрос и воспользовавшись тем, что общее число перестановок из n элементов равно $n!$, решаем задачу по классическому определению вероятности.

Используем для решения алгоритм генерации всех перестановок из n элементов с подсчетом количества k неподвижных точек в каждой перестановке.

Предлагаем фрагмент программной реализации этого алгоритма на языке программирования Pascal.

Структуры данных:

- перестановка из элементов $1, 2, ..., n$ – одномерный массив a;
- количество перестановок с k неподвижными точками – целая переменная c;
- искомая вероятность – вещественная переменная p.

```
Procedure P (var a:mas);         {формирование следующей
                                  перестановки из n элементов}
var i,j:byte;
begin
   i:=n-1;
while (i>0) and (a[i]>a[i+1]) do dec(i);
   j:=n;
   while a[j]<a[i] do dec(j);
   < a[i]↔a[j] >;
For j:=1 to (n-i) div 2 do <a[i+j]↔a[n-j+1]> ;
end;

{основная программа}

Begin
   f:=<факториал числа n >;
   if n=k then p:=1/f
    else begin
      < генерация первой перестановки из n элементов,
      а именно, перестановки 1, 2, 3, …,n>;
      c:=0;
      For i:=2 to f do
       begin
            P(a);
            x:=< количество неподвижных точек в перестановке a >;
            if x=k then inc(c);
       end;
```

```
      p:=c/f;
      end;
   Writeln('Ответ: ', p:5:3);
   Readln;
End.
```

Задания для самостоятельной работы

1. В урне находятся 10 пронумерованных шаров. Шары перемешиваются и последовательно, один за другим, извлекаются из урны. С какой вероятностью хотя бы для двух шаров их номера совпадут с номерами извлечения?

2. На званый обед приглашены 10 гостей. Хозяин решил, как именно нужно рассадить гостей за стол, и приготовил карточки с именами. С какой вероятностью хотя бы четверо из приглашенных сядут на выбранные хозяином места, если он забыл расставить карточки.

3. Найти все решения в задаче «о победе футбольной команды» для значений n и k от 0 до 10. Результат работы программы – двумерная таблица размерности 11x11, элементами которой являются количества неподвижных точек.

Задание 3

Имеется 2n шаров, из них 2k белых. Шары раскладывают в две коробки поровну. С какой вероятностью в коробках окажется одинаковое число белых шаров?

Решение задачи с помощью классического определения: $p=\dfrac{C_{2k}^{k}\cdot C_{2n-2k}^{n-k}}{C_{2n}^{n}}$.

Для n=4, k=3 получаем $p=\dfrac{C_6^3\cdot C_2^1}{C_8^4}=\dfrac{4}{7}\approx 0{,}57142857$.

Алгоритмический подход к решению задачи

Заполним одномерный массив a размерности *2n* целыми числами (все имеющиеся шары), среди которых *2k* чисел отрицательные (белые шары), а остальные положительные. Для моделирования разбиения шаров на две равные группы сгенерируем все последовательности из нулей и единиц длиной *2n* (одномерный массив *b*). При этом элементы, помеченные 0, относятся к первой группе, а помеченные 1 – ко второй. Если количество единиц и нулей в массиве *b* совпадает и равно *n*, следовательно, произошло разбиение шаров на равные части (количество таких разбиений подсчитываем в переменной *v*). Если количество отрицательных чисел (белых шаров) в одной из двух равных частей совпадает с *k*, следовательно, в обеих коробках по *k* белых шаров (количество таких разбиений подсчитываем в переменной *m*). Искомую вероятность найдем по классическому определению *p=m/v*.

```
const n=4;k=3;
type mas=array[1..2*n] of integer; var a,b:mas; m,v:byte;p:real;
procedure Input (var a:mas);
var i:byte;
begin
for i:=1 to 2*k do
  a[i]:=(-1)*i;
for i:=2*k+1 to 2*n do
  a[i]:=i-2*k;
end;
procedure Rec(i,c,x:byte);
begin
   If i<=2*n then
    begin
b[i]:=0; Rec(i+1,c,x);
              b[i]:=1; If a[i]<0 then Rec(i+1,c+1,x+1) else Rec(i+1,c+1,x);
    end
   else
    if c=n then
       begin
              inc(v);
              if x=k then inc(m);
       end
end;
Begin
Input(a);
m:=0; v:=0;
Rec(1,0,0);
p:=m/v;
writeln('p=',p:5:12);
readln;
End.
```

Задания для самостоятельной работы

1. Группу, в которой $3n$ студентов и из них $3k$ девушек ($k\leq n$), произвольным образом разбивают на три равные подгруппы. С какой вероятностью в каждой из подгрупп окажется одинаковое число девушек? Проверьте результат, например, для $n=5$, $k=3$.

2. Колода из 52 карт делится на две равные части. С какой вероятностью в одной из частей будет 3 туза?

3. Было куплено 90 лотерейных билетов, среди которых оказалось 9 выигрышных (3 билета с выигрышем по 100 рублей, 3 билета – по 200 рублей и 3 билета – по 300 рублей). Три друга поровну поделили все билеты между собой. С какой вероятностью все получат равные по сумме выигрыши?

Глава 2. Случайные величины и их распределения

§ 2.1. *Числовые характеристики случайных величин*

2.1.1. Основные понятия и примеры решения задач

Случайные величины. Закон распределения дискретной и непрерывной случайных величин, числовые характеристики (математическое ожидание, дисперсия, среднее квадратичное отклонение) и их свойства. Функция распределения и плотность вероятностей непрерывной случайной величины, ее числовые характеристики.

Числовые характеристики распределения случайной величины

Случайная величина X задается:

- числовыми значениями $x_1,\ldots, x_n$;
- вероятностями $p_1,\ldots, p_n$, с которыми она принимает эти числовые значения;
- плотностью распределения вероятностей – дифференциальной функцией f(x).

Математическое ожидание

дискретной случайной величины $M(X)=\sum\limits_{i=1}^{n} p_i \cdot x_i$.

непрерывной случайной величины $M(X)=\int\limits_{-\infty}^{+\infty} x\cdot f(x)dx$.

Дисперсия

дискретной случайной величины $D(X)=M(X^2)-M^2(X)$.

непрерывной случайной величины $D(X)=\int\limits_{-\infty}^{+\infty} x^2\cdot f(x)dx-M^2(X)$.

Среднее квадратическое отклонение

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

Мода (Мо) – значение признака, наиболее часто встречающееся в исследуемой совокупности. Это одна из вариант признака, которая в ряду распределения имеет наибольшую частоту (или частость)

Модой Мо(X) случайной величины X называется ее наиболее вероятное значение, т. е. значение случайной величины, для которого вероятность (для дискретной величины) или плотность вероятности (для непрерывной величины) достигает максимума.

Если вероятность или плотность вероятности достигает максимума не в одной, а в нескольких точках, распределение называется полимодальным, если мода единственна, то распределение называется унимодальным.

➢ В **дискретном** ряду мода определяется визуально по максимальной частоте (или частости).

Пример 1: распределение женской обуви, проданной за месяц в магазине:

Размер женской обуви (x_i)	Число проданных пар, % к итогу (w_i)
33	4
34	12
35	18
36	26
37	20
38	13
39	6
40	1
Итого:	100

Наибольшая частость равна 26 и соответствует варианте 36. Мо=36 Наибольшим спросом у женщин пользовался 36-й размер.

➢ В **интервальном** ряду по наибольшей частоте определяется модальный интервал, а конкретное значение моды в интервале вычисляется по формуле

$$Mo = x_0 + h \cdot \frac{f_{Mo} - f_{Mo-1}}{\left(f_{Mo} - f_{Mo-1}\right) + \left(f_{Mo} - f_{Mo+1}\right)},$$

где x_0 – нижняя граница модального интервала, h – величина модального интервала, f_{Mo}, f_{Mo-1}, f_{Mo+1} – частоты (или частости) модального, предмодального и послемодального интервалов.

Пример 2: данные о содержании влаги в поступившей партии товара в магазин:

Влажность, % (xi)	Число образцов (fi)
До 14	20
14–16	30
16–18	25
18–20	15
20 и более	10
Итого:	100

Наибольшая частота равна 30 и соответствует интервалу 14–16 (модальный интервал).

$$Mo = 14 + 2 \cdot \frac{30 - 20}{(30 - 20) + (30 - 25)} \approx 15{,}3\%$$

Медиана (Ме) – значение признака (варианта), приходящееся на середину ранжированной (упорядоченной) совокупности, т. е. это вариант, который делит ряд распределения на две равные по объему части.

Медианой Me(X) непрерывной случайной величины X называется такое ее значение, для которого P(X<Me(X)) = P(X>Me(X))=1/2, т. е. вероятность того, что случайная величина X примет значение, меньшее медианы Me(X) или большее ее, одна и та же и равна 1/2. Геометрически вертикальная прямая x=Me(X), проходящая через точку с абсциссой, равной Me(X), делит площадь фигуры под кривой распределения на две равные части. Очевидно, что в точке x=Me(X) функция распределения равна ½, т. е. F(Me(X))=1/2.

Алгоритм нахождения медианы:

- Ряд ранжируют.
- Вычисляют номер медианы $N = \frac{n+1}{2}$ (n – число единиц в совокупности).
- Вычисляют накопленные частоты (или частости).

➢ В **дискретном** ряду

Показатели признака: 3, 7, 2, 5, 2, 3, 9 (n=7, N=4).

Ранжируем: 2, 2, 3, [redacted], 5, 7, 9.

Me=3.

Пример 1: n=100, N=101/2=50,5. Накапливаем частоты до тех пор, пока кумулятивная частость не будет больше или равна номеру N.

Размер женской обуви (x_i)	Число проданных пар, % к итогу (w_i)	Накопленная частость (S_i)
33	4	4
34	12	16
35	18	34
36	26	60
37	20	–
38	13	–
39	6	–
40	1	–
Итого:	100	

Me=36.

Одна половина обуви (50%) была продана 36-го и менее размера, другая половина (50%) – 36-го и более размера.

➢ В **интервальном** ряду значение медианы вычисляется по формуле

$$Me = x_0 + h \cdot \frac{\frac{1}{2}\sum_{i=1}^{m} f_i - S_{Me-1}}{f_{Me}},$$

где x_0 – нижняя граница медианного интервала, h – величина медианного интервала, f_{Me} – частота медианного интервала, S_{Me-1} – накопленная частота предмедианного интервала.

Пример 2: n=100, N=101/2=50,5

Влажность, % (xi)	Число образцов (fi)	Накопленная частота (S_i)
До 14	20	20
14–16	30	50
16–18	25	75
18–20	15	–
20 и более	10	–
Итого:	100	

$$Me = 16 + 2 \cdot \frac{\frac{1}{2} \cdot 100 - 50}{25} = 16.$$

Одна половина (50%) партии товара имеет влажность 16% и менее, другая половина (50%) – 16% и более.

Форма распределения

В симметричных рядах распределения значения моды и медианы совпадают со средней величиной ($\bar{x}$=Mo=Me), в умеренно асимметричных они соотносятся таким образом: 3($\bar{x}$-Me)= $\bar{x}$-Mo.

Для примера 1 $\bar{x}$=36,14; Mo=Me=36 – симметричный ряд.

Для примера 2 $\bar{x}$=16,3; Mo=15,3; Me=16 – умеренно асимметричный ряд.

(3($\bar{x}$-Me)=0,9; $\bar{x}$-Mo=1)

Квартили (Q) – значения вариантов, которые делят упорядоченный ряд по объему на 4 равные части. Выделяют три квартиля. Второй квартиль – это медиана (Me=Q_2).

Первый квартиль $$Q_1 = x_{Q_1} + h \cdot \frac{\frac{1}{4}\sum_{i=1}^{m} f_i - S_{Q_1-1}}{f_{Q_1}}.$$

Третий квартиль $$Q_3 = x_{Q_3} + h \cdot \frac{\frac{3}{4}\sum_{i=1}^{m} f_i - S_{Q_3-1}}{f_{Q_3}}.$$

$$Q_k = x_{Q_k} + h \cdot \frac{\frac{k}{4}\sum_{i=1}^{m} f_i - S_{Q_k-1}}{f_{Q_k}}$$ – общая формула *k*-го квартиля.

Пример 2

Влажность, % (xi)	Число образцов (fi)	Накопленная частота (S_i)
До 14	20	20
14–16	30	50
16–18	25	75
18–20	15	90
20 и более	10	–
Итого:	100	

N_1=25,75; N_2=50,5; N_3=75,25.

$$Q_1 = 14 + 2 \cdot \frac{\frac{1}{4}\cdot 100 - 20}{30} \approx 14{,}3, \quad Q_3 = 18 + 2 \cdot \frac{\frac{3}{4}\cdot 100 - 75}{15} = 18.$$

Вывод: для ранжированных данных 25% товаров содержат процент влажности, не превышающий 14,3%, у 75% товаров процент влажности не превышает 18%.

Обобщение: **Квантилем уровня q** (или q-квантилем) называется такое значение x_q случайной величины, при котором интегральная функция ее распределения принимает значение, равное q, т. е. $F(x_q)=P(X<x_q)=q$.

Очевидно, что введенная выше медиана случайной величины есть квантиль уровня 0,5 (2-й квартиль), т. е. $Me(X)=x_{0,5}$. Квантили $x_{0,25}$ и $x_{0,75}$ (1-й и 3-й квартили) получили название соответственно верхнего и нижнего квантилей.

Начальным моментом k-го порядка случайной величины X называется математическое ожидание k-й степени этой величины $m_k=M(X^k)$.

Центральным моментом k-го порядка случайной величины X называется математическое ожидание k-й степени отклонения случайной величины X от ее математического ожидания. $\mu_k=M[X-M(X)]^k$.

Момент	Случайная величина	
	дискретная	**непрерывная**
Начальный	$m_k = \sum_{i=1}^{n} x_i^k \cdot p_i$	$m_k = \int_{-\infty}^{+\infty} x^k \cdot f(x)\,dx$
Центральный	$\mu_k = \sum_{i=1}^{n} (x_i - a)^k \cdot p_i$	$\mu_k = \int_{-\infty}^{+\infty} (x-a)^k \cdot f(x)\,dx$

В таблице: x_i – значения, которые принимает дискретная случайная величина с вероятностями p_i, f(x) – плотность вероятности непрерывной случайной величины, a – математическое ожидание.

Нетрудно заметить, что при k=1 первый начальный момент случайной величины X есть ее математическое ожидание $m_1=M(X)=a$, при k=2 второй центральный момент – дисперсия $\mu_k=D(X)$.

Центральные моменты могут быть выражены через начальные моменты по формулам:

$\mu_1 = 0$

$\mu_2 = m_2 - (m_1)^2$

$\mu_3 = m_3 - 3m_1m_2 + 2(m_1)^3$

$\mu_4 = m_4 - 4m_1m_3 + 6(m_1)^2m_2 - 3(m_1)^4$

Первый начальный момент (математическое ожидание) характеризует среднее значение или положение распределения случайной величины на числовой оси.

Второй центральный момент (дисперсия) характеризует степень рассеяния распределения случайной величины относительно математического ожидания.

Третий центральный момент служит для характеристики **асимметрии (скошенности) распределения**. Он имеет размерность куба случайной величины. Чтобы получить безразмерную величину, ее делят на куб среднего квадратического отклонения случайной величины. Полученная величина называется **коэффициентом асимметрии** случайной величины $A = \frac{\mu_3}{\sigma^3}$. Если распределение симметрично относительно математического ожидания, то A=0. Кривая, более пологая справа, имеет положительную асимметрию (A>0), более пологая слева – отрицательную (A<0).

Четвертый центральный момент служит для характеристики **крутости (островершинности или плосковершинности) распределения.**

Эксцессом (коэффициентом эксцесса) случайной величины называется число $E = \frac{\mu_4}{\sigma^4} - 3$ (число 3 вычитается, так как для нормального распределения $\frac{\mu_4}{\sigma^4} = 3$). Кривые, более островершинные, чем нормальная, обладают положительным эксцессом, более плосковершинные – отрицательным эксцессом.

2.1.2. Задачи для самостоятельной работы

Задание 1

Один студент получает только 4 и 5. Его средний балл успеваемости равен 4,6. Найти закон распределения оценок студента. Описать и построить график интегральной функции распределения оценок студента.

Задание 2

В магазине имеются 10 телевизоров, из которых 4 с дефектом. Пусть X – случайная величина – число исправных телевизоров среди трех выбранных. Найти закон распределения случайной величины X.

Задание 3

Найти математическое ожидание и дисперсию (двумя способами) дискретной случайной величины, распределенной следующим образом:

x_i	0	1	2
p_i	0.4	0.4	0.2

Задание 4

Непрерывная случайная величина имеет плотность распределения вероятностей $f(x)=\begin{cases}1/k, & x\in(0;4)\\ 0, & x\notin(0;4)\end{cases}$. Найти:

1) Значение параметра k, при котором эта функция действительно может быть функцией плотности.

2) Математическое ожидание.

3) Дисперсию.

4) P(1<=X<=3).

Задание 5

Найти числовые характеристики непрерывной случайной величины, заданной функцией распределения.

$$F(x)=\begin{cases}0, & x\le 0\\ x^2/9, & 0<x\le a\\ 1, & x>a\end{cases}.$$

Задание 6

Пусть случайная величина X имеет плотность $p(x)=\begin{cases}\dfrac{6}{11}(x^2+x+1), & x\in(0;1)\\ 0, & x\notin(0;1)\end{cases}$ Найти математическое ожидание и дисперсию величины X.

Задание 7

Случайная величина X имеет функцию распределения

$F(x)=\begin{cases}0, & x\le 0\\ x^4, & 0<x\le 1\\ 1, & x>1\end{cases}$. Найти математическое ожидание случайной величины X.

Задание 8

Может ли функция $p(x)=\begin{cases}\cos x, & x\in\left[-\dfrac{\pi}{2};\dfrac{\pi}{2}\right]\\ 0, & x\notin\left[-\dfrac{\pi}{2};\dfrac{\pi}{2}\right]\end{cases}$ быть плотностью некоторой случайной величины?

Задание 9

Случайная величина X задана законом распределения

$p_k=P(X=2^k)=\dfrac{3}{4^{k+1}}$, $k=0,1,2,\ldots$ Найти p=P(X>0) и M(X).

Задание 10

В магазин поступила обувь с двух фабрик в соотношении 2:3. Куплено 4 пары обуви. Найти закон распределения числа купленных пар обуви, изготовленных первой фабрикой. Найти M(X), D(X), σ(X).

Задание 11

Для рекламы фирма вкладывает в каждую 10-ю единицу продукции приз в 1000 руб. Пусть X – случайная величина – размер выигрыша при 5 купленных изделиях. Изобразить график функции распределения X и найти M(X).

§ 2.2. Виды распределений случайной величины

2.2.1. Основные понятия

Биномиальное, геометрическое, гипергеометрическое, равномерное, стандартное нормальное распределение, распределение Пуассона. Числовые характеристики случайных величин с различными распределениями.

Законы распределения случайных величин

Биномиальное распределение

Пусть имеются n испытаний Бернулли с вероятностью успеха p и вероятностью неуспеха q, p+q=1. Дискретная случайная величина X – число успехов – имеет распределение $p_k=P(X=k)=C_n^k p^k q^{n-k}, \quad k=0,1,...,n$.

При этом $\sum\limits_{k=0}^{n} p_k=(p+q)^n=1$.

Математическое ожидание и дисперсия случайной величины: $M(X)=np$, $D(X)=npq$.

Геометрическое распределение

Дискретная случайная величина X имеет геометрическое распределение, если она принимает значения k=1,2,3,... (счетное множество значений) с вероятностями $p_k=P(X=k)=pq^{k-1}, \quad 0<p<1, \quad q=1-p$. При этом $\sum\limits_{k=1}^{\infty} p_k=p\dfrac{1}{1-q}=1$.

Случайная величина X, имеющая геометрическое распределение, представляет собой число испытаний Бернулли до первого успеха.

Математическое ожидание и дисперсия случайной величины: $M(X)=\dfrac{1}{p}, D(X)=\dfrac{q}{p^2}$.

Гипергеометрическое распределение

Дискретная случайная величина X имеет гипергеометрическое распределение, если она принимает значения m с вероятностями

$$p_m=P(X=m)=\frac{C_M^m C_{N-M}^{n-m}}{C_N^n}, \quad m=0,1,...,k;\ k=\min(n,M);\ M\le N;\ n\le N.$$

Вероятность p_m является вероятностью выбора m объектов, обладающих заданным свойством, из множества n объектов, случайно извлеченных (без возврата) из совокупности N объектов, среди которых M объектов обладают заданным свойством.

Математическое ожидание и дисперсия случайной величины: $M(X)=n\frac{M}{N}, D(X)=n\frac{M}{N-1}(1-\frac{M}{N})(1-\frac{n}{N})$.

Распределение по закону Пуассона

Дискретная случайная величина X имеет распределение по закону Пуассона, если она принимает значения k=0,1,2,... с вероятностями $p_k=P(X=k)=\frac{e^{-a}a^k}{k!}$, $a>0$ – *параметр распределения.* При этом $\sum\limits_{k=0}^{\infty} p_k=e^{-a}e^{a}=1$.

Математическое ожидание и дисперсия случайной величины: $M(X)=a, D(X)=a$.

Равномерное распределение

Непрерывная случайная величина X распределена равномерно на отрезке [a;b], если ее плотность вероятности p(x) постоянна на этом отрезке и равна нулю вне его, $p(x)=\begin{cases}\frac{1}{b-a}, & a\le x\le b\\ 0, & x<a, x>b\end{cases}$.

Функция распределения случайной величины, распределенной по равномерному закону, имеет вид $F(x)=\begin{cases}0, & x\le a\\ \frac{x-a}{b-a}, & a<x\le b\\ 1, & x>b\end{cases}$.

Математическое ожидание и дисперсия случайной величины: $M(X)=\frac{a+b}{2}, D(X)=\frac{(b-a)^2}{12}$.

Показательное (экспоненциальное) распределение

Непрерывная случайная величина X имеет показательное распределение с параметром $\lambda>0$, если ее плотность вероятности p(x) имеет вид, $p(x)=\begin{cases}\lambda e^{-\lambda x}, & x\ge 0\\ 0, & x<0\end{cases}$.

Функция распределения случайной величины, распределенной по показательному закону, имеет вид $F(x)=\begin{cases}1-e^{-\lambda x}, & x\ge 0\\ 0, & x<0\end{cases}$.

Математическое ожидание и дисперсия случайной величины: $M(X)=\frac{1}{\lambda}, D(X)=\frac{1}{\lambda^2}$.

Нормальное распределение

Непрерывная случайная величина X имеет нормальный закон распределения с параметрами a и σ>0, если ее плотность вероятности p(x) имеет вид, $p(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}$.

Обозначим через N(a;σ) множество случайных величин, распределенных по нормальному закону с параметрами a и σ. Функция распределения нормальной случайной величины X∈N(a;σ) имеет вид $F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int\limits_{-\infty}^{x}e^{-\frac{(t-a)^2}{2\sigma^2}}dt$.

Математическое ожидание и дисперсия случайной величины: $M(X)=a, D(X)=\sigma^2$.

При a=0, σ=1 распределение называется стандартным нормальным. Множество таких распределений обозначается N(0;1). Для стандартного распределения плотность вероятности равна $\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$. Введенная ранее функция Лапласа Ф(x) задает вероятность попадания случайной нормальной величины X в интервал (0,x).

Пусть X∈ N(a;σ), тогда вероятность того, что X принимает значения из интервала (x1,x2), можно вычислить, используя функцию Ф(x), по формуле $P(x_1<X<x_2)=\Phi(\frac{x_2-a}{\sigma})-\Phi(\frac{x_1-a}{\sigma})$.

2.2.2. Задания для самостоятельной работы

1. Всхожесть семян данного растения определяется вероятностью 0,8. Посажено 100 семян. В каких пределах будет расположено число взошедших семян.

2. В большой партии изделий вероятность брака равна *p*. Контроль качества производится до первого появления бракованного изделия. В результате серии проверок обнаружилось, что бракованное изделие впервые появлялось в среднем при 10-м испытании. Оценить вероятность *p*.

3. Среди продукции цеха электронных плат 10 из партии в 100 штук не удовлетворяют стандарту. При приемке продукции проверяются 10 плат. Какое среднее количество нестандартных плат обнаружат?

4. Установлено, что время *T* горения электрической лампочки является случайной величиной, распределенной по показательному закону. Считая, что среднее значение этой величины равно 6 месяцам, найти вероятность того, что лампочка будет гореть не меньше года.

5. Текущая цена ценной бумаги представляет собой нормально распределенную случайную величину X со средним значением 100 усл. ед. и дисперсией 9. Найти вероятность того, что цена актива будет находиться в пределах от 91 до 109 усл. ед.

6. Интервал движения автобуса равен 15 мин. Какова вероятность того, что пассажир на остановке будет ждать автобус не более 5 минут, если время ожидания автобуса имеет равномерное распределение.

7. В ходе проверки аудитор в среднем обнаруживает 3 счета, содержащих ошибки. Число счетов с ошибками распределено по закону Пуассона. Найти вероятность, с которой будет найден один счет с ошибками.

2.2.3. Решение задач с применением EXCEL

Для загрузки пакета анализа в Excel выполните следующие действия:

1. Выполните команду **Сервис \ Надстройки**. На экране появится окно диалога «Надстройки».
2. Выберите **Пакет анализа**, а затем нажмите кнопку **ОК**.

После окончания загрузки в списке опций пункта **Сервис** основного меню появится строка **Анализ данных**. При выборе этой строки появляется окно диалога «Анализ данных».

В окне диалога «Анализ данных» отображается список инструментов.

При статистическом моделировании используется инструмент **Генерация случайных чисел.**

Задание № 1

1. Сгенерируйте значения случайных величин, распределенных

a) равномерно на отрезке [5; 15];

b) нормально со средним a=25 и стандартным отклонением σ=4;

c) биномиально для p=0,2; n=10;

d) по закону Пуассона для λ=5.

Для этого выберите число переменных: 1, число случайных чисел: N=1000, параметры в зависимости от распределения.

Каждое из распределений расположите на отдельном листе Excel.

2. Вычислите с помощью статистических функций СРЗНАЧ() и ДИСПР() математическое ожидание и дисперсию каждой сгенерированной случайной величины.

3. Убедитесь в правильности полученных значений, вычислив математическое ожидание и дисперсию каждой случайной величины по формулам, известным для каждого из видов распределений (см. теорию).

4. Дискретные случайные величины, распределенные биномиально и по закону Пуассона, представьте в табличной форме:

x_i				
p_i				

Для этого отсортируйте 1000 сгенерированных значений и сгруппируйте их, используя функцию СЧЁТЕСЛИ() для вычисления весов f_i, вероятности $p_i=f_i/N$.

5. Исследуйте, как изменяются кривые распределения дискретных случайных величин (графики в одной системе координат(x_i; p_i)) для биномиального распределения в зависимости от p (возьмите значения p, равные 0,2; 0,5; 0,8. Для значения 0,2 данные уже сгенерированы в пункте 1. Для остальных значений выполните генерацию аналогично), для распределения Пуассона в зависимости от λ (возьмите значения λ, равные 1; 8; 15).

6. Исследуйте, как изменяются кривые распределения непрерывных случайных величин (графики плотности распределения в одной системе координат (x; φ(x)) для нормального распределения в зависимости от математического ожидания a и среднего квадратического отклонения σ. Для этого возьмите значения x от –4 до 10 с шагом 0,5. В одной системе координат для одних и тех же значений x постройте четыре графика для различных a и σ.

a	3	3	3	5
σ	1	2	4	1

Плотность нормального распределения выражается формулой

$$\varphi(x,a,\sigma)=\frac{1}{\sigma\cdot\sqrt{2\pi}}\cdot e^{\frac{-(x-a)^2}{2\sigma^2}}$$ = **НОРМРАСП(x, a, σ, ложь)**

Задание № 2

Для сгенерированных в задании № 1 дискретных случайных величин, распределенных по закону Пуассона, вычислите моду, медиану, квартили, коэффициент асимметрии и коэффициент эксцесса двумя способами (с помощью встроенных функций МОДА(), МЕДИАНА(), КВАРТИЛЬ() и с помощью расчетных таблиц и формул из теории), сравните полученные результаты, сделайте выводы о виде кривых распределения для различных значений найденных характеристик.

§ 2.3. Закон больших чисел

2.3.1. Основные понятия

Неравенство Чебышева и закон больших чисел. Теоремы Чебышева и Бернулли, неравенство Маркова. Центральная предельная теорема (без доказательства). Метод Монте- Карло.

Неравенство Чебышева:

$$P(|X-M(X)|<\varepsilon)\geq 1-\frac{D(X)}{\varepsilon^2}$$

Неравенство из теоремы Чебышева:

$$P\left(\left|\frac{X_1+...+X_n}{n}-a\right|<\varepsilon\right)\geq 1-\frac{\frac{c}{n}}{\varepsilon^2},$$

где $X_1,...,X_n$ – попарно независимые случайные величины, дисперсии которых не превышают некоторой константы c, а все математические ожидания равны a; ε – наперед заданное сколь угодно малое положительное число.

Неравенство из теоремы Бернулли:

$$P\left(\left|\frac{m}{n}-p\right|<\varepsilon\right)\geq 1-\frac{\frac{p\cdot q}{n}}{\varepsilon^2},$$

где m – число наступлений события при n независимых испытаниях; p – вероятность наступления события при одном испытании; q – вероятность ненаступления события при одном испытании; ε – наперед заданное сколь угодно малое положительное число.

Неравенство Маркова:

$$\mathrm{P}(\mathrm{X}<\varepsilon)\geq 1-\frac{M(X)}{\varepsilon}$$

Метод Монте-Карло

Название метода связано с названием города Монте-Карло, где в казино играют в рулетку. Рулетка – одно из простейших устройств для получения случайных чисел, на использовании которых основан метод Монте-Карло.

Метод Монте-Карло используют:

- Для вычисления интегралов.
- Для решения систем алгебраических уравнений высокого порядка.
- Для исследования сложных систем: экономических, биологических, социальных.

Суть метода

Требуется найти значение a некоторой изучаемой величины. Для этого выбирается такая случайная величина X, что ее математическое ожидание $M(X)=a$.

Генерируется n значений случайной величины X, вычисляется их среднее арифметическое $\bar{x}=\frac{\sum x_i}{n}$. В качестве оценки искомого числа a принимают $\bar{x}$.

Метод Монте- Карло требует проведения большого числа испытаний n, поэтому называется еще **методом статистических испытаний.**

Отыскание возможных значений случайной величины X (моделирование) называют **разыгрыванием случайной величины.**

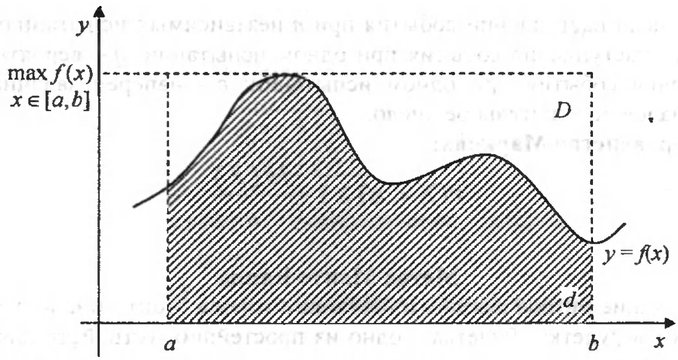
Вычисление определенного интеграла методом статистических испытаний (методом Монте-Карло)

Вычислим методом статистических испытаний следующий интеграл

$$\int_a^b f(x)dx,$$

где функция $y = f(x)$ непрерывна и положительна на отрезке $[a, b]$.

Из курса математики известно, что определенный интеграл $\int_a^b f(x)dx$ равен площади криволинейной трапеции, ограниченной прямыми $x = a$, $x = b$, осью абсцисс и графиком функции $y = f(x)$.



В курсе теории вероятностей приводится несколько определений вероятности. По геометрическому определению в плоскости вероятность попадания в область d точки, брошенной в область D ($d \subset D$), равна отношению площадей, то есть

$$P = \frac{S_d}{S_D}.$$

По статистическому определению вероятность наступления события приближенно равна отношению числа опытов m, в результате которых событие наступило, к общему числу всех опытов n, то есть

$$P = \frac{m}{n}.$$

В геометрическом определении в качестве области d будем рассматривать криволинейную трапецию, в качестве D – прямоугольник $\{a \le x \le b,\ 0 \le y \le \max\limits_{x \in [a,b]} f(x)\}$. В статистическом определении в качестве проводимого опыта возьмем бросание точки в прямоугольник D, а в качестве события – попадание точки в область d. Получаем формулу

$$P = \frac{m}{n} = \frac{S_d}{S_D} = \frac{S_{\text{криволин. трапеции}}}{S_{\text{прямоугольника}}}.$$

Отсюда формула

$$\int_a^b f(x)dx = S_{\text{криволин. трапеции}} = \frac{m}{n} \cdot S_D.$$

Тогда для решения задачи нужно провести достаточное количество опытов – n. Опыт заключается в случайном выборе точки (x_i, y_i) из D:

$$a \le x_i \le b,\ 0 \le y_i \le \max_{x \in [a,b]} f(x).$$

Также необходимо подсчитать количество опытов m, в которых точка (x_i, y_i) принадлежит криволинейной трапеции, то есть $y_i \le f(x_i)$. Тогда интеграл можно вычислить по формуле

$$\int_a^b f(x)dx = \frac{m}{n} \cdot S_D = \frac{m}{n} \cdot (b-a) \cdot \max_{x \in [a,b]} f(x).$$

Результат вычисления интеграла будет тем точнее, чем больше опытов будет проведено.

2.3.2. Задачи для самостоятельной работы

1. Номинальное значение диаметра втулки равно 5 мм, а дисперсия, вследствие погрешностей изготовления, не превосходит 0,01. Оценить вероятность того, что размер втулки будет отличаться от номинала не более чем на 0,5 мм.

2. Определить, сколько надо произвести замеров диаметров деревьев, чтобы средний диаметр деревьев отличался от истинного значения не более чем на 2 см с вероятностью, не меньшей 0,95. Известно, что на данном участке среднее квадратичное отклонение диаметров деревьев не превышает 10 см.

3. Сколько следует провести независимых испытаний, чтобы вероятность выполнения неравенства $\left|\frac{m}{n} - p\right| \le 0{,}05$ превысила 0,75, если вероятность появления данного события в отдельном испытании p=0,8?

4. Опыт страховой компании показывает, что на каждый пятый договор приходится страховой случай. Сколько договоров нужно заключить, чтобы с вероятностью 0,95 можно было утверждать, что доля страховых случаев отклонится от 0,2 не более чем на 0,05?

5. Среднее число дождливых дней в году в данном районе равно 80. Оценить вероятность того, что в этом районе будет не более 100 дождливых дней в году.

6. Известно, что $P(|X-M(X)|<0,1)=0,95$. В каких пределах находится среднее квадратическое отклонение?

7. Известно, что $P(X<1)=0,95$. В каких пределах находится математическое ожидание, если X – неотрицательная случайная величина?

8. Устройство состоит из 10 независимо работающих элементов. Вероятность отказа каждого элемента за время T равна 0,05. Оценить вероятность того, что абсолютная величина разности между числом отказавших элементов и средним числом отказов за время T окажется: а) меньше двух; б) не меньше двух.

9. Дискретная случайная величина X задана законом распределения

X	0,3	0,6
p	0,2	0,8

Оценить вероятность того, что $|X-M(X)|<0,2$.

10. Последовательность независимых случайных величин X_1, X_2, $\ldots, X_n, \ldots$ задана законом распределения

X_n	-na	0	na
p	$1/(2n^2)$	$1-1/n^2$	$1/(2n^2)$

Применима ли к заданной последовательности теорема Чебышева?

11. В каких пределах находится случайная положительная величина X с вероятностью 0,8, если ее математическое ожидание равно 2?

12. Среднее число клиентов, обслуживаемых банком за день, равно 30. Оценить вероятность того, что в течение дня будет обслужено не более 60 человек.

§ 2.4. Двумерные случайные величины

2.4.1. Основные понятия

Пусть $Z=(X,Y)$ – двумерная случайная величина. Тогда числовая функция двух переменных $F(x,y)=P[(X<x)*(Y<y)]$ – функция распределения величины X.

Для дискретной величины Z: $$F(x,y)=\sum_{x_i<x}\sum_{y_j<y} p_{i,j}$$

Свойства:

1. $0<=F(x,y)<=1$
2. Функция, не убывающая по каждому из своих аргументов, т. е.

$$x_1<x_2 \Rightarrow F(x_1,y)\le F(x_2,y)$$
$$y_1<y_2 \Rightarrow F(x,y_1)\le F(x,y_2)$$

3. $F(x,-\infty)=F(-\infty,y)=F(-\infty,-\infty)=0$

4. $F(x,+\infty)=F_1(x)$; $F(+\infty,y)=F_2(y)$, где $F_1(x)=P(X<x)$, $F_2(y)=P(Y<y)$

5. $F(-\infty,+\infty)=1$

Для непрерывной величины Z:

F(x,y) – непрерывна по обоим аргументам.

$$F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y}p(t_1,t_2)dt_1dt_2,$$

где $p(x,y)=\dfrac{\partial^2 F(x,y)}{\partial x\partial y}$ – плотность Z (совместная плотность X и Y).

Свойства:

1. $p(x, y)>=0$

2. $\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}p(x,y)dxdy=1$

3. Если S – прямоугольник ($a1<=x<=a2$, $b1<=y<=b2$), то вероятность попадания двумерной случайной величины Z=(X,Y) в S равна

$$P((X,Y)\in S)=\int_{a1}^{a2}\int_{b1}^{b2}p(x,y)dxdy$$

4. $p(x)=\int_{-\infty}^{+\infty}p(x,y)dy$; $q(y)=\int_{-\infty}^{+\infty}p(x,y)dx$, где p(x) и q(y) – одномерные плотности величин X и Y соответственно.

2.4.2 Задачи для самостоятельной работы

Задание 1

Известно, что вероятность того, что «наугад» взятая пара близнецов одного пола, равна 0,64. Как подсчитать вероятности всех вариантов, если вероятность рождения мальчика равна 0,51?

Задание 2

Пусть двумерная случайная величина (X,Y) задана законом распределения:

xi\yj	1	2	3	4
10	0,2	0,02	0,01	0
20	0,03	0,3	0,02	0
30	0,02	0,1	0,2	0,1

a) Найти законы распределения X и Y.

b) Найти условное распределение Y при условии, что X=30.

Задание 3

Два стрелка независимо один от другого производят по одному выстрелу, каждый по своей мишени. Случайная величина X – число попаданий первого стрелка; Y – второго стрелка. Вероятность попадания в мишень для первого стрелка p1, для второго – p2. Построить функцию распределения F(x,y) системы случайных величин (X,Y).

Задание 4

По мишени производится один выстрел. Вероятность попадания равна p. Рассматриваются две случайные величины: X – число попаданий; Y – число промахов. Построить функцию распределения F(x,y) системы случайных величин (X,Y).

Задание 5

Найти вероятность того, что в результате испытания составляющая X двумерной случайной величины (X,Y) примет значение X<2 и при этом составляющая Y<3, если известна функция распределения системы

$$F(x,y)=(\frac{1}{\pi}arctg\frac{x}{2}+\frac{1}{2})\cdot(\frac{1}{\pi}arctg\frac{y}{3}+\frac{1}{2}).$$

Задание 6

Найти вероятность попадания случайной точки (X,Y) в прямоугольник, ограниченный прямыми $x=\frac{\pi}{6}$, $x=\frac{\pi}{2}$, $y=\frac{\pi}{4}$, $y=\frac{\pi}{3}$, если известна функция распределения F(x,y)=sin x sin y (0<=x<=π/2, 0<=y<= π/2).

Задание 7

Найти плотность совместного распределения f(x,y) системы случайных величин (X,Y) по известной функции распределения

F(x,y)=sin x sin y (0<=x<=π/2, 0<=y<= π/2).

Задание 8

Найти функцию распределения двумерной случайной величины по данной плотности совместного распределения

$$f(x,y)=\frac{1}{\pi^2\cdot(1+x^2)\cdot(1+y^2)}.$$

Задание 9

Плотность распределения двумерной случайной величины $f(x,y)=\frac{1}{\pi^2\cdot(1+x^2)\cdot(1+y^2)}$. Найти вероятность попадания случайной точки в прямоугольник с вершинами K(1;1), L($\sqrt{3}$;1), M(1;0), N($\sqrt{3}$;0).

Задание 10

Задана плотность совместного распределения двумерной случайной величины (X,Y): f(x,y)=C cos x cos y в квадрате 0<=x<=π/2, 0<=y<= π/2; вне этого квадрата f(x,y)=0. Найти постоянный параметр C.

Задание 11

Двумерная случайная величина (X,Y) задана плотностью совместного распределения: $f(x,y)=\begin{cases} 1/(6\pi) & при \quad x^2/9+y^2/4<1 \\ 0 & при \quad x^2/9+y^2/4>1 \end{cases}$. Найти плотности распределения составляющих X и Y.

§ 2.5. *Элементы теории случайных процессов*

2.5.1. Основные понятия

Случайным процессом X(t) называется процесс, значение которого при любом значении аргумента t является случайной величиной (чаще всего t – это время).

Примеры: движение молекул газа во времени, уровень воды в водохранилище.

Случайный процесс можно записать в виде функции двух переменных X(t,ω), где ω – элементарное событие.

Реализацией (траекторией) случайного процесса X(t,ω) называется неслучайная функция x(t), т. е. фиксируется ω.

Сечение случайного процесса X(t,ω) – это случайная величина, полученная при фиксировании t, т. е. $X(t_0,\omega)$ – это случайная величина.

Таким образом, случайный процесс X(t,ω) – это совокупность всех сечений при возможных значениях t, т. е. многомерная случайная величина $(X(t_1); X(t_2); \ldots; X(t_n))$.

Математическое ожидание случайного процесса X(t,ω) – это неслучайная функция от t M[X(t)], которая при любом значении t равна значению математического ожидания соответствующего сечения. Оно характеризует среднюю траекторию всех возможных реализаций случайного процесса.

Дисперсия случайного процесса X(t,ω) – это неслучайная функция от t D[X(t)], которая при любом значении t равна значению дисперсии соответствующего сечения случайного процесса. Дисперсия характеризует разброс реализаций относительно средней траектории.

Среднее квадратическое отклонение случайного процесса X(t,ω) вычисляется как $\sigma[X(t)] = \sqrt{D[X(t)]}$.

Теснота зависимости между двумя сечениями случайного процесса X(t,ω) характеризуется **корреляционной функцией.** Это неслучайная функция от двух переменных $K_X(t_1,t_2) = M[(X(t_1)-M[X(t_1)])(X(t_2)-M[X(t_2)])]$.

Нормированная корреляционная функция

$$\rho(t_1,t_2) = \frac{K_X(t_1,t_2)}{\sigma[X(t_1)]\cdot\sigma[X(t_2)]}$$

Говорят, что в физической системе происходит случайный процесс, если она с течением времени может под влиянием случайных факторов переходить из состояния в состояние.

Если система имеет счетное (в частном случае конечное) множество возможных состояний и переход из одного состояния в другое осуществляется скачком, случайный процесс называется **дискретным**, т. е. X(t) принимает дискретные значения.

Если число состояний бесконечно большое и функция X(t) принимает любые значения из заданного промежутка, то случайный процесс называется **непрерывным**.

Если переходы системы из состояния в состояние возможны только в определенные моменты времени (аргумент t дискретный), то случайный процесс называется **случайной последовательностью (процессом с дискретным временем).**

Если переходы системы из состояния в состояние возможны в любой момент времени (аргумент t непрерывный), то случайный процесс называется **процессом с непрерывным временем.**

Случайный процесс называется **Марковским**, если для любого момента времени вероятностные характеристики процесса в будущем зависят только от его состояния в данный момент и не зависят от того, когда и как система пришла в это состояние («будущее определяется только настоящим»).

Марковский случайный процесс с дискретным числом состояний и с дискретным временем называется **Марковской цепью**.

Пусть $p_{ij}(s)$ – условная вероятность того, что после s-го испытания система будет находится в j-м состоянии, если после (s-1)- го испытания она находилась в i-м состоянии.

Цепь Маркова называется **однородной**, если вероятность $p_{ij}(s)$ не зависит от номера испытания s. Тогда p_{ij} называют **переходной вероятностью**, $\Pi=(p_{ij})$ – **матрицей переходов**. Заметим, что для любого номера состояния i выполняется условие $\sum_{j=1}^{k} p_{ij} = 1$ (k – число состояний).

Уравнения Колмогорова – Чемпена

$$P_n=P_{n-1}\cdot\Pi \qquad\qquad P_n=P_0\cdot\Pi^n$$

P_n – вектор вероятностей после n-го шага

P_{n-1} – вектор вероятностей после (n-1)-го шага

Π – матрица переходов

Π^n – матрица переходов через n шагов

P_0 – вектор начальных вероятностей

Теорема Маркова (о предельных вероятностях)

Если при некотором n все элементы матрицы перехода Π^n положительны, то существуют такие постоянные числа p_j^* $(j=1,..,k)$, что независимо от индекса i верно $\lim\limits_{n\to\infty} p_{ij}(n) = p_j^*$ (k – число состояний).

$(p_1^*, p_2^*, \ldots, p_k^*)$ – предельные вероятности.

Они находятся как решение системы $(x_1, x_2, \ldots, x_k)$

$$\begin{cases} p_{11}\cdot x_1 + p_{21}\cdot x_2 + \ldots + p_{k1}\cdot x_k = x_1 \\ p_{12}\cdot x_1 + p_{22}\cdot x_2 + \ldots + p_{k2}\cdot x_k = x_2 \\ \ldots \\ p_{1k}\cdot x_1 + p_{2k}\cdot x_2 + \ldots + p_{kk}\cdot x_k = x_k \end{cases}$$

Причем $\sum\limits_{i=1}^{k} x_i = 1$

Предельная вероятность p_j^* $(j=1,..,k)$ показывает среднее относительное время пребывания системы в j-м состоянии.

2.5.2. Задачи для самостоятельной работы

1. Пусть A1, A2, A3 – точки числовой оси с целочисленными координатами x=1, x=2, x=3.Частица движется по этим точкам следующим образом: если в какой-то момент времени t= 0,1,2,... частица находится в точке A2, то в следующий момент t +1 она переходит в A1 с вероятностью q(0<q<1) или в A3 с вероятностью p=1-q; если частица оказалась в левой граничной точке A3, то в следующий момент времени она там остается с вероятностью p или возвращается в A2 с вероятностью q; если частица оказалась в правой граничной точке A1, то в следующий момент времени она там остается с вероятностью q или возвращается в A2 с вероятностью p.

Найти:

1. Матрицу переходов.
2. Матрицу переходов за 2 шага.
3. Проверить существование предельных вероятностей.
4. Найти предельные вероятности, если они существуют.

2. Изменим задачу 2 так, что граничные точки A1 и A3 будут поглощающими. Найдем матрицу переходов, матрицу переходов за 2 шага.

3. Внесем изменение в задачу 2: пусть частица, находясь во внутренней точке A2, в следующий момент времени с одной и той же вероятностью или переходит в A1, или остается на месте, или переходит в точку A3. Если частица достигает любой из граничных точек A1 или A3, то она остается на месте с вероятностью или возвращается во внутреннюю точку A2 с равной вероятностью. Найти матрицу переходов и матрицу переходов за 2 шага; показать применимость к данной цепи теоремы Маркова и найти предельные вероятности.

4. В урне содержится 5 шаров, белые и черные. Испытание состоит в том, что каждый раз из урны случайно вынимается один шар и взамен в урну возвращается шар, но другого цвета. Найти матрицу переходных вероятностей для цепи Маркова, состояниями которой является количество белых шаров в урне. Найти вероятности перехода за два шага. Найти предельные вероятности. Найти вероятности после двух шагов, если известно, что первоначально в урне с равной вероятностью могло оказаться 2 белых, 3 черных или 3 белых, 2 черных.

5. Модель перемешивания колоды карт. Пусть имеется 3 карты с номерами 1, 2, 3. Состоянием системы назовем последовательность номеров этих карточек a1a2a3. Перемешивание происходит так: с равной вероятностью состояние a1a2a3 переходит в a3a1a2 или a1a3a2. Найти матрицу вероятностей перехода. Найти предельные вероятности.

§ 2.6. Проверочный тест

1. Доля взрослых среди 5000 зрителей, присутствующих в данный момент во Дворце спорта, – это

a) случайная величина b) случайное событие

2. Значение функции распределения случайной величины X от аргумента a равно

a) P(X=a) b) P(X<a) c) P(X<=a) d) P(X>a) e) P(X>=a)

3. Каким является множество значений непрерывной случайной величины?

a) бесконечным b) конечным c) счетным

4. Случайная величина X – количество раз, когда выпало 3 очка при однократном подбрасывании игрального кубика. Чему равно значение функции распределения F(1/3)?

a) 0 b) 1 c) 1/6 d) 5/6 e) 1/3 f) 3

5. Пусть случайные величины X и Y независимы и заданы законами распределения

xi	0	1	2	3
pi	0.2	0.4	0.3	0.1

yj	-1	0	1
pj	0.4	0.2	0.4

Случайная величина Z=X·Y. Найти P(Z=0)

a) 0 b) 0.2 c) 0.04 d) 0.36

6. Функция распределения непрерывной случайной величины X – это

a) $\int\limits_{-\infty}^{x} p(t)dt$ b) $P(a \le X \le b)$ c) $\int\limits_{a}^{b} p(x)dx$ d) $p'(x)$ e) $\int\limits_{-\infty}^{+\infty} p(x)dx$

где p(x) – плотность распределения случайной величины X.

7. Пусть закон распределения случайной величины X имеет вид $p_i=P(X=x)$. Случайная величина $Z=kX$, где k – постоянная величина. Найти $P(Z=kx_1)$.

a) $\sum_{i=1}^{k} p_i$ b) k c) p_1 d) kp_1

8. Найти математическое ожидание и дисперсию случайной величины, распределенной следующим образом:

xi	0	1	2
pi	0.4	0.4	0.2

9. Чему равно наименьшее значение функции распределения?

10. Какая характеристика оценивает степень рассеивания случайной величины?

a) Математическое ожидание
b) Дисперсия
c) Среднее квадратическое отклонение

11. Какое распределение имеет случайная величина, представляющая собой число испытаний до первого успеха?

a) Биномиальное
b) Геометрическое
c) Гипергеометрическое
d) Равномерное

12. Дано распределение случайной величины. Чему равно значение функции распределения от аргумента, равного 2?

xi	-1	2	3
pi	1/8	1/2	3/8

a) 1/8 b) 3/8 c) 1/2 d) 1 e) 5/8

13. При каком значении C функция f(x) является плотностью вероятности некоторой случайной величины?

$$f(x)=\begin{cases} C\cdot e^{-2x}, & x>0 \\ 0, & x<0 \end{cases}$$

14. Для случайной величины, имеющей плотность вероятности $f(x)=\begin{cases} 1/4, & x\in(0;4) \\ 0, & x\notin(0;4) \end{cases}$. Найти

a) Математическое ожидание
b) Дисперсию
c) $P(1<=X<=3)$

15. Студент получает оценки от 3 до 5. Его средний балл успеваемости равен 4,2. Какой процент удовлетворительных оценок (3) получает студент, если с одинаковой вероятностью он может получить 4 и 5.

16. В экзаменационном билете один теоретический вопрос и одна задача. Распределение оценок студента занесено в таблицу.

Задача \ Теор.вопр.	4	5
3	0.09	0.12
4	0.18	0.11
5	0.3	0.2

a) С какой вероятностью за теоретический вопрос студент получит 4?

b) С какой вероятностью за теоретический вопрос студент получит 4 при условии, что задача решена на 5?

Ключ для проверки

1	a	**5**	d	**9**	0	**13**	c=2
2	b	**6**	a	**10**	b,c	**14**	M(X)=2 D(X)=4/3 P(1<=X<=3)=0,5
3	a	**7**	c	**11**	b	**15**	20%
4	d	**8**	M(X)=0,8 D(X)=0,56	**12**	a	**16**	a)0,57 b)0,6

Глава 3. Математическая статистика

§ 3.1. Характеристики вариационного ряда

3.1.1. Основные понятия

Задачи математической статистики. Точечные и интервальные оценки. Понятие о несмещенности, состоятельности и эффективности оценок.

Вариационный ряд, его геометрическое изображение (полигон, гистограмма, кумулята, огива), абсолютные и относительные показатели вариации.

3.1.1.1. Задачи математической статистики

Математическая статистика – это часть прикладной математической дисциплины «Теория вероятностей и математическая статистика», которая изучает случайные явления, использует одинаковые с теорией вероятностей методы и понятия и основана на аксиоматике А. Н. Колмогорова.

Задачи математической статистики:

1. Определение способов сбора и группировки статистической информации.

2. Разработка методов анализа статистических данных, соответствующих целям исследования.

Сравнение характеристик областей применения аппарата теории вероятностей и математической статистики

Теория вероятностей	Математическая статистика
Модель, описывающая изучаемое явление или объект, известна до опыта. Есть сведения обо всей генеральной совокупности, описывающей исследуемое явление	Модель, описывающая изучаемое явление или объект: • либо неизвестна до опыта, тогда для определения модели проводятся пробные испытания, т. е. формируется выборка из генеральной совокупности; • либо модель задана до опыта с точностью до неизвестных параметров, значение которых получают приближенно по выборке
Выводы о поведении исследуемого объекта или явления делаются по всей генеральной совокупности	Выводы о поведении исследуемого объекта или явления делаются по выборке ограниченного объема и распространяются на всю генеральную совокупность

Этапы решения задачи описания эмпирических (полученных в результате опыта) данных вероятностными моделями

1. Предварительная обработка данных выборки из генеральной совокупности (анализ объема выборки, засоренность выборки, независимости элементов выборки).

2. Оценивание (точечное и интервальное) числовых и функциональных характеристик случайных величин.

3. Выбор типа вероятностной модели, описывающей эмпирические данные.

4. Точечное и интервальное оценивание неизвестных параметров модели.

5. Проверка гипотез о согласии модели и эмпирического распределения.

Все **методы математической статистики** можно разделить:

- на **параметрические** методы, основаные на использовании знаний о вероятностной модели (применяют для выборки объема не менее 60);
- **непараметрические** – используются, если модель неизвестна до опыта.

Статистика как наука исследует не отдельные факты, а массовые социально-экономические явления и процессы, выступающие как множество отдельных фактов, обладающих как индивидуальными, так и общими признаками.

Выделим здесь три основных момента:

1) Это **множество** явлений, а не одно.

2) Эти явления объединены общим качеством, представляют собой одну и ту же **закономерность.**

3) Эти явления **варьирующие**, т. е. отличающиеся по своим характеристикам.

Закономерность – повторяемость, последовательность и порядок изменений в явлениях.

Вариация – колеблемость, многообразие, изменяемость значения признака у отдельных единиц совокупности явлений.

Именно наличие вариации предопределяет необходимость статистики.

Закономерности, в которых необходимость неразрывно связана в каждом отдельном явлении со случайностью и лишь во множестве явлений проявляет себя как закон, называются **статистическими**.

Свойство статистических закономерностей проявляется лишь при обобщении данных по достаточно большому числу исследуемых единиц, что позволяет сделать **закон больших чисел.**

Динамическая закономерность проявляется в каждом отдельном явлении. Например, площадь круга зависит от его радиуса. Эта связь выражается формулой и справедлива для любого круга.

Оценки неизвестных параметров

Если установлено, какое именно распределение имеет изучаемый признак в генеральной совокупности, возникает задача оценки параметров, которыми определяется это распределение.

Например, если известно, что признак распределен в генеральной совокупности нормально, то необходимо оценить (найти приближенное значение) математическое ожидание (a) и среднее квадратическое отклонение (σ). Эти два параметра полностью определяют нормальное распределение. Для распределения Пуассона необходимо оценить один параметр, которым оно определяется: $\lambda=a=M(X)=D(X)$.

Пусть по выборке $X_1, X_2, \ldots, X_n$, полученной в результате n наблюдений (опытов), требуется оценить неизвестный параметр θ.

Заметим, что $X_1, X_2, \ldots, X_n$ – это случайные величины (X_i – результат i-го наблюдения), которые имеют такое же распределение, что и случайная величина X (вся генеральная совокупность). Конкретная выборка $x_1, x_2, \ldots, x_n$ (вариационный ряд) – это значения независимых случайных величин $X_1, X_2, \ldots, X_n$

Статистической оценкой θ* параметра θ теоретического распределения называют его приближенное значение, зависящее от выборки.

Оценка θ* параметра θ – это значение некоторой функции (функции выборки, или **статистики**) от результатов наблюдений над случайной величиной X. Следовательно, оценка θ* параметра θ – это статистика, которая близка к истинному значению параметра θ.

Несмещенной называют статистическую оценку θ*, математическое ожидание которой равно оцениваемому параметру θ при любом объеме выборки.

$$M(\theta^*)=\theta.$$

Смещенной называют статистическую оценку θ*, если $M(\theta^*)\neq\theta$.

Несмещенная оценка θ* параметра θ называется **эффективной**, если она имеет наименьшую дисперсию среди всех возможных несмещенных оценок параметра θ, т. е. оценка θ* эффективна, если ее дисперсия минимальна.

Состоятельной называют статистическую оценку, которая при $n\to\infty$ стремится по вероятности к оцениваемому параметру, т. е. для $\forall\varepsilon>0$ выполняется $\lim\limits_{n\to\infty} P\left(\left|\theta^*-\theta\right|<\varepsilon\right)=1$.

С увеличением объема выборки θ* все больше приближается к истинному значению параметра θ, т. е. практически достоверно $\theta^*\approx\theta$.

Свойство состоятельности обязательно для любого правила оценивания. Несостоятельные оценки не используются.

Состоятельность оценки θ* может быть установлена с помощью следующей теоремы: «Если оценка θ* параметра θ является несмещенной и $D(\theta^*)\to 0$ при $n\to\infty$, то θ* – состоятельная оценка».

Точечной называют оценку, которая определяется одним числом. При выборке малого объема точечная оценка может значительно отличаться от оцениваемого параметра, т. е. приводить к грубым ошибкам.

Точечные оценки

- В качестве оценки генеральной средней (а) принимают **выборочную среднюю** ($\bar{x}=\frac{\sum x_i \cdot f_i}{n}$). Она является **несмещенной** и **состоятельной** оценкой генеральной средней.

Свойство устойчивости выборочных средних: если по нескольким выборкам достаточно большого объема из одной и той же генеральной совокупности будут найдены выборочные средние, то они будут приближенно равны между собой.

- Пусть требуется по данным выборки оценить (найти приближенное значение) неизвестную генеральную дисперсию σ^2.

Выборочная дисперсия $S^2=\frac{\sum(x_i-\bar{x})^2 \cdot f_i}{\sum f_i}$ ($\sum f_i = n$) является **смещенной** оценкой генеральной дисперсии (т. е. $M(S^2)\neq\sigma^2$).

Исправленная дисперсия $S^2_{испр.}=\frac{n}{n-1}\cdot S^2=\frac{n}{n-1}\cdot\frac{\sum(x_i-\bar{x})^2 \cdot f_i}{\sum f_i}=\frac{\sum(x_i-\bar{x})^2 \cdot f_i}{n-1}$

является **несмещенной** оценкой генеральной дисперсии.

Следовательно, в качестве оценки генеральной дисперсии принимают исправленную дисперсию. При достаточно больших объемах выборки исправленная и выборочная дисперсии различаются незначительно. На практике исправленной дисперсией пользуются, если объем выборки меньше 30 (малая выборка).

- Для оценки среднего квадратического отклонения генеральной совокупности используют **исправленное среднее квадратическое отклонение**, равное квадратному корню из исправленной дисперсии, но оно не является несмещенной оценкой.

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

Интервальные оценки позволяют установить точность и надежность оценок. Пусть найденная по данным выборки статистическая характеристика θ^* служит оценкой неизвестного параметра θ. Тогда положительное число δ, такое, что $|\theta^*-\theta| < \delta$ характеризует **точность оценки**. Чем меньше δ, тем точнее оценка θ^*.

Надежностью (доверительной вероятностью) γ оценки θ^* называют вероятность, с которой выполняется неравенство $|\theta^*-\theta| < \delta$.

$$\gamma=P(|\theta^*-\theta| < \delta)$$

Обычно надежность γ задается и имеет значение, близкое к 1. Часто задают надежность 0,95; 0,99; 0,999.

Доверительным называют интервал (θ^*-δ; θ^*+δ), который покрывает неизвестный параметр θ с заданной надежностью γ.

3.1.1.2. Вариационный ряд. Его основные показатели

Задача математической статистики состоит в создании методов сбора и обработки статистических данных для получения научных и практических выводов.

Статистические данные (т. е. полученные из наблюдений или опытов числа, отражающие значения изучаемого признака) записываются в форме ***вариационного ряда***.

Примеры вариационных рядов:

1. Опыт – измерение роста 10 человек (изучаемый признак – рост)

152	160	156	156	167	170	155	152	167	165

2. Опыт – подсчет максимально возможного количества станков, которое может обслужить один ткач (изучаемый признак – число станков)

Количество станков	2	4	6
Число ткачей	2	64	154

3. Опыт – взвешивание телят (изучаемый признак – вес)

Вес	100–119	120–139	140–159	160–179
Число телят	2	20	60	18

Значение изучаемого признака в вариационном ряде называют ***вариантой*** и обозначают x_i. В примере № 1 вариационный ряд состоит только из вариант, в примере № 2 вариантами является количество станков, в примере № 3 – вес (варианты могут задаваться в виде интервалов, ***при этом для расчетов берут среднее значение интервала***).

Число единиц наблюдения, обладающих одинаковым значением изучаемого признака, в вариационном ряде называют ***частотой (весом)*** и обозначают f_i. В примере № 1 все веса равны 1 ($\forall i \quad f_i=1$) и поэтому они не указываются, в примере №2 весами является число ткачей, в примере № 3 – число телят.

Общее число единиц наблюдения называют ***объемом*** изучаемой совокупности и обозначают n. Для примера № 1 n=10, для примера № 2 n=220, для примера № 3 n=100.

Для обработки данных часто используют среднее значение признака, т. е. среднее арифметическое чисел x_i. Если в вариационном ряде не указаны веса ($\forall i \quad f_i = 1$), то находят простое среднее ($\bar{x}$). ***Если веса указаны, находят взвешенное среднее*** ($\bar{x}_{вз.}$).

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \qquad \bar{x}_{вз.} = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{n} \quad (n = \sum_{i=1}^{k} f_i)$$

(n – число изучаемых в совокупности единиц, k – число интервалов или отдельных различных значений изучаемого признака в вариационном ряду, образовавшихся после группировки статистических данных).

Вариация – колеблемость, многообразие, изменяемость величины признака у отдельных единиц совокупности.

Основные показатели вариационного ряда (вариации) – обобщенные характеристики степени колеблемости (вариации) признака в совокупности

Абсолютные	**Относительные**
❖ Размах вариации	❖ Коэффициент осцилляции
❖ Среднее линейное отклонение	❖ Коэффициент вариации
❖ Дисперсия	❖ Линейный коэффициент вариации
❖ Среднее квадратическое отклонение	

Размах вариации (R) – разность между наибольшим и наименьшим значениями изучаемого признака.

$$R = x_{\max} - x_{\min}$$

Среднее линейное отклонение ($\bar{d}$) – среднее арифметическое абсолютных значений отклонений вариант признака от средней величины признака. Как и среднее, может быть простым и взвешенным.

$$\bar{d} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}, \qquad \bar{d}_{вз.} = \frac{\sum_{i=1}^{k} |x_i - \bar{x}_{вз.}| \cdot f_i}{n} \quad (n = \sum_{i=1}^{k} f_i).$$

Дисперсия (σ^2) – средняя величина квадратов отклонений индивидуальных значений признака от средней величины признака.

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}, \qquad \sigma^2_{вз.} = \frac{\sum_{i=1}^{k} (x_i - \bar{x}_{вз.})^2 \cdot f_i}{n} \quad (n = \sum_{i=1}^{k} f_i).$$

Если все единицы наблюдения в изучаемой совокупности разбиты на группы, то появляются понятия групповой, межгрупповой и внутригрупповой дисперсий.

Групповой дисперсией называют дисперсию значений признака, принадлежащих данной группе, относительно среднего значения признака внутри этой группы.

$$\sigma^2{}_{j\,гр.} = \frac{\sum\limits_i \left(x_i - \bar{x}_{j\,гр.}\right)^2 \cdot f_i}{n_{j\,гр.}}$$ – групповая дисперсия j-й группы.

(где i – номера лишь тех единиц совокупности, которые содержатся в j-й группе; $n_{j\,гр}$ - объем j-й группы, $\bar{x}_{j\,гр.}$ - среднее значение признака внутри j-й группы).

Внутригрупповой дисперсией называют среднее арифметическое всех групповых дисперсий данной совокупности.

$$\sigma^2{}_{вн.\,гр.} = \frac{\sum\limits_j n_{j\,гр.} \cdot \sigma^2{}_{j\,гр.}}{n}$$ (где n – объем всей совокупности, $\sigma^2{}_{j\,гр.}$ – групповая дисперсия j-й группы, j – номера групп).

Межгрупповой дисперсией называют дисперсию средних значений признака в каждой группе относительно среднего значения признака во всей совокупности.

$$\sigma^2_{межгр.} = \frac{\sum\limits_j \left(\bar{x}_{j\,гр.} - \bar{x}\right)^2 \cdot n_{j\,гр.}}{n}$$

(где n – объем всей совокупности, j – номера групп, $\bar{x}_{j\,гр.}$ – среднее значение признака внутри j-й группы, $n_{j\,гр.}$ - объем j-й группы).

Общей дисперсией называют дисперсию значений признака всей совокупности относительно среднего значения признака во всей совокупности.

$$\sigma^2 = \frac{\sum\limits_{i=1}^{k} \left(x_i - \bar{x}\right)^2 \cdot f_i}{n}$$

Теорема

Если совокупность состоит из нескольких групп, то общая дисперсия равна сумме внутригрупповой и межгрупповой дисперсий:

$$\sigma^2 = \sigma^2{}_{вн.гр.} + \sigma^2{}_{межгр.}$$

Среднее квадратическое отклонение (σ) – квадратный корень из дисперсии $\sigma = \sqrt{\sigma^2}$. Оно выражается в тех же единицах измерения, что и сам признак (дисперсия же выражается в квадратных единицах измерения признака). Найдя σ, все единицы наблюдения можно разбить на три группы:

- Со средним значением признака: $\bar{x} - \sigma \le x_i \le \bar{x} + \sigma$
- Со значением признака выше среднего: $x_i > \bar{x} + \sigma$

- Со значением признака ниже среднего: $x_i < \bar{x} - \sigma$

Относительные признаки вариации используются при сравнении колеблемости различных признаков в одной совокупности или одного признака в нескольких совокупностях.

Коэффициент осцилляции (V_R) – процентное отношение размаха вариации к средней величине признака.

$$V_R = \frac{R}{\bar{x}} \cdot 100\%$$

Коэффициент вариации (V_σ) – процентное отношение среднего квадратического отклонения к средней величине признака. Если коэффициент вариации не превышает 33%, то совокупность считается ***однородной***.

$$V_\sigma = \frac{\sigma}{\bar{x}} \cdot 100\%$$

Линейный коэффициент вариации ($V_{\bar{d}}$) – процентное отношение среднего линейного отклонения к средней величине признака.

$$V_{\bar{d}} = \frac{\bar{d}}{\bar{x}} \cdot 100\%$$

Анализ рядов распределения наглядно можно проводить на основе их графического изображения. Для этой цели строят графики – полигон, гистограмма, огива, кумулята.

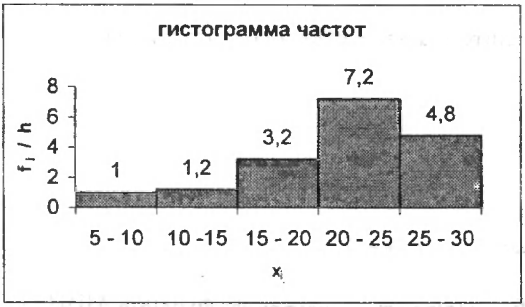
Полигон – ломаная линия, отрезки которой соединяют точки (x_1,f_1), (x_2,f_2), . . ., (x_n,f_n), где x_i – варианты, f_i – частоты. Используется для изображения дискретных вариационных рядов (варианты принимают только целые значения).

Полигоном относительных частот называют ломаную, отрезки которой соединяют точки (x_1,w_1), (x_2,w_2), . . .,(x_n,w_n), где x_i – варианты, w_i – относительные частоты ($w_i = \frac{f_i}{n}$).

Гистограммой называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат интервалы значений признака длиною h, а высоты пропорциональны частотам соответствующих интервалов значений. Используется для изображения интервальных вариационных рядов (варианты заданы интервалами). При построении гистограммы вариационного ряда с равными интервалами по оси ординат наносят частоты, с неравными интервалами – плотность распределения (частота интервала, деленная на длину интервала).

Если середины верхних сторон прямоугольников соединить ломаной, то получим полигон.

Пример



x_i	f_i/h
5–10	1
10–15	1,2
15–20	3,2
20–25	7,2
25–30	4,8

h= 5

При помощи **кумуляты** изображается ряд накопленных частот ($m_k = \sum_{i=1}^{k} f_i$). Это ломаная линия, отрезки которой соединяют точки (x_1,m_1), (x_2,m_2), . . ., (x_n,m_n), где x_i – варианты, m_i – накопленные частоты.

Если при графическом изображении вариационного ряда в виде кумуляты оси поменять местами, получим **огиву**.

3.1.2. Решение задач с применением EXCEL

Выполните решение предложенных ниже задач в электронных таблицах Excel.

1. Имеются следующие данные об обслуживании покупателей в магазине (изучаемый признак – время обслуживания).

Время обслуживания (мин)	10–12	12–14	14–16	16–18	18–20	Итого
Число покупателей (чел.)	2	5	19	13	11	50

Определить все известные Вам показатели предложенного вариационного ряда.

Изобразите полигон частот и гистограмму относительных частот для данного вариационного ряда.

Указания к решению

Оформите решение в виде расчетной таблицы:

f_i	X_i	f_i*X_i	$\|X_i-X_{ср}\|$	$\|X_i-X_{ср}\|*f_i$	$(X_i-X_{ср})^2$	$(X_i-X_{ср})^2*f_i$

В последней строке таблицы в некоторых столбцах найдите сумму чисел столбца.

Ниже таблицы вычислите значение следующих показателей:

X ср =

R=

Ср. лин. откл.=

Дисперс.=

Ср. квадр. откл.=

Коэф. осцил.=

Лин. коэф. вариац. =

Коэф. вариац. =

Для этого используйте формулы, содержащие функции МИН, МАКС, ABS, КОРЕНЬ, СУММ, абсолютные и относительные ссылки на ячейки и копирование формул.

Рассмотрите встроенные в Excel функции из категории статистические. Найдите значение некоторых показателей (для которых это возможно) вторым способом, т. е. используя готовые функции. Сравните результаты. В задачах какого типа можно использовать готовые статистические функции?

Для построения гистограммы и полигона используйте мастер диаграмм. Изменяя параметры диаграмм, добейтесь правильного и наглядного их изображения.

2. Совместные предприятия одной из отраслей сгруппированы по стоимости реализованной продукции и услуг за год. Результаты представлены в следующей таблице.

Стоимость реализованной продукции и услуг (млрд руб.)	**Число предприятий**	**Средняя стоимость реализованной продукции и услуг по группе (млрд руб.)**	**Групповая дисперсия**
3,5–6,5	9	5,59	6,13
6,5–9,5	10	7,06	6,51
9,5 и выше	11	12,20	72,16

Определите:

- Внутригрупповую дисперсию
- Межгрупповую дисперсию
- Общую дисперсию

3. Известны оценки за контрольную работу 9 учащихся.

1) Найти средний балл, полученный за контрольную работу, и степень колеблемости оценок (т. е. дисперсию).

2) Сгруппировать данные, выделив 2 группы: девушки и юноши. Оценить средний балл и колеблемость оценок в каждой группе.

3) Выполнить проверку полученных данных с помощью правила сложения дисперсий.

Абрамов Ю. П. – 5	Боброва Т. С. – 3	Гришин У. К. – 2
Ананьева Л. И. – 2	Бровкин Е. А. – 4	Дронова Т. И. – 5
Бирюков В. И. – 4	ВикуловВ. И. – 3	Прохина В. П. – 4

Используйте мастер функций (готовые статистические функции).

§ 3.2. Выборочное наблюдение

3.2.1. Основные понятия

Генеральная совокупность и выборка. Случайность и репрезентативность выборки.

Оценка параметров генеральной совокупности по ее выборке. Конечные оценки генеральной средней и генерального среднего квадратичного отклонения. Доверительные интервалы для параметров нормального распределения.

Определение оптимального объема выборки.

Определение границ показателей генеральной совокупности

Выборочное наблюдение – это такое несплошное наблюдение, при котором отбор подлежащих обследованию единиц осуществляется в случайном порядке, отобранная часть изучается, а результаты распространяются на всю исходную совокупность.

Совокупность, из которой производится отбор, называется **генеральной**, а совокупность отобранных для обследования единиц – **выборочной.**

Символы основных характеристик параметров генеральной и выборочной совокупностей

Изучаемый признак называется **количественным**, если его значениями являются числа. **Альтернативный** признак имеет лишь два значения (выполняется или не выполняется условие). Например, нестандартность изделий, малообеспеченность семей, всхожесть семян (либо есть, либо нет).

Количественный признак

Характеристика	**Генеральная совокупность**	**Выборочная совокупность**
Объем совокупности (численность единиц)	N	$n = \sum f_i$
Средний размер признака	a	$\bar{x} = \dfrac{\sum x_i \cdot f_i}{n}$
Дисперсия количественного признака	σ^2	$S^2 = \dfrac{\sum (x_i - \bar{x})^2 \cdot f_i}{n}$

Альтернативный признак

Характеристика	**Генеральная совокупность**	**Выборочная совокупность**
Численность единиц, обладающих обследуемым признаком	M	m
Доля единиц, обладающих обследуемым признаком	$p=\frac{M}{N}$	$w=\frac{m}{n}$
Дисперсия доли	$\sigma^2 = p\cdot(1-p)$	$s^2 = w\cdot(1-w)$

Основная задача выборочного наблюдения состоит в том, чтобы на основе характеристик выборочной совокупности получить достоверные суждения о показателях в генеральной совокупности. При этом следует иметь в виду, что при любых статистических исследованиях возникают ошибки двух видов: **регистрации и репрезентативности.**

Ошибки регистрации можно избежать при правильной организации и проведении наблюдения.

Ошибки репрезентативности присущи только выборочному наблюдению и возникают в силу того, что выборочная совокупность не полностью воспроизводит генеральную. Они представляют собой расхождение между значениями показателей, полученных по выборке, и значениями показателей этих же величин, которые были бы получены при сплошном наблюдении, т. е. между величинами выборных и соответствующих генеральных показателей.

Ошибка выборки (ошибка репрезентативности) вычисляется как модуль разности соответствующих генеральных и выборочных показателей:

- для среднего значения (количественный признак) $\varepsilon = |\bar{x} - a|$;
- для доли (альтернативный признак) $\varepsilon = |p - w|$.

Так как для каждой конкретной выборки в генеральной совокупности вычисляется своя ошибка выборки, то принято находить среднюю из этих ошибок.

Средняя ошибка выборки обозначается μ и зависит:

➢ от числа единиц в выборочной совокупности n (чем больше n, тем меньше μ);

➢ от степени вариации значения признака σ^2 (чем больше σ^2, тем больше μ).

Таким образом, μ – есть зависимость от n и от σ^2, т. е. μ=f(n, σ^2).

Эта формула зависит от вида выборки и от метода отбора единиц в выборочную совокупность.

***Методы*:**

- повторная («отбор по схеме возвращенного шара»),
- бесповторная («отбор по схеме невозвращенного шара»).

Виды:

- собственно-случайная,
- механическая,
- типическая,
- серийная.

Собственно-случайная выборка состоит в том, что выборочная совокупность образуется в результате случайного отбора отдельных единиц из генеральной совокупности. При этом количество отобранных в выборочную совокупность единиц обычно определяется исходя из принятой доли выборки.

Доля выборки есть отношение числа единиц выборочной совокупности n к численности единиц генеральной совокупности N, т. е. $\frac{n}{N}$.

Так, при 5%-ной выборке из партии товара в 2000 ед. численность выборки n составляет 100 ед. ($\frac{5 \cdot 2000}{100}$), а при 20%-ной выборке она составит 400 ед. ($\frac{20 \cdot 2000}{100}$) и т. д.

Важным условием репрезентативности собственно-случайной выборки является то, что каждой единице генеральной совокупности предоставляется равная возможность попасть в выборочную совокупность.

Одним из примеров использования собственно-случайной выборки является проведение тиражей лотереи, при которых обеспечивается равная возможность попадания в тираж любого номера лотерейного билета.

Механическая выборка применяется в случаях, когда генеральная совокупность каким-либо образом упорядочена, т. е. имеется определенная последовательность в расположении единиц (табельные номера работников, списки избирателей, телефонные номера респондентов, номера домов и квартир и т. п.).

Для проведения механической выборки устанавливается пропорция отбора, которая определяется соотнесением объемов выборочной и генеральной совокупностей. Отбор единиц осуществляется в соответствии с установленной пропорцией через равные интервалы. Например, при пропорции 1:50 (2%-ная выборка) отбирается каждая 50-я единица, при пропорции 1:20 (5%-ная выборка) – каждая 20-я единица и т. д.

Таким образом, генеральная совокупность разбивается на равные интервалы (группы), при этом размер интервала в генеральной совокупности равен обратной величине доли выборки. Из каждой такой группы в выборку отбирается лишь одна единица.

Типическая выборка используется в тех случаях, когда все единицы генеральной совокупности можно разбить на несколько типических групп. При обследованиях населения такими группами могут быть, например, районы, социальные, возрастные или образовательные группы, при обследовании предприятий - отрасль и подотрасль, форма собственности и т. п. Типический отбор предполагает выборку единиц из каждой типической группы собственно-случайным или механическим способом.

При **серийной выборке** из генеральной совокупности отбираются не отдельные единицы, а целые их серии (группы). Внутри же каждой из попавшей в выборку серии обследуются все без исключения единицы, т. е. применяется сплошное наблюдение.

Применение серийной выборки в торговле обусловлено тем, что многие товары для их транспортировки, хранения и продажи упаковываются в пачки, коробки, ящики и т. п. Поэтому при контроле качества поступившего в упаковке товара рациональнее проверить несколько отдельных упаковок (серий), чем из всех упаковок отобрать необходимое количество единиц товара.

Отбор отдельной серии в выборочную совокупность осуществляется либо посредством собственно-случайной выборки, либо механическим отбором. Практически серийная выборка производится, как правило, по схеме бесповторного отбора.

Формулы расчета средней ошибки выборки μ для различных способов формирования выборочной совокупности

- При **повторной** выборке: $\mu=\sqrt{\frac{\sigma^2}{n}}$.
- При **бесповторной** выборке: $\mu=\sqrt{\frac{\sigma^2}{n}\cdot\left(1-\frac{n}{N}\right)}$.

✓ Если изучаемый признак количественный, то цель задачи – определение границ **генерального среднего** и $\sigma^2=\frac{\sum(x_i-\bar{x})^2\cdot f_i}{n}$.

✓ Если изучаемый признак альтернативный, то цель задачи – определение границ **генеральной доли** и $\sigma^2=w\cdot(1-w)$.

❖ Для **собственно-случайной** и **механической** выборки σ^2 – выборочная дисперсия.

❖ Для **типической** выборки (единицы разбиты на группы или типы) σ^2 – среднее арифметическое из всех групповых дисперсий (т. е. внутригрупповая дисперсия).

❖ Для **серийной** выборки (выбираются не единицы совокупности, а целые серии единиц – коробки, пачки) n – число серий в выборке, N – число серий в совокупности, σ^2 – межгрупповая дисперсия.

Каждой конкретной выборке соответствует свое среднее выборочное значение признака ($\bar{x}$). Следовательно, для генеральной совокупности $\bar{x}$ – это случайное число, которое появляется с некоторой вероятностью. Среднее значение признака для генеральной совокупности a – это постоянное число (оно неизвестно).

Наибольшее отклонение **Δ** выборочной средней (или доли) от генеральной средней (или доли), которое возможно с заданной **доверительной вероятностью γ**, называется **предельной ошибкой выборки**.

Значения доверительной вероятности γ=Ф(t) для различных значений **коэффициента доверия t** вычисляются с помощью функции Лапласа Ф(t) (функции нормального распределения) и приводятся в специальных математических таблицах. В электронных таблицах Excel можно использовать встроенную статистическую функцию НОРМСТОБР.

По заданной вероятности γ=Ф(t) коэффициент доверия можно найти следующим образом: t=НОРМСТОБР(γ/2+0,5).

Предельная ошибка выборки вычисляется по формуле $\Delta = t \cdot \mu$.

Зная выборочную среднюю величину признака ($\bar{x}$) и предельную ошибку выборки (Δ), можно определить границы (пределы), в которых заключена генеральная средняя (a): $\bar{x} - \Delta \le a \le \bar{x} + \Delta$.

Аналогично для генеральной доли (p): $w - \Delta \le p \le w + \Delta$.

Постановка задачи

Определить границы, в которых находится значение генерального признака (a или p), с заданной доверительной вероятностью γ.

Алгоритм решения

1 – нахождение выборочного признака ($\bar{x}$ или w).

2 – нахождение средней ошибки (μ).

3 – определение, какому коэффициенту доверия t соответствует заданная вероятность γ=Ф(t) (по функции Лапласа).

4 – нахождение предельной ошибки выборки ($\Delta = t \cdot \mu$).

5 – определение границ, в которых заключен показатель генеральной совокупности.

($\bar{x} - \Delta \le a \le \bar{x} + \Delta$ или $w - \Delta \le p \le w + \Delta$).

Определение оптимального объема выборки

Вторая задача выборочного наблюдения заключается в определении объема выборки (n), при котором с заданной вероятностью средняя ошибка выборки не превосходит некоторой заранее заданной величины.

Для определения необходимой численности выборки исследователь должен задать уровень точности выборочной совокупности с определенной вероятностью. Эта вероятность γ=Ф(t) позволяет с помощью таблиц найти соответствующее ей значение ***коэффициента доверия*** t, используемого в формулах.

Величина, которую не должна превосходить средняя ошибка выборки, – это уже известная нам ***предельная ошибка выборки***, обозначаемая Δ.

Необходимый объем выборки определяется по формулам:

- при повторной выборке $n = \dfrac{t^2\sigma^2}{\Delta^2}$;
- при бесповторной выборке $n = \dfrac{t^2\sigma^2 N}{\Delta^2 N + t^2\sigma^2}$.

Из формулы видно, что с увеличением предполагаемой ошибки выборки Δ значительно уменьшается необходимый объем выборки n. Так, увеличение допустимой ошибки выборки в 2 раза уменьшает необходимый ее объем в 4 раза. Необходимая численность выборки прямо пропорциональна дисперсии признака и коэффициенту доверия.

Одним из наиболее важных и в то же время сложных вопросов определения необходимого объема выборки в исследованиях является расчет показателя вариации изучаемого признака (σ). При подготовке выборочного наблюдения у его организаторов часто отсутствуют необходимые для этих вычислений данные. Основой оценки степени колеблемости изучаемого признака служат, как правило, материалы предыдущих обследований. Обращение к ним при отсутствии какой-либо другой информации вполне оправдано. Однако следует иметь в виду, что использование данных прошлых обследований имеет смысл только тогда, когда за прошедший до нового обследования период в генеральной совокупности не произошло значительных изменений.

Часто более точное представление об изучаемой совокупности может дать пробное обследование. По его данным можно рассчитать среднее квадратическое отклонение и дисперсию для последующего обоснования необходимого объема выборки.

Зная примерную величину средней, дисперсию можно найти из соотношения $\sigma \approx \dfrac{1}{3}\bar{x}$. Если известны x_{max} и x_{min}, то можно определить среднее квадратическое отклонение в соответствии с правилом «трех сигм»: $\sigma = \dfrac{1}{6}(x_{max} - x_{min})$, так как в нормальном распределении в размахе вариации «укладывается» 6σ ($\bar{x} \pm 3\sigma$).

При расчете n не следует гнаться за большими значениями t и малыми значениями Δ, так как это приведет к увеличению объема выборки, а следовательно, к увеличению затрат средств, труда и времени, вовсе не являющемуся необходимым.

3.2.2. Примеры решения задач

Пример 1

Предположим, что в результате выборочного обследования жилищных условий жителей города, осуществленного на основе собственно-случайной повторной выборки, получен следующий ряд распределения:

Результаты выборочного обследования жилищных условий жителей города

Общая (полезная) площадь жилищ, приходящаяся на 1 человека, $м^2$	До 5	5–10	10–15	15–20	20–25	25–30	30 и более
Число жителей	8	95	204	270	210	130	83

Для определения средней ошибки выборки нам необходимо, прежде всего, рассчитать выборочную среднюю величину и дисперсию изучаемого признака.

Общая (полезная) площадь жилищ, приходящаяся на 1 человека, $м^2$	Число жителей f_i	Середина интервала x_i	$x_i \cdot f_i$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 f_i$
До 5,0	8	2,5	20,0	272,4	2179,3
5,0–10,0	95	7,5	712,5	132,4	12574,7
10,0–15,0	204	12,5	2550,0	42,3	8632,3
15,0–20,0	270	17,5	4725,0	2,3	611,6
20,0–25,0	210	22,5	4725,0	12,2	2565,2
25,0–30,0	130	27,5	3575,0	72,2	9381,5
30,0 и более	83	32,5	2697,5	182,1	15115,5
Итого	1000		19005	715,9	51060,0

$$\bar{x} = \frac{19005}{1000} = 19{,}005.$$

$$\sigma^2 = \frac{51060{,}0}{1000} = 51{,}06.$$

$$\sigma = \sqrt{51{,}06} = 7{,}15.$$

Средняя ошибка выборки составит:

$$\mu = \frac{7{,}15}{\sqrt{1000}} = 0{,}23 \text{ м}^2.$$

Определим предельную ошибку выборки с вероятностью $\gamma=0{,}954$ ($t=2$):

$$\Delta = t \cdot \mu = 2 \cdot 0{,}23 = 0{,}46 \text{м}^2.$$

Установим границы генеральной средней:

$$19{,}005 - 0{,}46 \le a \le 19{,}005 + 0{,}46.$$

$$18{,}545 \le a \le 19{,}465.$$

Таким образом, на основании проведенного выборочного обследования с вероятностью 0,954 можно заключить, что средний размер общей площади, приходящейся на одного человека, в целом по городу лежит в пределах от 18,5 до 19,5 м^2.

Пример 2

В микрорайоне города проживает 5000 семей. В порядке случайной бесповторной выборки предполагается определить средний размер семьи. Какое количество семей нужно обследовать, если ошибка выборочной средней не должна превышать 0,8 человек с вероятностью 0,954; а среднее квадратическое отклонение, полученное из материалов предыдущих обследований, равно 3 человека.

Решение: при γ=0,954 по функции Лапласа t=2. Тогда необходимая численность выборки n равна:

$$n = \frac{t^2\sigma^2 N}{\Delta^2 N + t^2\sigma^2}.$$

$$n = \frac{2^2 \cdot 3^2 \cdot 5000}{0{,}8^2 \cdot 5000 + 2^2 \cdot 3^2} = \frac{4 \cdot 9 \cdot 5000}{0{,}64 \cdot 5000 + 4 \cdot 9} = \frac{180000}{3236} = 56 \text{ семей}.$$

Пример 3

Для определения средней длины детали следует провести выборочное обследование методом случайного повторного отбора. Какое количество деталей надо отобрать, чтобы ошибка выборки не превышала 3 мм с вероятностью 0,997 при среднем квадратическом отклонении 6 мм (ошибка и среднее квадратическое отклонение заданы исходя из технических нормативов).

Решение: при γ=0,997 по функции Лапласа t=3, объем выборки рассчитывается следующим образом:

$$n = \frac{t^2\sigma^2}{\Delta^2}.$$

$$n = \frac{3^2 \cdot 6^2}{3^2} = 36 \text{ деталей.}$$

Пример 4

В фермерских хозяйствах Кировской области 10000 коров. Из них в Зуевском районе – 5000, в Белохолуницком районе – 3000, в Лебяжском районе – 2000. С целью определения средней удойности предполагается провести типическую выборку коров с пропорциональным отбором внутри групп (механическим). Какое количество коров следует отобрать, чтобы с вероятностью 0,954 ошибка выборки не превышала 5 л, если на основе предыдущих обследований известно, что дисперсия типической выборки (внутригрупповая дисперсия) равна 1600?

Решение: рассчитаем необходимую численность типической выборки:

$$n = \frac{t^2\sigma^2 N}{\Delta^2 N + t^2\sigma^2}.$$

$$n = \frac{2^2 \cdot 1600 \cdot 10000}{5^2 \cdot 10000 + 2^2 \cdot 1600} = \frac{64000000}{250000 + 6400} = 249{,}6 \approx 250 \text{ коров.}$$

Необходимо отобрать 250 коров, из них:

в Зуевском районе: $n_1 = 250 \cdot \frac{5000}{10000} = 125$ коров;

в Белохолуницком районе: $n_2 = 250 \cdot \frac{3000}{10000} = 75$ коров;

в Лебяжском районе: $n_3 = 250 \cdot \frac{2000}{10000} = 50$ коров.

Пример 5

На склад завода поступило 100 ящиков готовых изделий по 80 шт. в каждом. Для установления среднего веса деталей следует провести серийную выборку деталей методом механического отбора так, чтобы с вероятностью 0,954 ошибка выборки не превышала 2 г. На основе предыдущих обследований известно, что дисперсия серийной выборки равна 4. Определить необходимый объем выборки.

Решение: $n = \frac{t^2\sigma^2 N}{\Delta^2 N + t^2\sigma^2}$.

$$n = \frac{2^2 \cdot 4 \cdot 100}{2^2 \cdot 100 + 2^2 \cdot 4} = \frac{1600}{416} \approx 4 \text{ ящика.}$$

Замечание: n и N выражаются в ящиках, а не в штуках изделий.

Пример 6

Определите, сколько семей необходимо охватить собственно-случайной выборкой для определения доли семей, не имеющих детей с вероятностью 0,954 и предельной ошибкой 2%. Известно, что в регионе проживают 600 семей, и по результатам ранее проведенных обследований доля семей, не имеющих детей, составляет 25%.

Решение: при γ=0,954 по функции Лапласа t=2.

Δ=2%=0,02; N=600; w=25%=0,25. Следовательно, дисперсия при изучении доли признака вычисляется по формуле $\sigma^2 = w \cdot (1 - w)$=0,25*0,75=0,1875.

$$n = \frac{t^2\sigma^2 N}{\Delta^2 N + t^2\sigma^2}.$$

$$n = \frac{2^2 \cdot 0{,}1875 \cdot 600}{0{,}02^2 \cdot 600 + 2^2 \cdot 0{,}1875} = \frac{450}{0{,}99} \approx 455 \text{ семей.}$$

3.2.3. Решение задач с применением EXCEL

Решите предложенные ниже задачи в электронных таблицах Excel.

Пример оформления решения:

Задача №___

Способ формирования выборки: ____________________

Вид отбора: ____________________

Цель выборки: определение границ генерального среднего

Дано:	Решение:
N=______	t=_______
n=_______	□=_______
$\bar{x}$ =_______	□=_______
σ=_______	Ответ: ______<=a<=______
γ=_______	

Задача №1

Для определения скорости расчетов с кредиторами предприятий корпорации в коммерческом банке была проведена случайная бесповторная 20%-ная выборка, состоявшаяся из 100 платежных документов, по которым средний срок перечисления и получения денег оказался равным 22 дням со средним квадратическим отклонением 6 дней. Необходимо с вероятностью 0,954 определить доверительные пределы средней продолжительности расчетов предприятий данной корпорации.

Задача № 2

Для определения зольности угля месторождения в порядке случайной повторной выборки взято 200 проб. В результате лабораторных исследований установлена средняя зольность угля в выборке, равная 17% при среднем квадратическом отклонении 3%. С вероятностью 0,866 определите пределы, в которых находится средняя зольность угля месторождения.

Задача № 3

10%-ный бесповторный типический отбор рабочих предприятия, пропорциональный размерам цехов, проведенный с целью оценки потерь из-за временной нетрудоспособности, привел к следующим результатам.

Результаты обследования рабочих предприятия

Цех	Всего рабочих, человек	Обследовано, человек	Число дней временной нетрудоспособности за год	
			средняя	дисперсия
I	1000	100	18	49
II	1400	140	12	25
III	800	80	15	16

С вероятностью 0,954 определите, в каких пределах находится среднее число дней временной нетрудоспособности одного рабочего в целом по предприятию.

Задача № 4

Для определения урожайности зерновых культур проведено выборочное обследование 100 хозяйств региона различных форм собственности, в результате которого получены сводные данные.

Распределение урожайности по хозяйствам региона, имеющим различную форму собственности

Хозяйства (по формам собственности)	Количество обследованных хозяйств	Средняя урожайность, ц/га	Дисперсия урожайности в каждой группе
Коллективные	30	18	15
Акционерные общества	50	20	25
Крестьянские (фермерские)	20	28	40
Итого	100		

Необходимо с вероятностью 0,954 определить доверительные пределы средней урожайности зерновых культур по всем хозяйствам региона.

Задача № 5

Для контроля всхожести партия семян была разбита на 25 равных по величине серий. Затем на основе случайного бесповторного отбора было проверено на всхожесть 5 серий. В результате установлено, что процент взошедших семян составляет 68. Межгрупповая дисперсия равна 400. С вероятностью 0,683 определите пределы, в которых находится доля взошедших семян всей партии.

Задача № 6

Для определения среднего возраста 1200 студентов факультета необходимо провести выборочное обследование методом случайного бесповторного отбора. Предварительно установлено, что среднее квадратическое отклонение возраста студентов равно 10 годам. Сколько студентов нужно обследовать, чтобы с вероятностью 0,954 средняя ошибка выборки не превышала 3 года?

Задача № 7

В коммерческом банке 160 персональных компьютеров 4 типов, в том числе I типа – 32, II типа – 48, III типа – 64, IV – 16. В целях изучения эффективности их использования предполагается организовать выборочное обследование на основе типической пропорциональной выборки. Отбор внутри типов ПЭВМ механический. Какое количество компьютеров каждого типа необходимо отобрать, чтобы с вероятностью 0,683 ошибка не превышала 5 единиц ПЭВМ? По материалам предыдущего обследования известно, что дисперсия типической выборки равна 729.

Задача № 8

Какое количество деталей необходимо отобрать из партии в 2000 штук, чтобы оценить качество продукции с предельной ошибкой в оценке доли деталей, не соответствующих стандарту, не превышающей 0,2% с вероятностью 0,954, если по данным предыдущей проверки доля нестандартной продукции составила 1,5%?

§ 3.3. Критерии согласия

3.3.1. Основные понятия

Понятие о случайных процессах и цепях Маркова. Критерии согласия (Пирсона, Колмогорова).

Построение нормального распределения по эмпирическим данным

Имея дело с эмпирическим распределением, можно предположить, что данному распределению соответствует определенная, характерная для него теоретическая кривая. Выдвинув гипотезу о той или иной форме распределения, стремятся описать эмпирический ряд с помощью математической модели, выражающей некоторый теоретический закон распределения. Среди различных кривых распределения особое место занимает нормальное распределение.

Нормальным N($\bar{x}$,σ) называют распределение непрерывной случайной величины x, если ее плотность распределения выражается формулой

$$f(x) = \varphi(x, \bar{x}, \sigma) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{\frac{-(x-\bar{x})^2}{2\sigma^2}}$$ – НОРМРАСП(x, $\bar{x}$, σ, ложь),

где x – значение изучаемого признака.

Или $f(x) = \varphi(t) = \frac{1}{\sqrt{2\pi}} \cdot e^{\frac{-t^2}{2}}, \quad где \quad t = \frac{x - \bar{x}}{\sigma}$.

Интегральная функция непрерывной случайной величины, распределенной нормально:

$$F(x, \bar{x}, \sigma) = \int_{-\infty}^{x} \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{\frac{-(t-\bar{x})^2}{2\sigma^2}} dt$$ – НОРМРАСП(x, $\bar{x}$, σ, истина)= НОРМСТРАСП(НОРМАЛИЗАЦИЯ(x, $\bar{x}$, σ))

Стандартное нормальное распределение N(0,1) при $\bar{x}$ =1, σ=0

Плотность (дифференциальная функция)

$$f(x) = \varphi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{\frac{-x^2}{2}}$$ – НОРМРАСП(x, 0, 1, ложь)

Интегральная функция

$$F(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}\cdot e^{\frac{-t^2}{2}}dt$$

НОРМСТРАСП(x)= НОРМРАСП(x, 0, 1, истина)

Критерии согласия

Все предположения о характере того или иного распределения – это гипотезы, а не категорические утверждения. Поэтому должны быть подвергнуты статистической проверке с помощью показателей, которые называют критериями согласия.

Для установления теоретического закона распределения нужно определить:

1) вид закона распределения (определяется из теоретических предпосылок, опыта аналогичных предшествующих исследований, или на основании графического изображения эмпирического распределения);

2) параметры распределения (их заменяют наилучшими оценками по выборке).

Но, как бы хорошо ни был подобран теоретический закон распределения, между ним и эмпирическим распределением неизбежны расхождения.

Задача: определить, являются ли эти расхождения существенными (неслучайными) и теоретический закон распределения выбран неудачно, или расхождения несущественны, зависят лишь от случайных обстоятельств, связанных с ограниченным числом наблюдений.

Для решения этой задачи используются случайные величины с известными законами распределения, называемые **критериями согласия**.

Рассмотрим три из них:

- Критерий Пирсона (χ^2)
- Критерий Романовского (C)
- Критерий Колмогорова (λ)

Алгоритм решения задачи

Проверяется гипотеза H_0: эмпирическое распределение подчиняется определенному теоретическому закону распределения.

Для этого выбирается случайная величина U (критерий согласия), характеризующая степень расхождения теоретического и эмпирического распределений. Закон распределения U при больших n известен и не зависит от эмпирического закона распределения.

1. Найти $u_{набл}$ – фактически наблюдаемое в опыте расхождение теоретических и эмпирических распределений.

2. Зная закон распределения U, найти вероятность $P(U>=u_{набл.})=\alpha$ по таблице распределения.

3. Вывод: если α мала, то H_0 отвергают, т. е. отклонения от теоретического закона распределения, большие, чем полученные в опыте, практически невозможны. Следовательно, закон распределения для X выбран не верно. Если α не мала, то H_0 можно считать правдоподобной, т. е. не противоречащей опытным данным. Следовательно, расхождение между эмпирическими и теоретическими распределениями несущественно.

χ^2 – критерий Пирсона

Пусть $x_1, x_2, ..., x_n$ – выборка из некоторой генеральной совокупности X (X – случайная величина), $F(x)$ – предполагаемая функция теоретического распределения X.

$F(x)$ – это интегральная функция распределения случайной величины X. Пусть в качестве теоретического распределения выбрано нормальное распределение со средним квадратическим отклонением σ и математическим ожиданием a.

$$F(x,a,\sigma)=\int_{-\infty}^{x}\frac{1}{\sigma\cdot\sqrt{2\pi}}\cdot e^{\frac{-(t-a)^2}{2\sigma^2}}dt$$ – НОРМРАСП(x, a, σ, истина)

H_0: F(x) – это интегральная функция распределения случайной величины X.

1) На основании выборки построим интервальный вариационный ряд $\{\Delta_i, f_i\}$ i=1,...,m, где f_i – число элементов выборки, попавших в интервал $\Delta_i = [a_i, a_{i+1})$; m – количество интервалов.

При этом f_i называют **эмпирическими частотами**,

2) Для каждого интервала $\Delta_i = [a_i, a_{i+1})$ вычислим теоретические вероятности попадания случайной величины X в этот интервал $p_i = P(a_i <= X <= a_{i+1}) = F(a_{i+1}) - F(a_i)$.

При этом $n \cdot p_i$ – называют **теоретическими частотами** (n – число элементов в выборке, $\sum_{i=1}^{m} f_i$).

3) В качестве критерия согласия берем случайную величину $U=\sum_{i=1}^{m}\frac{(f_i-n\cdot p_i)^2}{n\cdot p_i}$. Известно, что она имеет χ^2 распределение (распределение Пирсона) с числом степеней свободы k=m-r-1, где m – число интервалов вариационного ряда, r – число параметров теоретического распределения (Например, у нормального распределения два параметра: математическое ожидание и дисперсия, у распределения Пуассона один параметр – математическое ожидание).

Вычисляем $\chi^2_{набл.}=\sum_{i=1}^{m}\frac{(f_i-n\cdot p_i)^2}{n\cdot p_i}$.

4) Для заданного уровня значимости α и числа степеней свободы k=m-r-1 по специально составленной таблице χ^2-распределения найдем $\chi^2_{критич.}(\alpha,k)$. (статистическая функция Excel ХИ2ОБР(α;k)).

5) Если $\chi^2_{набл} > \chi^2_{критич.}$, то H_0 отвергается, т. е. функция распределения F(x) выбрана неверно, если $\chi^2_{набл} <= \chi^2_{критич}$, то H_0 принимается, т. е. с вероятностью (1-α) можно утверждать, что расхождения между теоретическими и эмпирическими частотами случайны.

Замечания

1. Критерий Пирсона можно применять, только если n>30.

2. В каждом интервале $\Delta_i = [a_i, a_{i+1})$ должно быть по крайней мере 5 наблюдений ($f_i >= 5$), иначе соседние интервалы объединяют и m уменьшается.

Критерий Романовского

Основан на использовании критерия Пирсона, т. е. уже найденных $\chi^2_{набл}$ и числа степеней свободы k.

$$C = \frac{|\chi^2 - k|}{\sqrt{2k}}$$

При C<3 расхождения между теоретическим и эмпирическим распределениями считаются случайными, если C>3, то неслучайными, и, следовательно, теоретическое распределение не может служить моделью для изучаемого эмпирического распределения.

Критерий Колмогорова

Дан вариационный ряд

x_i				
f_i				

1. Строятся эмпирическая функция распределения $Fn(x) = \frac{m_i}{n} = \frac{\sum_{j=1}^{i} f_j}{n}$

(накопленные частости, n – число элементов в выборке, $\sum_{i=1}^{m} f_i$) и предполагаемая теоретическая функция распределения F(x) (параметры закона распределения F(x) считаются известными).

2. Определяется мера расхождения между теоретическим и эмпирическим распределениями $\lambda_{набл} = \sqrt{n} \cdot \max|Fn(x) - F(x)|$. Известно, что эта случайная величина распределена по закону Колмогорова.

3. Для заданного уровня значимости α по таблице распределения Колмогорова найдем $\lambda_{критич}(\alpha)$

α	0,4	0,3	0,2	0,1	0,05	0,025	0,01	0,005	0,001	0,0005
λ(α)	0,89	0,97	1,07	1,22	1,36	1,48	1,63	1,73	1,95	2,03

4. Если $\lambda_{набл.} > \lambda_{критич.}$, то H_0 отвергается, закон распределения F(x) выбран не верно, если $\lambda_{набл.} <= \lambda_{критич.}$, то H_0 не противоречит опытным данным.

3.3.2. Решение задач с применением EXCEL

Задача № 1

Рассчитать теоретические частоты ряда распределения на основании эмпирических данных о росте призывников, представленных в таблице.

Группы призывников по росту, см.	Число призывников
143–146	1
146–149	2
149–152	8
152–155	26
155–158	65
158–161	120
161–164	181
164–167	201
167–170	170
170–173	120
173–176	64
176–179	28
179–182	10
182–185	3
185–188	1
Итого	1000

Указания к решению

Выдвинув гипотезу о нормальном распределении, определим по эмпирическим данным параметры этой кривой. Для этого:

1. Найти средний рост призывников $\bar{x} = a$.
2. Найти среднее квадратическое отклонение σ.
3. Определить нормированное отклонение $t_i = \dfrac{x_i - \bar{x}}{\sigma}$ для каждого варианта x_i (статистическая функция в Excel t_i=НОРМАЛИЗАЦИЯ(x_i,a, σ)).
4. По таблице нормального распределения найти значение функции $\varphi(t) = \dfrac{1}{\sqrt{2\pi}} \cdot e^{\frac{-t^2}{2}}$ – НОРМРАСП(t, 0, 1, ложь).

5. Определить теоретические частоты $f' = \dfrac{k \cdot \sum f_i}{\sigma} \cdot \varphi(t)$, где k – длина интервала (так как вариационный ряд имеет равные интервалы, то $\dfrac{k \cdot \sum f_i}{\sigma}$ – это константа).

6. Сравнить на графике эмпирические f' и теоретические f частоты.

Задача № 2

Для эмпирического распределения рабочих цеха по выработке по данным таблицы подобрать соответствующее теоретическое распределение и на уровне значимости 0,05 проверить гипотезу о согласованности двух распределений с помощью критерия «хи-квадрат».

$[x_i, x_{i+1}]$	94–100	100–106	106–112	112–118	118–124	124–130	130–136	136–142
n_i	3	7	11	20	28	19	10	2

Указания к решению

1. Построить гистограмму распределения рабочих по выработке ($[x_i, x_{i+1}]$; $w_i = n_i/n$). По виду гистограммы убедитесь, что можно предположить нормальный закон распределения признака.

2. Параметры нормального закона: математическое ожидание и дисперсия – неизвестны, поэтому их заменяют на выборочную среднюю и «исправленную» выборочную дисперсию. Так как в данной задаче число наблюдений 100 достаточно велико, то вместо «исправленной» дисперсии можно взять обычную выборочную дисперсию. Найдите выборочную среднюю (а) и выборочную дисперсию (σ^2).

3. Выдвигается гипотеза: случайная величина X – выработка рабочих цеха – распределена нормально с параметрами а и σ^2, т. е. $X \in N(a, \sigma^2)$. Для определения наблюдаемого значения критерия «хи-квадрат» удобно составить таблицу.

Интервал $[x_i, x_{i+1}]$	**Эмпирические частоты, n_i**	**Вероятности, p_i**	**Теоретические частоты, np_i**	**$(n_i - np_i)^2$**	$\dfrac{(n_i - np_i)^2}{np_i}$

Учитывая, что в рассматриваемом эмпирическом распределении частоты первого и последнего интервалов меньше 5, при использовании критерия «хи-квадрат» целесообразно объединить указанные интервалы с соседними.

Для расчета вероятностей p_i попадания случайной величины X в интервал $[x_i, x_{i+1}]$ используйте функцию Лапласа в соответствии со свойством нормального распределения:

$$p_i\left(x_i \le X \le x_{i+1}\right) = F\left(x_{i+1}\right) - F\left(x_i\right), \text{ где}$$

$$F(x, \bar{x}, \sigma) = \int_{-\infty}^{x} \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{\frac{-(t-\bar{x})^2}{2\sigma^2}} dt$$ – НОРМРАСП(x, $\bar{x}$, σ, истина)

4. Найдите критическое значение критерия «хи-квадрат» по таблицам (в Excel статистическая функция ХИ2ОБР(α, k=m-r-1), где m – новое число интервалов, после объединения, r – число параметров нормального закона распределения) и сделайте вывод о том, согласуется ли выбранный теоретический нормальный закон с опытными данными.

Задача № 3

Имеются следующие статистические данные о числе вызовов специализированных бригад скорой помощи в час в некотором населенном пункте в течение 300 часов. Подобрать соответствующее теоретическое распределение и на уровне значимости 0,05 проверить гипотезу о согласованности двух распределений с помощью критерия «хи-квадрат».

Число вызовов в час, x_i	0	1	2	3	4	5	6	7	8
Частота, n_i	15	71	75	68	39	17	10	4	1

Указания к решению

1. Построить полигон частот дискретной случайной величины X.
2. Вычислить выборочную среднюю и выборочную дисперсию.
3. Выдвигаем гипотезу: случайная величина X – число вызовов скорой помощи в час – распределена по закону Пуассона с параметром λ равным выборочной средней.

Причины выбора в качестве теоретического закона распределения Пуассона:

- вызов скорой помощи для каждого жителя – событие в целом достаточно редкое;
- по виду полигон частот дискретной случайной величины X напоминает полигон пуассоновского распределения вероятностей при небольших значениях λ;
- для распределения Пуассона характерно равенство дисперсии и среднего значения, а в пункте 2 мы получили приближенно равные выборочную среднюю и выборочную дисперсию.

4. Для определения наблюдаемого значения критерия «хи-квадрат» составим таблицу (см. задачу № 2). Вероятность значений случайной величины X найдем по формуле $p_i = P(X = x_i = m) = \frac{\bar{x}^m \cdot e^{-\bar{x}}}{m!}$ Для этого восполь-

зуйтесь статистической функцией ПУАССОН(x_i; $\bar{x}$; ложь). При вычислении наблюдаемого значения критерия «хи-квадрат» объедините интервалы с частотой, меньшей 5, с соседними интервалами.

5. Найдите критическое значение критерия «хи-квадрат» по таблицам (в Excel статистическая функция ХИ2ОБР(α,k=m-r-1), где m – новое число интервалов, после объединения, r – число параметров распределения Пуассона (r=1)) и сделайте вывод о том, согласуется ли выбранный теоретический закон с опытными данными.

Задача № 4

По данным задачи № 2 с помощью критерия Колмогорова на уровне значимости 0,05 проверить гипотезу о том, что случайная величина X – выработка рабочих предприятия – имеет нормальный закон распределения с параметрами a=119,2; σ^2=87,48, т. е. N(119,2; 87,48).

Указания к решению:

1. В качестве вариант возьмем среднее значение в каждом интервале данного в задаче интервального вариационного ряда.

2. Значение эмпирической функции распределения $F_n(x)$ вычислим как накопленные частости $\dfrac{m_i}{n} = \dfrac{\sum_{j=1}^{i} n_j}{n}$.

3. Для построения теоретической функции распределения для нормального закона с параметрами a=119,2; σ^2=87,48 F(x) воспользуйтесь встроенной статистической функцией НОРМРАСП(xi; 119,2; 9,35; истина).

4. Для вычислений заполните таблицу.

x_i	n_i	m_i	$F_n(x)$	F(x)	$\lvert F_n(x)-F(x)\rvert$

Для определения наблюдаемого значения критерия Колмогорова вычислите значение $\lambda = \sqrt{n} \cdot \max\lvert F_n(x) - F(x)\rvert$.

5. Найдите критическое значение критерия Колмогорова по таблицам для уровня значимости $\alpha = 0{,}05$ и сделайте вывод о том, согласуется ли выбранный теоретический закон с опытными данными.

Задача № 5

По данным задачи № 1 с помощью критериев

a) Пирсона,

b) Романовского,

c) Колмогорова

на уровне значимости 0,05 проверить гипотезу о распределении призывников по росту по нормальному закону распределения.

§ 3.4. Дисперсионный анализ

3.4.1. Основные понятия

Дисперсионный анализ – это статистический метод, предназначенный для оценки влияния различных факторов на результат эксперимента. Результатом эксперимента является некоторая случайная величина X, называемая **результативным признаком**. На значение случайной величины X влияет **фактор** F (или несколько факторов), состоящий из нескольких **уровней** (групп) F_i (i=1,2,…,p).

Например, директора фирмы интересует зависимость выполненных за смену работ от работающей в эту смену бригады (всего работают, сменяя друг друга, 5 бригад). Объем выполненных работ – это результативный признак X. На него влияет фактор F – работающая в эту смену бригада. F_i – i-й уровень фактора – это номер бригады (i=1,2,…,p, p=5). Все значения случайной величины X разбиваются на p групп, каждая i-я группа содержит n_i значений случайной величины X. Чаще всего на практике для любого i=1,2,…,p количество наблюдений (число значений случайной величины X) одинаково n_i=q. Результаты наблюдений заносятся в таблицу:

№ испытания	№ бригады				
	1	2	3	4	5
1					
2					
…					
q					

Идея дисперсионного анализа

Основная дисперсия раскладывается в сумму составляющих ее дисперсий, каждое слагаемое которой соответствует действию определенного фактора, влияющего на результативный признак.

Например, в двухфакторном дисперсионном анализе получим разложение вида

$\sigma_C^2 = \sigma_A^2 + \sigma_B^2 + \sigma_{AB}^2 + \sigma_Z^2$, где

σ_C^2 – общая дисперсия изучаемого признака C;

σ_A^2 – доля дисперсии, вызванная влиянием фактора A;

σ_B^2 – доля дисперсии, вызванная влиянием фактора B;

σ_{AB}^2 – доля дисперсии, вызванная взаимодействием факторов A и B;

σ_Z^2 – доля дисперсии, вызванная неучтенными случайными причинами (случайная дисперсия).

В дисперсионном анализе рассматривается гипотеза H_0: ни один из рассматриваемых факторов ***не оказывает влияния*** на изменчивость результирующего признака.

Однофакторный дисперсионный анализ

Пусть на количественный нормально распределенный признак X воздействует фактор F, который имеет p постоянных уровней. Число наблюдений на каждом уровне одинаково и равно q. Следовательно, всего наблюдалось $n=p*q$ значений x_{ij} признака X, где i – номер испытания ($i=1,2,\ldots,q$), j – номер уровня фактора ($j=1,2,\ldots,p$). Результаты наблюдений приведены в таблице.

№ испытания (i)	уровни фактора F (j)			
	F_1	F_2	...	F_p
1	x_{11}	x_{12}	...	x_{1p}
2	x_{21}	x_{22}	...	x_{2p}
...	...	...	...	...
q	x_{q1}	x_{q2}	...	x_{qp}
групповая средняя	$\bar{x}_{гр\,1}$	$\bar{x}_{гр.2}$	...	$\bar{x}_{гр\,p}$

Модель однофакторного дисперсионного анализа

$$x_{ij}=a+m_j+\varepsilon_{ij}$$

a – генеральное среднее всех возможных результатов наблюдения, т. е. M(X). Его оценка – общая выборочная средняя $\bar{x}=\frac{\sum_{i=1}^{q}\sum_{j=1}^{p}x_{ij}}{p\cdot q}$

(1)

$a_{гр\,j}$ – генеральное среднее всех возможных результатов наблюдения j-ой группы. Его оценка – групповая выборочная средняя $\bar{x}_{гр\,j}=\frac{\sum_{i=1}^{q}x_{ij}}{q}$

(2)

$m_j=a_{гр.j}-a$ – эффект влияния на x_{ij}, вызванный j-м уровнем фактора F. Его оценка – $\bar{x}_{гр\,j}-\bar{x}$.

ε_{ij} – влияние на x_{ij} всех других неконтролируемых факторов.

Применение дисперсионного анализа предполагает, что

1) ε_{ij} взаимно независимы для любых i и j;
2) $\varepsilon_{ij}\in N(0,\sigma)$, где $\sigma=const$.

По основной идее дисперсионного анализа представим общую дисперсию как сумму дисперсии, вызванной влиянием фактора F, и дисперсии, вызванной случайными причинами.

Известно, что оценкой дисперсии является сумма квадратов отклонений наблюдаемых значений от среднего значения. Рассмотрим эти три дисперсии.

1. Общая сумма квадратов отклонений наблюдаемых значений от общей средней $\bar{x}$. Характеризует влияние и фактора F и случайных причин.

$$S_{общ} = \sum_{i=1}^{q}\sum_{j=1}^{p}\left(x_{ij} - \bar{x}\right)^2$$

2. Факторная сумма квадратов отклонений групповых средних от общей средней. Характеризует рассеяние между группами, а следовательно, воздействие фактора F.

$$S_{факт.} = q \cdot \sum_{j=1}^{p}\left(\bar{x}_{гр.j} - \bar{x}\right)^2$$

3. Остаточная сумма квадратов отклонений наблюдаемых значений группы от своей групповой средней. Характеризует рассеяние внутри группы, т. е. отражает влияние случайных причин.

$$S_{ост.} = \sum_{i=1}^{q}\sum_{j=1}^{p}\left(x_{ij} - \bar{x}_{гр\,j}\right)^2$$

По основной идее дисперсионного анализа

$$S_{общ.} = S_{факт.} + S_{ост.}$$

Если поделить обе части равенства на число наблюдений n, то получим известную теорему о сумме дисперсий, где $S_{факт.}$ превратится в межгрупповую дисперсию, а $S_{ост.}$ – во внутригрупповую дисперсию.

Элементарными преобразованиями можно получить формулы, более удобные для расчетов:

$$S_{общ.} = \sum_{j=1}^{p} P_j - \frac{1}{p \cdot q}\left(\sum_{j=1}^{p} R_j\right)^2, \quad S_{факт.} = \frac{1}{q}\sum_{j=1}^{p} R_j^2 - \frac{1}{p \cdot q}\left(\sum_{j=1}^{p} R_j\right)^2,$$

$$S_{ост.} = S_{общ.} + S_{факт.}, \text{ где } P_j = \sum_{i=1}^{q} x_{ij}^2, \quad R_j = \sum_{i=1}^{q} x_{ij}$$

Разделив суммы квадратов отклонений на соответствующее число степеней свободы, получим общую, факторную и остаточную дисперсии.

Число степеней свободы определяется как общее число наблюдений минус число связывающих их уравнений. Поэтому для среднего квадрата $S_{факт.}^2$, являющегося несмещенной оценкой межгрупповой дисперсии, число степеней свободы $k1=p-1$, так как при его расчете используются p групповых средних, связанных между собой одним уравнением (1). А для среднего квадрата $S_{ост.}^2$, являющегося несмещенной оценкой внутригрупповой дисперсии, число степеней свободы $k2=pq-p$, так как при ее расчете используются все $pq=n$ наблюдений, связанных между собой p уравнениями (2).

Для среднего квадрата $S_{общ}^2$, являющегося несмещенной оценкой общей дисперсии, число степеней свободы $k=pq-1=n-1$, так как при ее расчете используются все $pq=n$ наблюдений, связанных между собой одним уравнением (1).

$$S^2_{общ} = \frac{S_{общ}}{pq-1} \qquad S^2_{факт.} = \frac{S_{факт.}}{p-1} \qquad S^2_{ост.} = \frac{S_{ост.}}{p(q-1)}$$

Задача однофакторного дисперсионного анализа

Проверим гипотезу H_0: фактор F ***не влияет*** на значение результирующего признака X на уровне значимости α.

Переформулируем гипотезу H_0. Если F не влияет на X, то средние значения в разных группах $\bar{x}_{гр\, j}$ равны (значения признака не зависят от уровня фактора, на котором они были получены).

H_0: $\bar{x}_{гр.1} = \bar{x}_{гр.2} = \ldots = \bar{x}_{гр.p}$

Еще раз переформулируем гипотезу H_0. При равенстве групповых средних ($\bar{x}_{гр.1} = \bar{x}_{гр.2} = \ldots = \bar{x}_{гр.p}$) дисперсии $S^2_{факт.}$ и $S^2_{ост.}$ являются несмещенными оценками генеральной дисперсии σ^2 случайной величины X.

Следовательно, проверка нулевой гипотезы H_0 свелась к проверке равенства дисперсий $S^2_{факт}$ и $S^2_{ост.}$ Получаем H_0: $S^2_{факт} = S^2_{ост.}$

Для проверки нулевой гипотезы используем критерий $F = \frac{S^2_{факт}}{S^2_{ост.}}$, имеющий распределение Фишера с числом степеней свободы $k1=p-1$, $k2=p(q-1)$. Если $F_{набл.} > F_{крит.}$, нулевая гипотеза отвергается, т. е. фактор F ***влияет*** на признак X.

Замечание: если $S^2_{факт.} < S^2_{ост.}$, то уже отсюда следует справедливость нулевой гипотезы и нет надобности прибегать к критерию F.

Если установлено, что фактор F влияет на результирующий признак X, то можно измерить степень этого влияния. Для этого используют **выборочный коэффициент детерминации**.

$$d = \frac{S_{факт.}}{S_{общ}}$$

Он показывает, какую долю выборочной дисперсии составляет дисперсия групповых средних, т. е. какая доля общей дисперсии объясняется зависимостью результативного признака X от фактора F.

Дисперсионный анализ применяется также для установления **однородности нескольких совокупностей**. Дисперсии этих совокупностей по предположению одинаковы. Если дисперсионный анализ покажет, что и средние значения равны (выполняется гипотеза H_0: $S^2_{факт} = S^2_{ост.}$), то совокупности однородны. Однородные совокупности можно объединить в одну и тем самым получить о ней более полную информацию и более надежные выводы.

3.4.2. Решение задач с использованием EXCEL

Задача 1

Проверить статистическую существенность влияния катализатора А на химическую реакцию, результаты измерений при 5 уровнях фактора А приведены в таблице.

A1	A2	A3	A4	A5
3,2	2,6	2,9	3,7	3
3,1	3,1	2,6	3,4	3,4
3,1	2,7	3	3,2	3,2
2,8	2,9	3,1	3,3	3,5
3,3	2,7	3	3,5	2,9
3	2,8	2,8	3,3	3,1

Указания к решению

1. Проверим гипотезу H_0: катализатор А не влияет на химическую реакцию на уровне значимости $\alpha = 0{,}05$. ($H_0: \quad S^2_{факт} = S^2_{ост}$).

2. Для проверки гипотезы о равенстве дисперсий воспользуемся критерием F-распределения Фишера – Снедекора. Вычислим $F_{набл} = \dfrac{S^2_{факт}}{S^2_{ост}}$ и $F_{кр}(\alpha;k1=p-1;k2=p(q-1))$, где число испытаний q=6, число уровней изучаемого фактора p=5.

3. Для вычисления Sфакт. и Sост. по формулам $S_{факт} = q \cdot \sum\limits_{j=1}^{p}\left(\bar{x}_{гр.j} - \bar{x}\right)^2$ и $S_{ост} = \sum\limits_{j=1}^{p}\sum\limits_{i=1}^{q}\left(x_{ij} - \bar{x}_{гр.j}\right)^2$ постройте расчетные таблицы в EXCEL.

$$S^2_{факт} = \frac{S_{факт}}{p-1} \quad S^2_{ост} = \frac{S_{ост}}{p(q-1)}$$

4. Сравнив Fнабл и Fкр, ответьте на вопрос задачи.

Задача 2

В таблице приведены данные по объемам работ, выполненных на стройке за смену для четырех бригад. Проверить гипотезу дисперсионного анализа о равенстве средних. Оценить степень зависимости объема ежедневной выработки от работающей бригады. Найти оценки параметров модели дисперсионного анализа.

Номер бригады	Объем выполненной работы				Групповое среднее	Выборочная дисперсия
1	140	144	142	145	142,75	3,69
2	150	149	152	152	150,25	1,19
3	148	149	146	147	147,50	1,25
4	150	155	154	152	152,75	3,69

Задача 3

Решите задачу 1, используя инструмент пакета анализа электронных таблиц EXCEL «Однофакторный дисперсионный анализ».

Для активизации Пакета анализа необходимо выполнить команду Сервис – Надстройки и выбрать Пакет анализа, после этого в меню Сервис появится строка Анализ данных.

Сравните результаты решения с результатами решения задачи 1.

Задача 4

У 60 рабочих фиксировались среднечасовая выработка в натуральных единицах продукции. Данные обследования отражены в таблице. Оценить существенность влияния возраста и стажа на производительность труда.

Стаж	Возраст		
	от 25 до 35 лет	от 35 до 45 лет	от 45 до 55 лет
От 1 до 4 лет	19	19	18
	20	20	19
	20	20	20
	20	23	21
	22	25	23
От 4 до 7 лет	30	20	19
	31	29	25
	32	30	25
	32	31	26
	34	31	26
От 7 до 10 лет	35	36	24
	35	40	24
	39	41	24
	40	42	25
	41	45	25
Свыше 10 лет	40	28	20
	40	31	24
	41	35	25
	41	36	31
	42	40	32

Решите задачу, используя инструмент пакета анализа электронных таблиц EXCEL «Двухфакторный дисперсионный анализ с повторениями».

Дисперсионный анализ с повторениями позволяет оценить существенность влияния фактора А (стажа), В (возраста), и их взаимодействия (факторов А и В) на среднечасовую выработку продукции в натуральных единицах.

Рассмотрите решение задачи, используя таблицу 4x3, в каждой ячейке которой содержится среднее значение наблюдений. К таблице примените инструмент пакета анализа электронных таблиц EXCEL «Двухфакторный дисперсионный анализ без повторений». Дисперсионный анализ без повторений позволяет оценить существенность влияния фактора А (стажа) и В (возраста) на результирующий фактор без учета взаимодействия факторов А и В.

§ 3.5. Корреляционно-регрессионный анализ

3.5.1. Основные понятия и пример решения задач

Корреляция – это статистическая зависимость между случайными величинами, не имеющими строго функционального характера, при которой изменение одной из случайных величин приводит к изменению математического ожидания другой.

Основной задачей корреляционного анализа является выявление тесноты связи между переменными X и Y и количественная оценка тесноты этой связи.

Корреляционный анализ следует применять только в том случае, когда данные наблюдений или эксперимента можно считать случайными и выбранными из нормальной совокупности.

В корреляционном анализе экспериментальные данные задаются как несгруппированные данные, т. е. набор пар чисел (x_i, y_i), $i = \overline{1,n}$, где $\{x_1, \ldots, x_n\}$ – выборка значений переменной X, $\{y_1, \ldots, y_n\}$ – выборка значений переменной Y. Однако очень часто экспериментальные данные задаются корреляционной таблицей. Вариант такого задания рассмотрен на следующем примере.

Пример. Для исследования зависимости годового объема производства Y от основных фондов X получены статистические данные по 20 предприятиям.

Таблица 1

y_i \ x_i		12,5	17,5	22,5	27,5	n_j
20–21	20,5	1	–	–	–	1
21–22	21,5	–	2	–	–	2
22–23	22,5	–	1	2	–	3
23–24	23,5	–	–	3	3	6
24–25	24,5	–	–	–	8	8
n_i		1	3	5	11	n = 20

В первой строке таблицы записаны значения переменной X, в первом столбце – интервалы изменения переменной Y, во втором – середины этих интервалов. Для дальнейших расчетов необходимы только середины интервалов, поэтому сами интервалы могут отсутствовать. Центральную часть таблицы занимают частоты n_{ij} (число предприятий), соответствующие значениям переменных $X = x_i$ и $Y = y_j$. В последней строке записаны частоты

$$n_i = \sum_{j=1}^{m} n_{ij},$$

а в последнем столбце – частоты

$$n_j = \sum_{i=1}^{l} n_{ij},$$

здесь $l = 4$ – число значений величины X, $m = 5$ – число значений величины Y, $n = \sum n_i = \sum n_j = 20$ – число всех значений.

Таблица такого вида называется **корреляционной таблицей.**

Основной оценкой для тесноты связи между переменными X и Y служит **выборочный коэффициент корреляции r**, который определяется формулой

$$r = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{S_x S_y},$$

где $\overline{x}$ и $\overline{y}$ – средние арифметические значения признаков X и Y соответственно, S_x и S_y – дисперсии переменных X и Y соответственно.

К основным **свойствам коэффициента корреляции** необходимо отнести следующие:

1) Коэффициент корреляции принимает значения на отрезке [-1,1].

В зависимости от того, насколько $|r|$ приближается к 1, различают слабую, умеренную и сильную связь, т. е. чем ближе $|r|$ к 1, тем теснее связь.

2) Если переменные X и Y умножить на одно и то же число, то коэффициент корреляции не изменится.

3) При положительной, или прямой, связи, когда с увеличением значений одного признака возрастают значения другого, коэффициент корреляции приобретает положительный (+) знак и находится в пределах от 0 до +1.

4) При отрицательной, или обратной, связи, когда с увеличением значений одного признака соответственно уменьшаются значения другого, коэффициент корреляции сопровождается отрицательным (–) знаком и находится в пределах от 0 до –1.

5) Чем сильнее связь между признаками, тем ближе величина коэффициента корреляции к 1. Если r = ±1, то корреляционная связь переходит в функциональную линейную зависимость, т.е. каждому значению признака X

будет соответствовать одно строго определенное значение признака Y. (Y=a+bX)

Запишем более подробные формулы для вычисления коэффициента корреляции.

$$r=\frac{n\sum_{i=1}^{l}\sum_{j=1}^{m}x_iy_jn_{ij}-\left(\sum_{i=1}^{l}x_in_i\right)\left(\sum_{j=1}^{m}y_jn_j\right)}{\sqrt{n\sum_{i=1}^{l}x_i^2n_i-\left(\sum_{i=1}^{l}x_in_i\right)^2}\sqrt{n\sum_{j=1}^{m}y_j^2n_j-\left(\sum_{j=1}^{m}y_jn_j\right)^2}}$$

$$r=\frac{\sum_{i=1}^{l}\sum_{j=1}^{m}(x_i-\bar{x})(y_j-\bar{y})n_{ij}}{nS_xS_y}.$$

Пример. Для данных таблицы 1 найти выборочный коэффициент корреляции.

Для вычислений составим таблицу (таблица 2).

Находим суммы

$$\sum_i x_in_i=480,\ \sum_j y_jn_j=468;$$

$$\sum_i x_i^2n_i=11925,\ \sum_j y_j^2n_j=10979$$

и заносим их в таблицу. Вычислим

$$\sum_i\sum_j x_iy_jn_{ij}=20.5\cdot12.5+21.5\cdot17.5\cdot2+22.5\cdot17.5+22.5\cdot22.5\cdot2+23.5\cdot22.5\cdot3+$$

$$+23.5\cdot27.5\cdot3+24.5\cdot27.5\cdot8=11330.$$

Таблица 2

j,i	x_i	n_i	y_j	n_j	x_in_i	y_jn_j	x_i^2	$x_i^2n_i$	y_j^2	$y_j^2n_j$
1	12,5	1	20,5	1	12,5	20,5	156,25	156,25	420,25	420,25
2	17,5	3	21,5	2	52,5	43,0	306,25	918,75	462,25	924,5
3	22,5	5	22,5	3	112,5	67,5	506,25	2531,25	506,25	1518,75
4	27,5	11	23,5	6	302,5	141,0	756,25	8318,75	552,25	3313,5
5			24,5	8		196,0			600,25	4802
$\sum$					480	486		11925		10979

Найдем выборочный коэффициент корреляции

$$r=\frac{20\cdot11330-480\cdot468}{\sqrt{20\cdot11925-480^2}\sqrt{20\cdot10979-468^2}}=\frac{1960}{90\cdot23.58}=0.924.$$

В **регрессионном анализе** изучаются модели вида

$$Y = \varphi(X) + \varepsilon,$$

где X – неслучайная независимая переменная, называемая фактором; Y – случайная зависимая переменная (результирующий признак); ε – случайная переменная, характеризующая отклонение от линии регрессии (остаточная переменная).

Уравнение регрессии записывается в виде

$$y_x = \varphi\left(x, b_0, b_1, \ldots, b_p\right),$$

где x – значения величины X; b_0, b_1, ..., b_p – параметры функции регрессии.

Таким образом, **задача регрессионного анализа** состоит в определении функции φ, ее параметров b_0, b_1, ..., b_p и дальнейшего статистического исследования уравнения регрессии.

Если функция φ линейна по x, т. е.

$$y_x = a + bx,$$

то говорят, что имеет место линейная регрессия Y по X.

Параметры a и b определяются по методу наименьших квадратов следующим образом:

$$a = \frac{\sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}.$$

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n\sum_{i=}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}.$$

Учитывая, что

$$\sum_{i=1}^{n} x_i = n\overline{x}, \quad \sum_{i=1}^{n} y_i = n\overline{y}, \quad \sum_{i=1}^{n} x_i^2 = n\overline{x_i^2}, \quad \sum_{i=1}^{n} x_i y_i = n\overline{xy},$$

получим

$$a = \frac{\overline{y} \cdot \overline{x^2} - \overline{x} \cdot \overline{xy}}{\overline{x^2} - \left(\overline{x}\right)^2}, \qquad b = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \left(\overline{x}\right)^2}.$$

Пример. Для зависимости Y от X, заданной корреляционной табл. 1, найти параметры a и b уравнения линейной регрессии $y=a+bx$.

$$\overline{x} = \frac{1}{n}\sum_{i} x_i n_i = \frac{1}{20} \cdot 480 = 24; \quad \overline{y} = 23.4;$$

$$\overline{x^2}=\frac{1}{n}\sum_i x_i^2 n_i=\frac{1}{20}\cdot 11925=596.25;\ \overline{y^2}=\frac{1}{n}\sum_j y_j^2 n_j=\frac{1}{20}\cdot 10979=548.95;$$

$$\overline{xy}=\frac{1}{n}\sum_i\sum_j x_i y_j n_{ij}=\frac{1}{20}\cdot 11330=566{,}5\cdot$$

Уравнение регрессии будет иметь вид

$$y_x=17.59+0.24x,$$

$$a=17{,}59;\ b=0{,}24.$$

3.5.2. Решение задач с использованием EXCEL

Задача 1

Для исследования зависимости годового объема производства Y от основных фондов X получены статистические данные по 20 предприятиям. Для данных таблицы найти выборочный коэффициент корреляции. Для зависимости Y от X, заданной корреляционной таблицей, найти оценки параметров a и b уравнения линейной регрессии $y=a+bx$.

x_i \ y_j		**12,5**	**17,5**	**22,5**	**27,5**	n_j
20–21	20,5	1	–	–	–	1
21–22	21,5	–	2	–	–	2
22–23	22,5	–	1	2	–	3
23–24	23,5	–	–	3	3	6
24–25	24,5	–	–	–	8	8
n_i		1	3	5	11	n=20

Указания к решению

Вычислить выборочный коэффициент корреляции, используя:

1) расчетную таблицу в Excel

i, j	x_i	n_i	y_i	n_j	$x_i n_i$	$y_j n_j$	x_i^2	$x_i^2 n_i$	y_i^2	$y_i^2 n_j$

2) встроенную статистическую функцию КОРРЕЛ();

3) инструмент Анализа данных Корреляция (Сервис – Анализ данных – Корреляция).

Сравнить полученные результаты.

Для вычисления коэффициентов a и b по методу наименьших квадратов

1) воспользуйтесь данными, полученными в расчетной таблице, подставив их в формулу;

2) используйте инструмент Пакета анализа Регрессия в электронных таблицах Excel. Для этого в диалоговом окне «Регрессия»:

- в качестве входного интервала Y выберите столбец значений Y_i (i=1..20);
- в качестве входного интервала X выберите столбцы значений X_i (i=1..20);
- выберите уровень надежности 95%;
- остатки отобразите в виде графика подбора.

Сравните результаты решения.

Задача 2

Определить тесноту связи между себестоимостью продукции Y (тыс. руб.) и количеством выпускаемой продукции X (тыс. штук) по данным 7 предприятий.

X	2	3	4	5	6	7	8
Y	2	1,9	2,2	2,4	2,3	2,5	2,5

Найти линейное уравнение регрессии.

Решить задачу, используя расчетную таблицу и встроенный Пакет Анализа. Сравнить полученные результаты.

Задача 3

В таблице содержатся данные о распаде 10 г. радиоактивного вещества, где *t* – время (в месяцах), *X* – количество (г) оставшегося вещества в момент времени *t*. Показать, что процесс распада подчиняется экспоненциальному закону, найти его параметры.

t	**1**	**2**	**3**	**4**	**5**	**6**
X	8,453	7,666	5,047	3,628	3,464	2,434
t	**7**	**8**	**9**	**10**	**11**	**12**
X	2,905	1,167	2,142	2,028	2,992	0,715

Указания к решению

1. Предположим, что функция регрессии имеет вид $x=ae^{bt}$. Приведите уравнение к линейному виду $y=a'+bt$, прологарифмировав его. Для вычисления параметров линейного уравнения регрессии a' и b составьте расчетную таблицу в Excel

i	t_i	x_i	y_i	t_i^2	y_i^2	y_it_i

По значению a' вычислите параметр a и запишите экспоненциальное уравнение регрессии.

2. Решить задачу, используя инструмент Пакета анализа Регрессия в электронных таблицах Excel.

§ 3.6. Кластерный анализ

3.6.1. Основные понятия

Многомерный статистический анализ (МСА)

Многомерный статистический анализ – раздел математической статистики, посвященный методам сбора, систематизации, обработки и интерпретации сложных совокупностей данных, нацеленный на выявление неявных (латентных) закономерностей в структуре и тенденциях развития исследуемых многомерных процессов.

Формы представления исходных статистических данных

1) ***Матрица объект-свойство***

$$X_{n\times p}=\begin{bmatrix} X_{11}(t)\ X_{12}(t)\dots X_{1p}(t) \\ X_{21}(t)\ X_{22}(t)\dots X_{2p}(t) \\ \dots\quad \dots\quad \dots \\ X_{n1}(t)\ X_{n2}(t)\dots X_{np}(t) \end{bmatrix},\quad t=t_1,\dots,t_N \qquad (1)$$

$X_{ij}(t)$ – значение j-ого признака, характеризующего состояние i-ого объекта в момент времени t. (число объектов n, число признаков – p)

2) ***Матрица парных сравнений***

Состоит из характеристик попарных сравнений объектов по некоторому свойству (число объектов n).

$$\Gamma_{n\times n}=\begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1n} \\ \dots & \dots & \dots \\ \gamma_{n1} & \cdots & \gamma_{nn} \end{bmatrix} \qquad (2)$$

Задачи многомерного статистического анализа

1) Статистическое исследование зависимостей – выявление и описание множественных статистических ***связей***, существующих между множеством признаков. ***Используемые методы***: корреляционный, регрессионный анализ, анализ временных рядов.

2) Классификация объектов и признаков – разбиение совокупности объектов, представленной в виде матриц (1) или (2) на сравнительно небольшое число (оно может быть и неизвестно заранее) однородных групп или классов. Исходными данными при классификации объектов являются строки матрицы (1) или (2), при классификации признаков – столбцы матрицы (1) или (2). ***Методы***: дискриминантный анализ, кластерный анализ.

3) Снижение размерности признакового пространства – переход от исходного набора признаков к вспомогательному набору меньшего числа признаков в задачах отбора наиболее информативных показателей, сжатия больших массивов информации, визуализации многомерных данных. ***Методы***: факторный анализ, компонентный анализ (метод главных компонент), многомерное шкалирование.

Основные этапы многомерного статистического анализа

1) Предварительный анализ исследуемой реальной системы:

a) Определение цели и задач исследования.

b) Выбор объектов и признаков.

c) Выбор формы для сбора информации.

d) Оценка необходимого времени и трудозатрат на проведение исследования.

2) Составление детального плана сбора исходной статистической информации.

3) Сбор и контроль исходных статистических данных и их преобразование в электронную форму.

4) Первичная статистическая обработка данных:

a) Отображение вербальных признаков в номинальной или порядковой шкале.

b) Статистическое описание исходных совокупностей.

c) Анализ выбросов.

d) Восстановление пропущенных наблюдений.

e) Проверка однородности выборки.

f) Проверка статистической независимости последовательности наблюдений, составляющих выборку.

g) Экспериментальный анализ закона распределения исследуемой генеральной совокупности.

5) Уточнение методов анализа, используемых для моделирования исследуемой проблемы. Составление детального плана вычислительного анализа информации.

6) Вычислительная реализация основной части статистической обработки данных.

7) Подведение итогов исследования, интерпретация результатов, выводы.

Некоторые из этапов могут объединяться или исключаться.

Многомерный статистический анализ обобщает большое число методов, которые подразделяются:

- по виду признаков (количественные, неколичественные),
- по виду шкал измерения,
- по числу зависимых переменных,
- по виду структуры зависимости.

Шкала – система чисел или иных элементов, принятых для оценки каких-либо величин.

Виды шкал

Номинальные (классификационные)

Объекты невозможно измерить количественно, поэтому им присваиваются метки. ***Например***, пол, профессия, регион проживания. Признак с двумя значениями называется **дихотомическим (бинарным)**, если число значений больше двух – **категориальным.**

Порядковые (ранговые)

Объекты можно обозначить порядковыми числительными (их называют рангами), которые можно подвергать возведению в степень, извлечению корня. Можно сравнить объекты по принципу «больше–меньше», но без указания, насколько больше. ***Например***, школьные оценки, оценки качества продукции.

Количественные (метрические)

Например, объем продукции определенного вида в соответствующих единицах измерения (тонны, рубли).

С оценками можно выполнять операции сложения, умножения, деления.

- ***интервальные***

Обладают всеми качествами порядковой шкалы и точно определяют величину интервала между точками на шкале в принятых единицах измерения.

- ***пропорциональные***

Кроме возможностей интервальных шкал имеют фиксированную нулевую точку отсчета. Поэтому позволяют выяснить, насколько или во сколько раз один объект меньше другого.

Кластерный анализ
(классификация без обучения)

В зависимости от наличия и характера априорной информации о природе искомых классов и от конечных целей многомерной классификации при отсутствии обучающих выборок используют один из трех **подходов**:

1) ***методы расщепления смесей вероятностных распределений*** (каждый класс интерпретируется как параметрически заданная одномодальная генеральная совокупность при неизвестном значении определяющего ее параметра, а классифицируемые наблюдения – как выборки из смеси таких генеральных совокупностей);

2) ***методы кластерного анализа;***

3) ***классификационные процедуры иерархического типа*** (главная цель – получение наглядного представления о структуре всей классифицируемой совокупности).

Непараметрический случай классификации без обучения
(кластерный анализ)

Постановка задачи

Пусть мы не располагаем обучающими выборками и отсутствует информация о характере распределения наблюдений X_i $(i=1, ..., n)$ внутри каждого из классов.

Необходимо всю совокупность объектов разбить на сравнительно небольшое число (заранее известное или нет) однородных классов.

Исходные данные X_i $(i=1, \ldots, n)$ – это точки в p-мерном признаковом пространстве (p-мерные вектора).

Геометрическая близость точек означает, что они находятся на сравнительно небольших расстояниях друг от друга. Полученные в результате разбиения классы называют **кластерами** (cluster) – это группа элементов, обладающих каким-то общим свойством. Метод нахождения кластеров называется **кластерным анализом**.

Однородность объектов определяется либо заданием правила вычисления расстояния между объектами, либо заданием некоторой меры степени близости (сходства).

Виды расстояний между объектами

- **Евклидово расстояние**

$$d_E(X_i, X_j) = \sqrt{\sum_{k=1}^{p} (x_i^k - x_j^k)^2}$$

Используется, если

✓ наблюдения извлекаются из генеральной совокупности, которая описывается многомерным нормальным законом распределения,

✓ X взаимно независимы и имеют одинаковую дисперсию,

✓ компоненты наблюдений $1, 2, \ldots, p$ одинаково важны для классификации,

✓ признаковое пространство совпадает с геометрическим (т. е. p может принимать значения 1, 2 или 3).

- **Взвешенное евклидово расстояние**

$$d_{BE}(X_i, X_j) = \sqrt{\sum_{k=1}^{p} w_k \cdot (x_i^k - x_j^k)^2}$$

Используется, если

✓ компоненты наблюдений 1, 2, ..., p неоднородны и степень их важности задают веса w_k, причем $0< w_k <1$, $k=1,\ldots, p$.

Определение весов требует дополнительного исследования (экспертные опросы, специальные модели), определение их по исходным данным не дает желаемого эффекта.

- **Расстояние Минковского**

$$d_M(X_i, X_j) = \left(\sum_{k=1}^{p} | x_i^k - x_j^k |^g \right)^{1/g}$$

Частные случаи:

✓ g=1 ***Хеммингово расстояние*** (манхеттенское, или расстояние городских кварталов)

$d_H(X_i, X_j) = \sum_{k=1}^{p} | x_i^k - x_j^k |$ используется, если объекты характеризуются дихотомическими признаками (имеют ровно два значения 0,1)

✓ g=2 ***евклидово расстояние***

Расстояние между классами объектов

Пусть S_i – i-я группа (класс) ($i=1, .., k$) k – число классов;
n_i – число объектов в i-й группе;
$\overline{X}(i)$ – среднее арифметическое наблюдений в i-й группе («центр тяжести» группы)
$\rho(S_l, S_m)$ – расстояние между группами S_l и S_m;
$d(X_i, X_j)$ – расстояние между объектами X_i и X_j.

Виды расстояний

- Расстояние по принципу **«ближайшего соседа»**

$$\rho_{\min}(S_l, S_m) = \min_{X_i \in S_l, X_j \in S_m} d(X_i, X_j)$$

- Расстояние по принципу **«дальнего соседа»**

$$\rho_{\max}(S_l, S_m) = \max_{X_i \in S_l, X_j \in S_m} d(X_i, X_j)$$

- Расстояние по **«центрам тяжести»** групп

$$\rho(S_l, S_m) = d(\overline{X}(l), \overline{X}(m))$$

- Расстояние по принципу **«средней связи»**

$$\rho_{cp}(S_l, S_m) = \frac{1}{n_l n_m} \sum_{X_i \in S_l} \sum_{X_j \in S_m} d(X_i, X_j)$$

Основные типы задач кластер-анализа и основные типы кластер-процедур

В зависимости от количества наблюдений n выделяют ***2 типа задач кластер-анализа***:

1) несколько десятков наблюдений,

2) несколько сотен или тысяч наблюдений.

В зависимости от априорной информации об окончательном числе классов выделяют ***3 типа задач кластер-анализа***:

a) число классов известно априори,

b) число классов неизвестно и подлежит оценке,

c) число классов неизвестно, но его определение не входит в условие задачи. Требуется построить иерархическое дерево исследуемой совокупности (**дендрограмму**).

Выделяют ***3 основных типа кластер-процедур***:

1. **Иерархические** процедуры (делятся на **агломеративные** и **дивизимные**). Предназначены для решения задач типа (1с и 2с).

2. Параллельные процедуры, предназначены для решения задач типов (1a) и (1b). Реализуются с помощью итерационных алгоритмов, на каждом шаге которых одновременно используются все имеющиеся наблюдения.

3. Последовательные процедуры предназначены для решения задач типов (2a) и (2b). Реализуются с помощью итерационных алгоритмов, на каждом шаге которых используется лишь небольшая часть исходных наблюдений и результат разбиения на предыдущем шаге.

Иерархические кластер-процедуры

Принцип работы **агломеративных (дивизимных)** процедур состоит в последовательном объединении (разделении) групп элементов сначала самых близких (далеких), а затем все более отдаленных друг от друга (приближенных друг к другу). В агломеративных процедурах сначала каждый элемент образует отдельный класс, в дивизимных – все элементы образуют один класс.

Результаты иерархических процедур:

- структура исследуемого множества наблюдений,
- наглядная интерпретация результатов.

Недостатки иерархических процедур:

- громоздкость вычислительной реализации,
- результаты разбиения могут быть весьма далеки от оптимальных, так как объекты распределяются по кластерам за один проход.

Пример. Агломеративный иерархический алгоритм «ближайшего соседа» («одиночной связи»)

Использует расстояние между классами по правилу «ближайшего соседа». На первом шаге алгоритма каждое наблюдение рассматривается как отдельный кластер. Строится матрица расстояний (например, евклидовых) между классами. Матрица квадратная, размер равен числу наблюдений. Далее ищем минимум в матрице расстояний и объединяем два самых близких кластера в один. Пересчитывается матрица расстояний, размерность которой снижается на единицу. Работа алгоритма заканчивается, когда все наблюдения объединены в один класс.

Агломеративные иерархические алгоритмы «средней связи» и «полной связи» («дальнего соседа») отличаются от этого алгоритма лишь способами вычисления расстояний между классами.

Последовательные кластер-процедуры

Когда число n классифицируемых наблюдений велико, иерархические процедуры крайне трудоемки, поэтому используют алгоритмы, на каждом шаге которых обсчитывается лишь небольшая часть исходных наблюдений, например, одно из них, как в **методе k-средних** (Дж. Мак-Куин, 1967).

Пусть имеется n наблюдений, характеризующимися p признаками. Наблюдения необходимо разбить на заданное число классов k<=n.

Сначала из n точек (объектов) отбираются случайно или задаются исследователем k точек, которые принимаются за эталоны (центры кластеров).

Каждому эталону присваивается порядковый номер, который и является номером кластера.

Смысл алгоритма k-средних в последовательном уточнении эталонных точек с соответствующим пересчетом приписываемых им весов.

На первом шаге из оставшихся (n-k) не эталонных объектов берется точка X_i и проверяется, к какому из эталонов она находится ближе всего (например, с помощью евклидова расстояния). Проверяемый объект X_i присоединяется к тому эталону, до которого расстояние минимально. Эталон заменяется новым, определяемым как центр тяжести старого эталона и присоединенной точки X_i, вес эталона увеличивается на единицу, все другие эталоны и веса не меняются.

На следующем шаге выбираем следующую точку X_i и повторяем ту же процедуру.

Через (n-k) шагов все объекты окажутся отнесенными к одному из k кластеров.

Затем, для того чтобы добиться устойчивости разбиения, все точки, кроме новых эталонных, опять по тому же алгоритму присоединяют к полученным кластерам. Новое разбиение сравнивают с предыдущим. Если они совпадают, то работа алгоритма заканчивается, в противном случае цикл повторяется.

3.6.2. Решение задач с использованием EXCEL

Задача 1

Необходимо провести классификацию 5 фирм, каждая из которых характеризуется тремя переменными: x_1 – среднегодовая величина оборотных средств, млн руб., x_2 – материальные затраты на 1 руб. произведенной продукции, коп., x_3 – объем произведенной продукции, млн руб. Проведите классификацию при помощи иерархического агломеративного метода.

№ фирмы	x_1	x_2	x_3
1	130	90	165
2	80	75	90
3	140	80	100
4	70	76	80
5	60	66	110

<u>Указания к решению:</u>

✓ Нормируйте исходные данные вычитанием среднего (СРЗНАЧ()) и делением на среднеквадратическое отклонение (СТАНДОТКЛОНП() или КОРЕНЬ(ДИСПР())). Можно использовать встроенную функцию НОРМАЛИЗАЦИЯ().

✓ Постройте матрицу расстояний (матрица квадратная, симметричная относительно главной диагонали, состоящей из нулей, имеет размерность

5x5), воспользовавшись евклидовым расстоянием (КОРЕНЬ(СУММКВРАЗН())).

✓ На каждом очередном шаге объединяйте два самых близких (расстояние в матрице минимальное) объекта в один кластер. При пересчете значений матрицы расстояний (размерность матрицы на каждом шаге уменьшается на один) используйте принцип «дальнего соседа» (МАКС()).

✓ Постройте дендрограмму кластеризации пяти объектов.

Задача 2

Классифицируйте объекты A, B, C и D, заданные двумя переменными, на две группы **методом k-средних**.

объекты	**A**	**B**	**C**	**D**
x_1	5	-1	1	-3
x_2	3	1	-2	-2

Указания к решению:

1. Разбейте объекты на два кластера, например (AB) и (CD). Вычислите координаты центров (СРЗНАЧ()) этих кластеров (($\overline{X}_1, \overline{X}_2$) для каждого кластера).

2. Вычислите евклидово расстояние (КОРЕНЬ(СУММКВРАЗН())) от каждого объекта A, B, C и D до центра каждого кластера (AB) и (CD).

3. Если какой-то из объектов A, B, C, D имеет расстояние до центра «чужого» кластера меньшее, чем расстояние до центра своего кластера, то этот объект перемещается в другой кластер, получаем два новых кластера.

4. Пересчитайте координаты центров новых кластеров. Повторяйте выполнение шагов, начиная со 2-го для новых кластеров до тех пор, пока ни один из объектов не будет перемещен в другой кластер. Итерационный процесс закончится, когда классификация совпадет с классификацией, полученной на предыдущем шаге.

Библиографический список

Основная литература

1. Вентцель, Е. С. Задачи и упражнения по теории вероятностей [Текст] / Е. С. Вентцель, Л. А. Овчаров. – М.: Высш. шк., 2000. – 366 с.

2. Вентцель, Е. С. Теория случайных процессов и ее инженерные приложения [Текст] / Е. С. Вентцель, Л. А. Овчаров. – М.: Высш. шк., 2000.

3. Вентцель, Е. С. Теория вероятностей [Текст] / Е. С. Вентцель. – М.: Высш. шк., 2002.

4. Волков, И. К. Случайные процессы [Текст] / И. К. Волков. – М.: Изд-во МГТУ им. Н. Э.Баумана, 2000.

5. Гмурман, В. Е. Теория вероятностей и математическая статистика [Текст] / В. Е. Гмурман. – М.: Высш. шк., 2000. – 479 с.

6. Гмурман, В. Е. Руководство к решению задач по теории вероятностей и математической статистике [Текст] / В. Е. Гмурман. – М.: Высш. шк., 1999. – 400 с.

7. Теория вероятностей [Текст] / под ред. В. С. Зарубина, А. П. Крищенко. – М.: Изд-во МГТУ им. Н. Э.Баумана, 1999.

Дополнительная литература

1. Андрухаев, Х. М. Сборник задач по теории вероятностей [Текст] / Х. М, Андрухаев. – М.: Просвещение, 1985.

2. Гнеденко, Б. В. Курс теории вероятностей [Текст] / Б. В. Гнеденко. – М.: Наука, 1988.

3. Колмогоров, А. Н. Основные понятия теории вероятностей [Текст] / А. Н, Колмогоров. – М.: Наука, 1974.

4. Мостеллер, Ф. Пятьдесят занимательных вероятностных задач с решениями [Текст] / Ф. Мостеллер. – М.: Наука, 1985.

5. Феллер, В. Введение в теорию вероятностей и ее приложения [Текст] / В. Феллер. – М.: Мир, 1984.

ДЛЯ ЗАМЕТОК

Учебное издание

Ашихмина Татьяна Викторовна

**ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА
В ПРИМЕРАХ И ЗАДАЧАХ**

Редактор *Т. Н. Котельникова*

Подписано в печать 06.05.2010 г.
Формат 60x84/16.
Бумага офсетная.
Усл. печ. л. 6,75.
Тираж 100 экз.
Заказ № 1401.

Издательство Вятского государственного гуманитарного университета,
610002, г. Киров, ул. Красноармейская, 26

Издательский центр Вятского государственного гуманитарного университета,
610002, г. Киров, ул. Ленина, 111, тел. (8332) 673-674

ISBN 978-5-93825-840-2