МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
УО «Белорусский государственный экономический университет»

М.П. Дымков

ВЫСШАЯ МАТЕМАТИКА

Третий семестр

Курс лекций
для студентов экономических
специальностей вузов

Минск 2014

Пространство элементарных событий. Операции над случайными событиями

Первичными понятиями в ТВ являются:
случайное событие и вероятность.

Случайным событием (*возможным событием* или просто *событием*) называется любой факт, который в результате испытания (эксперимента) может произойти или не произойти.

Под экспериментом понимается выполнение некоторого комплекса условий G, в которых наблюдается то или иное явление или факт.

Примеры случайного эксперимента: бросание монеты, бросание игральной кости, проведение лотереи, азартные игры, стрельба по цели, поступление звонков на телефонную станцию и т.п.

Событие—это не какое-то происшествие, а всего лишь *возможный исход*, *результат испытания* (опыта, эксперимента)

Различные результаты эксперимента называют **исходами**.

Определение 1. Множество всех взаимоисключающих исходов эксперимента называется **пространством элементарных событий**. Взаимоисключающие исходы – это те, которые не могут наступить одновременно.

Пространство элементарных событий будем обозначать буквой Ω, а его исходы – буквой ω.

Примеры

1)В опыте с бросанием монеты случайные события – это выпадение «герба» (А) или «цифры» $(B) \Rightarrow \Omega = \{A, B\}$. Эти два исхода в рамках данного опыта уже нельзя разбить на более мелкие составляющие, т.е. они в некотором роде являются «элементарными» (разумеется, здесь исключаются всякие гипотезы о падении на ребро и т.п.).
2) Выпадение на игральной кости: одного очка ω_1,….., выпадение шести очков ω_6. Это элементарные события и их уже *нельзя разбить* на более мелкие. $\Rightarrow \Omega = \{\omega_1, ..., \omega_6\}$.

Более сложный пример получим, если рассмотрим падение идеальной (т.е. не имеющей размера) частицы на плоскость. Тогда результат испытания представляет собой попадание частицы в определенную точку на плоскости, и этот результат можно отождествить с двумерным вектором в некоторой системе координат. В этом случае множество всевозможных исходов при таком испытании не является конечным— оно бесконечно (имеет мощность континуума). Еще пример такого рода — время безотказной работы лампочки.

Итак, событие – это результат испытания. Испытание – это осуществление определенного комплекса условий.

На практике интересуют события ***неэлементарные.***

Определение2. Произвольное подмножество из пространства элементарных событий называется **событием**.

Событие может состоять из одного или нескольких элементарных событий, а также состоять из счетного или несчетного числа элементарных событий. События обозначают обычно буквами: A, B, C…

Например, при бросании игральной кости: выпадение четного числа очков – это событие A происходит тогда и только тогда, когда появляется одно из элементарных событий $\omega_2, \omega_4, \omega_6.$, т.е. A=$\{\omega_2, \omega_4, \omega_6.\}$

Среди множества событий особо выделяют следующие:

Достоверное событие (его также обозначают буквой Ω) – это событие, которое происходит всегда в данном опыте.

Например, в урне имеются только белые шары. Тогда извлечение из урны белого шара – достоверное событие.

Событие Ω, состоящее из всех исходов эксперимента, очевидно является **достоверным событием**. Оно обязательно происходит, так как эксперимент всегда заканчивается каким-нибудь исходом.

Невозможное событие (его обозначают как $\varnothing$)– событие, которое никогда не произойдет в данном опыте.

Например, в предыдущем примере вытянуть черный шар (а в урне только белые шары !) – невозможно.

Пустое множество исходов эксперимента является **невозможным** событием

Для ТВ достоверные и невозможные события являются мало интересными. ТВ ищет закономерности в случайных массовых событиях.

Случайные события в зависимости от того, как они могут появиться в испытаниях, классифицируются на:

единственно возможные,
равновозможные,
несовместные.

События называются ***единственно возможными,*** если при испытании появление одного и только одного из них есть событие достоверное.

Случайные события называют ***равновозможными***, если наступление одного из них не является более возможным, чем другое.

Два события А и В называют ***несовместными,*** если наступление одного из них исключает наступление другого (выпадение «герба» исключает «цифру»). Если же появление одного из них не исключает возможность появления другого, то такие события называются совместными (два стрелка – в одну цель – попадание первого не исключает попадание второго).

Два события называются если появление одного из них ***противоположными,***равносильно не появлению другого («герб» – «цифра»).

Если A – событие, то $\overline{A}$ – обозначает противоположное событие.

Например, если A – герб $\Rightarrow$ $\overline{A}$ – цифра.

Определение 2. **Полной группой** событий называется совокупность событий, которые:
1)попарно несовместны;
2)и в результате данного опыта происходит одно и только одно из них.

(Будем полную группу событий обозначать также буквой Ω (в этом имеется определенный смысл).

По другому, полную группу событий можно определить как совокупность всех единственно возможных событий данного испытания.

Например, пусть в урне находятся красные, белые и черные шары. Полную группу событий по извлечению шаров составляют следующие три события: извлечен A – красный, B – черный, C – белый шар .

Например, при проведении опроса группы студентов по поводу, с какой буквы начинается фамилия, то полную группу событий составляют буквы алфавита (за исключением известных типа «ь» и т.д.) , т.е. $\Omega = \{A, Б, B, ..., Я\}$.

Если же нас интересует пол студентов, то $\Omega = \{M, F\}$, где M – мужской пол, F – женский.

Часто бывает полезным наглядное представление событий в виде так называемых **диаграмм Венна**. Будем изображать множество всех событий (или говорят еще — исходов испытания) Ω в виде прямоугольника. Тогда каждый элементарный исход испытания (или элементарное событие) ω соответствует точке внутри прямоугольника, а каждое событие (неэлементарное) A можно отождествить с некоторой областью (подмножеством) этого прямоугольника (см. рис. 1).

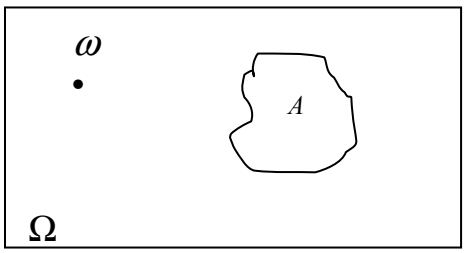


Рис. 1

Алгебра событий

Определение 3. **Суммой** двух событий A и B (обозначается $A+B$ или $A\bigcup B$) называется событие, состоящее из всех исходов, входящих либо в A, либо в B. Другими словами, под $A+B$ понимают следующее событие: произошло или событие A, или событие B, либо они произошли одновременно, т.е. произошло хотя бы одно из событий A или B.

Примеры.

1) Бросают игральную кость.

Пусть событие A – выпадет 1 или 3;

событие B – 3 или 5.

Тогда сумма $A+B=\{$выпадение нечетных чисел$\}$.

2) A – съели первое блюдо; B – съели второе блюдо; C – третье блюдо; $E = A + B + C -$ перекусили что-то.

Определение 4. **Произведением** двух событий A и B (обозначается AB *или* $A \cap B$) называется событие, состоящее из тех исходов, которые входят как в A, так и в B. Иными словами, AB означает событие, при котором события A и B наступают одновременно.

Примеры.

1) Предположим, что эксперимент (опыт) заключается в подбрасывании двух монет. Пусть событие B – выпадение их обеих одной стороной; A – выпадение хотя бы одного «герба».

Тогда $A \cap B = \{$выпадение двух «гербов»$\}$.

2) Рассмотрим предыдущий пример 2 (пример с обедом).

Тогда $D = ABC = \{$плотно «пообедали»$\}$.

3) A – он пришел; B – она пришла $\Rightarrow$ $A \cap B = \{$свидание$\}$.

Заметим, что для указания операции « + » и « · » используются также логические выражения , а именно « и » для обозначения операции умножения « · » и « или » для операции суммы « + » . Когда говоря «A и B», то это речь идет о произведение событий AB, когда говорят «A или B», то это равносильно A + B.

Заметим также, что если A и B несовместные события, то $A \cap B = \varnothing$.

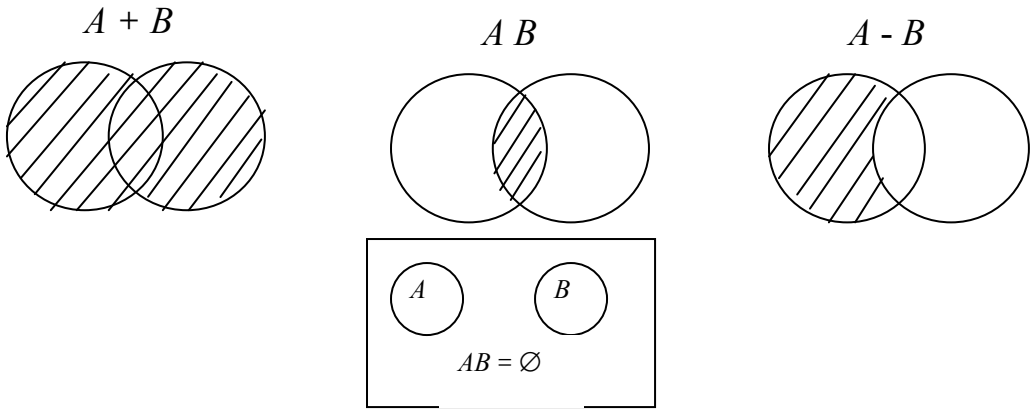
Определение 5. **Разностью** двух событий A и B (обозначается $A-B$ или $A \setminus B$) называется событие, состоящее из исходов, входящих в A, но не в B.

Смысл события $A-B$ состоит в том, что событие A наступает, но при этом не наступает событие B.

Пример.

Бросаем 2 монеты: A = {выпадение хотя бы одного «герба»}; B = {падение монет одной стороной}.
Тогда $A \setminus B$={выпадение ровно одного «герба»}.

Если события изобразить на плоскости, то результат определенных операций над событиями выглядит так:



Определение6. **Противоположным (дополнительным)** для события A (обозначается $\overline{A}$) называется событие, состоящее из всех исходов, которые не входят в A. Наступление события $\overline{A}$ означает просто, что событие A не наступило.

Определение 8. Говорят, что событие A **содержится** в событии B (обозначается $A \subset B$), если все исходы события A входят в событие B.

Пример 1. Два шахматиста играют подряд две партии. Под исходом опыта будем понимать выигрыш одного из них в i-ой партии или ничью. **Задача:** Построить пространство Ω элементарных исходов.

Решение. Обозначим события: A_i – в i-ой партии выиграл 1-ый игрок, B_i – 2-ой игрок, C_i – ничья.

Тогда возможные исходы игры:

- обе партии выиграл первый игрок – $A_1 \cdot A_2$;
- обе партии выиграл второй игрок – $B_1 \cdot B_2$;
- обе партии закончились вничью – $C_1 \cdot C_2$;
- в первой партии выиграл первый игрок, во второй выиграл второй игрок – $A_1 B_2$;
- в первой выигрыш 1-го игрока, во второй – ничья $A_1 C_2$;
- в первой партии победа игрока 2, во второй – первого $B_1 A_2$;
- в первой – победа второго игрока, во второй – ничья $B_1 C_2$;
- в первой – ничья, во второй – победа первого игрока – $C_1 A_2$;
- в первой – ничья, во второй – победа второго игрока – $C_1 B_2$.

Ответ:

$\Omega = \{ A_1 \cdot A_2,\ B_1 \cdot B_2,\ C_1 \cdot C_2,\ A_1 B_2,\ A_1 C_2,\ B_1 A_2,\ B_1 C_2,\ C_1 A_2,\ C_1 B_2 \}$.

Пример 2. Пусть A, B, C – три произвольных события. ***Найти выражения*** для событий, состоящих в том, что:

1) произошло только A;

Решение. Обозначим $\overline{B}$ и $\overline{C}$, что события B и C не произошли. Тогда событие: ***произошло только A*** можно записать в виде: $A \cdot \overline{B} \cdot \overline{C}$.

2) произошло A и B, но C не произошло;

$$A \cdot B \cdot \overline{C}$$

3) все три события произошли;

$$A \cdot B \cdot C$$

4) произошло, по крайней мере, одно из событий;

$$A + B + C.$$

5) произошло, по крайней мере, два события;

$$AB + AC + BC.$$

6) произошло одно и только одно событие;

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

7) произошло два и только два события;

$AB\overline{C} + A\overline{B}C + \overline{A}BC$, или $AB + AC + BC - ABC$.

8) ни одно событие не произошло;

$$\overline{A} \cdot \overline{B} \cdot \overline{C}$$

9) произошло не более двух событий:

$\overline{ABC}$, т.е. три события одновременно *не произошли*.

Свойства операций над событиями:

1) $A+B=A+B$; 2) $AB=BA$; 3) $A+\overline{A}=\Omega$;

4) $A\cdot\Omega=A$; 5) $AB\subset A$; 6) $A\overline{A}=\text{ø}$;

7) $\overline{\overline{A}}=A$; 8) $A-B=A\overline{B}$; 9) $(A+B)\cdot C=AC+BC$;

10) $\overline{A+B}=\overline{A}\cdot\overline{B}$; 11) $\overline{AB}=\overline{A}+\overline{B}$;

12) $(A-B)\cdot C=AC-BC$.

Замечания.

1. Если ***сложное*** событие записано в виде нескольких действий над различными элементарными событиями, то сначала выполняются дополнения, затем умножение, и, наконец, сложение и вычитание (сравните с арифметикой чисел). Так, например, сложное событие, задаваемое формулой $C=A_1\overline{A_2}B_1\cup A_3\overline{B_2}\setminus B_3$ эквивалентно событию, которое записывается по формуле $C=\left\{\left[A_1\left(\overline{A_2}\right)B_1\right]\cup\left[A_3\left(\overline{B_2}\right)\right]\right\}\setminus B_3$.

2. Все действия над событиями можно получить с помощью только двух действий – объединения и дополнения (или, пересечения и дополнения). Это утверждение основано на формулах де Моргана:

$$A\cap B=\overline{\overline{A}\cup\overline{B}},\quad A\cup B=\overline{\overline{A}\cap\overline{B}},\quad A\setminus B=A\overline{B},$$

которые можно доказать с помощью диаграмм Венна.

Упр.*1. Доказать, что $\overline{A\cup B}=\overline{A}\cap\overline{B}$.

Упр.• 2.

Пусть $A_i\,(i=1,2,3)$ – событие – i-ый посетитель магазина сделал покупку.

Составить следующие события:

а) второй посетитель ушел без покупки $(\overline{A_2})$;

б) только второй посетитель купил $(\overline{A_1}, \overline{A_2}, \overline{A_3})$;

в) двое ушли без покупки $(\overline{A_1}\,\overline{A_2} + \overline{A_2}\,\overline{A_3} + \overline{A_1}\,\overline{A_3})$;

г) кто-то купил $A_1 + A_2 + A_3$;

д) только один купил $(A_1\overline{A_2}\,\overline{A_3} +)$.

Классическое определение вероятности.

Еще не вводя никаких определений, а исходя лишь из повседневного опыта, каждый из нас согласится, что различные события можно сравнивать по степени их возможного появления. Например, попадание в цель из близкого расстояния более возможно, чем с дальнего; если в лотерее разыгрывается 1 авто и 100 мотоциклов, то выиграть мотоцикл – более возможное событие, чем авто. Вероятность события – это некоторая численная мера степени объективной возможности появления события. Таким образом, вероятность события сопоставляет каждому событию A некоторое действительное число $p(A)$, (т.е. $A \to p(A)$), которое тем больше, чем больше возможностей для появления этого события. Вероятность события А обозначается p (A) (‘probability’).

Классическое определение вероятности. Вероятностью $p(A)$ появления события А называется число, равное отношению числа m случаев, благоприятствующих появлению события А к общем числу n всех (единственно возможных, равновозможных и несовместных) исходов испытания:

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

Говорят, что случай (исход) благоприятствует событии А, если появление этого случая влечет за собой наступление данного события А.

Примеры.

1. В урне 3 шара – красный, белый, черный. Найти вероятность события А, состоящего в том, что мы извлечем белый шар (с одной попытки!)

$$\Rightarrow m=1,\ n=3;\ p(A)=\frac{1}{3}.$$

2. Игральная кость: А – выпадение четного числа

$$\Rightarrow m=3,\ n=6;\ p=\frac{1}{2}.$$

3. В партии из 200 деталей имеется 4 бракованные. Пусть событие А = {наугад взятая деталь – бракованная}. Имеем

$$m=4,\ n=200;\ p(A)=\frac{4}{200}=\frac{1}{50}=0,02.$$

Непосредственно из определения вытекают следующие

Основные свойства вероятности.

1. $0\le p(A)\le 1$.
2. $p(\Omega)=1$ (т.к. m = n).
3. p (∅) = 0 (т.к. m = 0).

Формула (1) – называется классическим определением вероятности. Это определение предполагает, что число n всевозможных исходов конечно, а исходы – равновозможные.

В действительности, равновозможность исходов наблюдается редко. Поэтому вместо (1) иногда используется ***статистическая вероятность*** (относительная частота, частость событий).

Пусть при проведении N испытаний некоторое событие A появилось m раз. Отношение $\frac{m}{N}$ называется частотой события A или частостью события

$$\frac{m}{n} = \omega(A). \qquad (2)$$

В ряде случаев эксперименты показывают, что при большом $N \to \infty$ это отношение остается примерно постоянным, что и позволяет принять его за определение статистической вероятности:

$$p^*(A) = \frac{m}{N}.$$

Геометрическая вероятность.

Уже упоминали, что множество всех случайных событий в некоторых экспериментах может заполнять собой некоторую ограниченную область Ω конечномерного (пусть k – мерного) пространства. (Например, падение идеальной частицы на стол представляет некое подмножество из R^2). Тогда любое событие A представляет собой некоторую подобласть этого множества Ω, т.е. $A \subset \Omega$.

Тогда вероятность (геометрическая) наступления события A есть $\boxed{p(A) = \frac{\mu(A)}{\mu(\Omega)},}$

где $\mu(A)$ – мера множеств A (длина отрезка, если опыт проводится в одномерном пространстве, площадь или объем – в зависимости от размерности пространства).

Пример. Два человека обедают в столовой, которая открыта с 12 до 13 часов. Каждый из них приходит в произвольный момент времени и обедает в течение 10 минут. Какова вероятность их встречи?

Решение. Пусть x — время прихода первого в столовую, а y — время прихода второго $(12 \le x \le 13; 12 \le y \le 13)$.

Можно установить взаимно-однозначное соответствие между всеми парами чисел $(x;y)$ (или множеством исходов) и множеством точек квадрата со стороной, равной 1, на координатной плоскости, где начало координат соответствует числу 12 по оси X и по оси Y, как изображено на рисунке 6. Здесь, например, точка A соответствует исходу, заключающемуся в том, что первый пришел в 12.30, а второй - в 13.00. В этом случае, очевидно, встреча не состоялась.

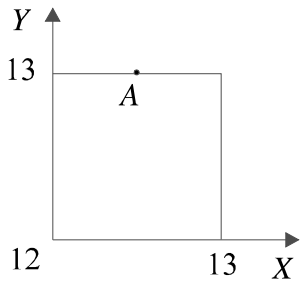
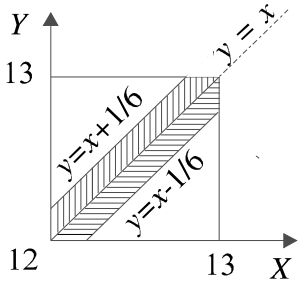


Рис.6

Если первый пришел не позже второго ($y \ge x$), то встреча произойдет при условии $0 \le y - x \le 1/6$ (10 минут – это 1/6 часа).



Если второй пришел не позже первого ($x \ge y$), то встреча произойдет при условии $0 \le x - y \le 1/6$..

Между множеством исходов, благоприятствующих встрече, и множеством точек области σ, изображенной на рисунке 7

в заштрихованном виде, можно установить взаимно-однозначное соответствие.

Искомая вероятность равна отношению заштрихованной площади области к площади всего квадрата. Площадь квадрата равна единице, а площадь заштрихованной области можно определить как разность единицы и суммарной площади двух треугольников, изображенных на рисунке. Отсюда следует:

$$p = 1 - \frac{25}{36} = \frac{11}{36}$$

Комбинаторные формулы

Нахождение вероятности по формуле $p(A)=\frac{m}{n}$ называют еще ***непосредственным подсчетом вероятности.***

Ясно, что эта формула требует умения искать числа m и n. Для их поиска в случае нетривиальных реализаций случайных событий часто используются элементы комбинаторики.

Комбинаторика – это раздел математики, в котором изучаются вопросы о том, сколько различных подмножеств, обладающих заданными свойствами, можно выбрать из данного множества.

Ниже приведем основные формулы комбинаторики и понятия такие как: перестановки, размещения и сочетания.

Основные правила комбинаторики.

1. **Правило суммы.**

Если два действия взаимно исключают друг друга, причем одно из них можно выполнить m способами, а другое – n способами, то выполнить **одно** (не важно какое) из этих действий можно $N=n+m$ способами.

Пример.

В коробке лежат 9 красных, 7 зеленых и 12 желтых карандашей. Выбор одного карандаша (любого цвета) можно сделать 28 способами ($N=9+7+12=28$)

2. **Правило умножения.**

Пусть требуется выполнить одно за другим какие-то k действий. Если первое действие можно выполнить n_1 способами, после этого второе действие можно

осуществить n_2 способами и т.д. и, наконец, после осуществления $(k-1)$-го действия, k-ое можно выполнить n_k способами, то все k действий вместе могут быть выполнены $N = n_1 \cdot n_2 \cdot n_3 \cdot \ldots \cdot n_k$ способами.

Эти правила дают удобные универсальные методы решения многих комбинаторных задач.

Примеры.

1) В коробке лежат 9 красных, 7 зеленых и 12 желтых карандашей. Сколько можно осуществить наборов из трех карандашей в заданном порядке «красный-зеленый-желтый» ?

Число наборов совпадает с числом выборов одного красного карандаша из 9, другого зеленого из 7 и третьего желтого из 12. Следовательно таких наборов будет $N = 9 \cdot 7 \cdot 12 = 756$)

2) Три человека *независимо друг от друга* решили поместить свои вклады в банк. Банков всего 8.

Тогда общее число способов выбора для всех троих равно $N = 8 \cdot 8 \cdot 8 = 8^3$

Пусть имеется множество A_n, состоящее из n различных элементов.

Множество элементов, для которого установлен порядок расположения элементов, называется **упорядоченным**.

Пример. Пусть множество А состоит из двух чисел 1 и 2. Тогда {1, 2} и {2, 1} – есть ***два различных упорядоченных множества.***

Определение. ***Перестановкой*** из n элементов называется любой упорядоченный набор этих элементов.

Таким образом, перестановки отличаются друг от друга только порядком элементов.

Примеры перестановок:

1) Множество A состоит из 1, 2 и 3. Тогда всевозможные перестановки есть:
{1, 2, 3}, {2, 3, 1}, {3, 1, 2}, {1, 3, 2}, {2, 1, 3}, {3, 2, 1}.

2) Распределение n различных должностей среди n человек;

3) Расположение n различных предметов в одном ряду.

Число перестановок обозначается P_n (читается "P из n").

Чтобы вывести формулу числа перестановок, представим себе n ячеек, пронумерованных числами 1,2,...n. Все перестановки будем образовывать, располагая элементы U_n в этих ячейках. В первую ячейку можно занести любой из n элементов (иначе: первую ячейку можно заполнить n различными способами). Заполнив первую ячейку, можно найти n–1 вариантов заполнения второй ячейки. Таким образом, существует $n(n–1)$ вариантов заполнения двух первых ячеек. При заполнении первых двух ячеек можно найти n–2 варианта заполнения третьей ячейки, откуда получается, что три ячейки можно заполнить $n(n\text{-}1)(n\text{-}2)$ способами. Продолжая этот

процесс, получим, что число способов заполнения n ячеек равно $n(n-1)(n-2)...3\cdot 2\cdot 1$. Отсюда

$$P_n = n(n-1)(n-2)...\cdot 3\cdot 2\cdot 1 = n! \quad (1!=1;\ 0!=1)$$

Произведение всех чисел от 1 до n обозначают n! и называют "n" факториал.

$1!=1,\ 2!=1\cdot 2,\ 5!=1\cdot 2\cdot 3\cdot 4\cdot 5.\ 0!=1$ (по определению).

Пример. Сколько существует вариантов замещения 5-ти различных вакантных должностей 5-ю кандидатами?

$$5!=5\cdot 4\cdot 3\cdot 2=120\ .$$

Размещениями из n элементов по k элементов будем называть упорядоченные подмножества, состоящие из k элементов множества A_n

Пример. Выпишем все размещения из трех чисел 1, 2, 3 по два: {1, 2},{1, 3}, {2, 1}, { 3, 1}, {2, 3}, {3, 2}.

Число размещений из n элементов по k элементов обозначается A_n^k (читается "A из n по k").

***Одно размещение из n элементов по k элементов может отличаться от другого* как набором *элементов*, так и порядком *их расположения*.**

Можно видеть, что размещение из n элементов по n является перестановкой из n элементов.

В задачах о размещениях полагается $k < n$. Число размещений A_n^m подсчитывается по схеме точно так же как и число перестановок: на первом месте может находиться любой из n элементов, на втором – любой оставшихся (n – 1) элементов и т.д., и, наконец, на m-ом месте может находиться любой из оставшихся (n – k + 1) элементов. Снова воспользуемся основной формулой комбинаторики (только в данном случае имеем m групп размеров $n,\ n-1,\ \ldots,\ n-k+1$). Имеем

$$A_n^m = n(n-1)\ldots(n-k+1) = \frac{n(n-1)\ldots 2\cdot 1}{(n-k)(n-k-1)\ldots 2\cdot 1} = \frac{n!}{(n-k)!}.$$

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

Примеры.

1)Сколько существует различных вариантов выбора 4-х кандидатур из 9-ти специалистов для поездки в 4 различные страны?

$$A_9^4 = 9\cdot 8\cdot 7\cdot 6 = \frac{9!}{(9-4)!} = \frac{9!}{5!} = 3024$$

2) Научное общество состоит из 25 человек. Надо выбрать президента общества, вице-президента, ученого секретаря и казначея. Сколькими способами может быть сделан этот выбор, если каждый член общества может занимать лишь один пост?

В этом случае надо найти число размещений (без повторений) из 25 элементов по 4, так как здесь играет роль и то кто будет выбран в руководство общества и то, какие посты займут выбранные

$$A_{25}^4 = 25\cdot 24\cdot 23\cdot 22 = 303600.$$

Заметим, что способ выбора, приводящий к перестановкам и размещениям называют ***еще выборкой без возвращений.***

Если при выборе k элементов из n, элементы возвращаются обратно и упорядочиваются, то говорят, что это размещение с повторениями. Число размещений с повторениями равно $\overline{A}_n^k = n^k$

Пример. Для запирания сейфов и автоматических камер хранения применяют секретные замки, которые открываются лишь тогда, когда набрано некоторое «тайное слово». Пусть на диск нанесено 12 букв, а секретное слово состоит из 5 букв. Сколько неудачных попыток может быть сделано человеком, не знающим секретного слова?

Решение. Общее число возможных комбинаций можно найти по формуле $N = \overline{A}_{12}^5 = 12^5 = 248832$.

Сочетаниями из n элементов по k элементов называются подмножества, состоящие из k элементов множества A_n

Одно сочетание от другого отличается только составом выбранных элементов (но не порядком их расположения, как у размещений).

Пример. Сочетаниями из четырех чисел 1, 2, 3, 4 по два являются: {1, 2}, {1, 3}, {1, 4}, { 2, 3}, {2, 4}, {3, 4}.

Число сочетаний из n элементов по k элементов обозначается C_n^k (читается "C из n по k").

Как подсчитать число сочетаний C_n^k? Заметим, что сочетание отличается от размещения только тем, что входящие в него элементы неупорядочены. Но мы уже знаем, что k элементов можно упорядочить (т.е. их по всякому переставить) $k!$ способами. Другими словами, имеем, что $A_n^k = k! \cdot C_n^k$. Отсюда находим, что $C_n^k = \frac{A_n^k}{k!}$ или

$$\boxed{C_n^k = \frac{n!}{(n-k)!k!}}$$

Примеры.

1)Сколько существует вариантов выбора 6-ти человек из 15 кандидатов для назначения на работу в одинаковых должностях?

$$C_{15}^6 = \frac{15!}{9!6!} = \frac{15 \cdot 14 \cdot 13 \cdot 12 \cdot 11 \cdot 10}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2} = 5005$$

2) Покупая карточку лотереи «Спортлото», игрок должен зачеркнуть 6 из 49 возможных чисел от 1 до 49. Сколько возможных комбинаций можно составить из 49 по 6, если порядок чисел безразличен?

Число возможных комбинаций можно рассчитать по формуле

$$N = C_{49}^6 = \frac{49!}{6!\,43!} = \frac{44 \cdot 45 \cdot 46 \cdot 47 \cdot 48 \cdot 49}{2 \cdot 3 \cdot 4 \cdot 5 \cdot 6} = 13983816.$$

Свойство числа сочетаний $\boxed{C_n^m = C_n^{n-m}}$

Сочетания с повторениями.

Сочетание с повторениями из n элементов по m элементов может содержать любой элемент сколько угодно раз от 1 до m включительно или не содержать его совсем, т.е. каждое сочетание из n элементов по m элементов может состоять не только из m различных элементов, но из m каких угодно и как угодно повторяющихся элементов.

Число сочетаний с повторениями из n элементов по m обозначают символом $\overline{C}_n^m$ и вычисляется по формуле:

$$\overline{C}_n^m = C_{n+m-1}^m = \frac{(n+m-1)!}{m!(n-1)!}.$$

В сочетаниях с повторениями m может быть и больше n .

Пример. В кондитерском магазине продавались 4 сорта пирожных: наполеоны, эклеры, песочные и слоеные. Сколькими способами можно купить 7 пирожных?

Решение. Число различных покупок равно числу сочетаний с повторениями из 4 по 7:

$$N = \overline{C_4^7} = C_{4+7-1}^7 = \frac{10!}{7!3!} = 120$$

Пусть множество из n элементов можно разбить на m упорядоченных частей (говорят еще на m групп или на m подмножеств), из которых первая содержит k_1 элементов, вторая —k_2 элементов и т.д ., m-ая — k_m элементов ($k_1 + ... + k_m = n$). Число таких способов разбиения есть

$$C_n(k_1, k_2, ..., k_m) = \frac{n!}{k_1! \cdot k_2! \cdot ... \cdot k_m!}$$

Пример. Десять человек размещаются в гостинице в 1-местный, 2-местный, 3-местный и 4-местный номера. Сколько существует способов их размещения?

Решение. По формуле $C_{10}(1,2,3,4)=\frac{10!}{1!2!3!4!}=12600$

Рассмотрим некоторые комбинаторные задачи.

1. Из 7 заводов организация должна выбрать 3 завода для размещения ***трех различных заказов***. Сколькими способами можно разместить заказы?

Так как из условия ясно, что каждый завод может либо получить один заказ, либо не получить ни одного, и что выбрав три завода, можно по-разному разместить среди них заказы, здесь нужно считать число размещений $A_7^3=\frac{7!}{4!}=7\cdot 6\cdot 5=210$

2. Если из текста задачи 1 убрать условие ***различия трех заказов,*** сохранив все остальные условия, получим другую задачу.

Теперь способ размещения заказов определяется только выбором тройки заводов, так как все эти заводы получат одинаковые заказы, и число вариантов определяется как число сочетаний $C_7^3=\frac{7!}{4!\cdot 3!}=35$

3. Имеются 7 заводов. Сколькими способами организация может разместить на них ***три различных*** производственных заказа? (Заказ нельзя дробить, то есть распределять его на нескольких заводах).

В отличие от условия первой задачи, здесь организация может отдать все три заказа первому заводу или, например, отдать два заказа второму заводу, а один - седьмому.

Задача решается так.

Первый заказ может быть помещен 7 различными способами (на первом заводе, на втором и т.д.). Поместив первый заказ, имеем 7 вариантов помещения **второго**.

Таким образом, существует $7 \cdot 7 = 49$ способов размещения первых двух заказов.

Разместив их каким-либо образом, можем найти 7 вариантов помещения **третьего**.

Следовательно, существуют $49 \cdot 7 = 7^3$ способов размещения трех заказов.

4. Добавим к условию задачи 1 **одну фразу**: организация также должна распределить три различных заказа на изготовление деревянных перекрытий среди 4-х лесопилок. Сколькими способами могут быть распределены ***все заказы***?

Каждый из A_7^3 способов распределения заказов на заводах может сопровождаться A_4^3 способами размещения заказов на лесопилках. Общее число возможных способов размещения всех заказов будет равно

$$A_7^3 \cdot A_4^3 = \frac{7!}{4!} \cdot \frac{4!}{1!} = 7!$$

Приведенные формулы из комбинаторного анализа позволяют осуществить непосредственное вычисление вероятности во многих задачах.

Примеры.

1. Имеется 6 карточек, на которых написаны буквы О, В, З, Д, У, Х. Карточки перемешаны. Какова вероятность того, что, доставая карточки наугад, мы получим слово «воздух».

Решение. Событие А (слово «воздух») содержит лишь один благоприятный вариант, т.е. m = 1. Число всех возможных элементарных исходов обозначает здесь число всех возможных упорядоченных наборов из 6 букв, т.е. число перестановок из 6 элементов. Следовательно,

$$n = 6! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 = 720 \;\Rightarrow\; p(A) = \frac{m}{n} = \frac{1}{720}.$$

2. Для участия в лотерее «Спортлото» на карточке, содержащей 49 чисел, нужно отметить 6 чисел. Какова вероятность выиграть?

Решение. Число благоприятных исходов m = 1. Число n всех элементарных исходов – это совокупность всех сочетаний из 49 чисел по 6 (нам неважно, в какой последовательности мы заполняем клетки), т.е. $n = C_{49}^{6}$.

Итак, $$p(A) = \frac{1}{C_{49}^{6}} = \frac{1}{13983816} = \frac{6!(49-6)!}{49!} = \frac{6!43!}{49!}.$$

3. На 9 вакантных мест по определенной специальности претендует 15 безработных, из них 7 женщин, остальные мужчины. Какова вероятность, что из 9 случайно отобранных безработных окажется 5 женщин?

Решение. A = {среди 9 отобранных ровно 5 женщин}. Число всех элементарных исходов $n = C_{15}^{9}$.

Число благоприятных событию A исходов найдем так: выбор 5 женщин из имеющихся 7, можно осуществить C_7^5 способами и одновременно еще выбор 4 мужчин и 8 (это можно сделать C_8^4 способами). Следовательно, по правилу умножения действий, имеем $m = C_7^5 \cdot C_8^4$.

Итак $$p(A) = \frac{m}{n} = \frac{C_7^5 \cdot C_8^4}{C_{15}^9} = \frac{42}{143}.$$

4. Первого сентября на первом курсе одного из факультетов запланировано по расписанию три лекции из 10 различных предметов. Студент, не успевший ознакомиться с расписанием, пытается его угадать. Какова вероятность успеха в данном эксперименте, если считать, что любое расписание из трех предметов равновозможно.

Решение. Студенту необходимо из 10 лекций, которые могут быть поставлены в расписание, причем в определенном порядке, выбрать три. Следовательно, число всех возможных исходов испытания равно числу размещений из 10 по 3, т.е.

$$n = A_{10}^3 = \frac{10!}{(10-3)!} = \frac{10!}{7!} = 10 \cdot 9 \cdot 8 = 720.$$

Благоприятных же случаев только один, т.е. m =1. Искомая вероятность будет равна $P = \frac{m}{n} = \frac{1}{720} \approx 0,0014$.

5. Найти вероятность того, что дни рождения 12 человек придутся на разные месяцы года.

Решение. Так как каждый из 12 человек может родиться в любом из 12 месяцев года, то число всех возможных вариантов можно посчитать по формуле размещений с повторениями: $n = \overline{A}_{12}^{12} = 12^{12}$.

Число благоприятных случаев получим, переставляя месяцы рождения у этих 12 человек, т.е. $m = P_{12} = 12!$.

Тогда искомая вероятность будет равна

$$P = \frac{m}{n} = \frac{12!}{12^{12}} = \frac{2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 \cdot 9 \cdot 10 \cdot 11 \cdot 12}{12 \cdot 12 \cdot 12 \cdot 12 \cdot 12 \cdot 12 \cdot 12 \cdot 12 \cdot 12 \cdot 12 \cdot 12 \cdot 12} = \frac{1925}{12^7} =$$

$$= \frac{1925}{35831808} \approx 0,00005372 \ .$$

Теоремы сложения и умножения вероятностей

Пользуясь алгеброй событий можно достаточно сложные случайные события представить в виде некоторого выражения от простейших случайных событий (для этого и нужна «алгебра»). Тогда естественно возникает вопрос: ***как, зная вероятности наступления простейших событий, вычислит вероятность сложного события?***

Начнем с простейших алгебраических выражений.

Найдем вероятность суммы событий $A \cup B$, используя классическое определение вероятности. Пусть в результате некоего опыта, состоящего из n исходов, событию A благоприятствует m_1 исходов, событию B – m_2 исходов, а одновременному появлению (AB) событий A и B благоприятствует k исходов. Тогда очевидно, сумме (объединению) событий $A \cup B$ будут благоприятствовать $(m_1 + m_2 - k)$ исходов. Следовательно,

$p(A+B) = \dfrac{m_1 + m_2 - k}{n}$. Но так как $p(A) = \dfrac{m_1}{n}$, $p(B) = \dfrac{m_2}{n}$ и $p(AB) = \dfrac{k}{n}$, то получаем формулу

$$p(A+B) = p(A) + p(B) - p(AB).$$

Теорема сложения вероятностей совместных событий. Вероятность появления хотя бы одного из двух совместных событий равна сумме вероятностей этих событий без вероятности их совместного появления:

$$P(A + B) = P(A) + P(B) - P(AB).$$

Если события A и B ***несовместны*** (т.е. $k=0$), то из (*) следует

$$p(A+B)=p(A)+p(B).$$

Теорема сложения вероятностей несовместных событий. Вероятность появления одного из двух несовместных событий, безразлично какого, равна сумме вероятностей этих событий:

$$P\,(A\,+\,B)\,=\,P\,(A)\,+P\,(B).$$

Отметим, что, так как $A+\overline{A}=\Omega$ – достоверное событие и $A\overline{A}=\varnothing$, то $p(A+\overline{A})=p(A)+p(\overline{A})=1 \Rightarrow p(A)=1-p(\overline{A})$.

Следствие. Сумма вероятностей противоположных событий равна 1: $P(A)+P(\overline{A})=1$.

Примеры.

1) В учебной группе 10 % студентов знает английский язык, 5 % – французский и 1 % – оба языка. Какова вероятность того, что наугад выбранный студент не знает ни одного языка?

Решение. p = 1 – (0,1 + 0,05 – 0,01) = 0,86.

2) Студент пришел на зачет, зная из 30 вопросов только 24. Преподаватель задает три вопроса. Зачет будет сдан, если студент ответит хотя бы на два из трех вопросов. Какова вероятность того, что этот студент сдаст зачет.

Решение. Пусть A_1 – событие, состоящее в том, что студент ответит на два из заданных трех вопросов, A_2 – он ответит на все три вопроса. Тогда, если A – студент сдаст зачет, то $A = A_1 + A_2$. События A_1 и A_2 несовместны. По классическому определению вероятности

$$P(A_1) = \frac{C_{24}^2 \cdot C_6^1}{C_{30}^3} = \frac{\frac{24!}{2! \cdot 22!} \cdot 6}{\frac{30!}{3! \cdot 27!}} = \frac{23 \cdot 24 \cdot 3}{28 \cdot 29 \cdot 5} = \frac{414}{1015} \approx 0,408,$$

$$P(A_2) = \frac{C_{24}^3}{C_{30}^3} = \frac{22 \cdot 23 \cdot 24}{28 \cdot 29 \cdot 30} = \frac{22 \cdot 23}{7 \cdot 29 \cdot 5} = \frac{506}{1015} \approx 0,499.$$

По теореме сложения для несовместных событий

$$P(A) = P(A_1) + P(A_2) = 0,408 + 0,499 = 0,907.$$

Пример. На стеллаже библиотеки в случайном порядке расставлено 15 учебников, причем пять из них в переплете. Библиотекарь берет наудачу четыре учебника. Найти вероятность того, что, по крайней мере, два из них в переплете.

Решение. Пусть A – событие, состоящее в том, что по крайней мере два из четырех взятых учебников будут в переплете. Это событие можно представить как сумму трех несовместных событий $A = A_2 + A_3 + A_4$, где A_2 – два учебника в переплете, A_3 – три учебника, A_4 – четыре учебника в переплете. Найдем вероятности этих событий. Число всех возможных исходов этого опыта

$$n = C_{15}^4 = \frac{15!}{4! \cdot 11!} = \frac{12 \cdot 13 \cdot 14 \cdot 15}{24} = 13 \cdot 7 \cdot 15 = 1365.$$

Для события A_2 число благоприятных исходов $m(A_2)=C_5^2\cdot C_{10}^2=10\cdot 45==450$,

для события A_3 – $m(A_3)=C_5^3\cdot C_{10}^1=10\cdot 10=100$,

для A_4 – $m(A_4)=C_5^4=5$.

Следовательно,

$$P(A_2)=\frac{450}{1365}=\frac{30}{91}, P(A_3)=\frac{100}{1365}=\frac{20}{273},\ P(A_4)=\frac{5}{1365}.$$

По теореме сложения для несовместных событий

$$P(A)=P(A_2)+P(A_3)+P(A_4)=\frac{30}{91}+\frac{20}{273}+\frac{1}{273}=\frac{111}{273}\approx 0,407$$

Условная вероятность – одно из основных понятий.

Именно условная вероятность оценивает то изменение в степени уверенности о наступлении некоторого события, которое происходит после получения дополнительной.

Вероятность события зависит от условий, при котором осуществляется опыт. На практике часто возникает ситуация, когда заранее известно, что некоторое событие В произошло. В этом случае В становится достоверным событием, а вероятности всех других событий некоторым образом меняются.

Определение. **Условной вероятностью** события A называется вероятность события A, вычисленная при условии, что произошло событие B. Условная вероятность события A при условии, что событие B произошло обозначается индексами $P(A/B)$ или $P_B(A)$.

В рамках классического определения вероятности условную вероятность естественно считать как отношение числа k благоприятных для совместного появления события AB, к числу исходов m,

благоприятствующих появлению (отдельно взятого) события B: $p(A/B)=\frac{k}{m}$.

Пусть n обозначает общее количество исходов данного опыта. Тогда после преобразований имеем

$$p(A/B)=\frac{k}{m}=\frac{k/n}{m/n}=\frac{p(AB)}{p(B)}.$$

Итак, **условная вероятность** события A при условии, что произошло событие B с $P(B)\neq 0$, определяется формулой $p(A/B)=\frac{p(AB)}{p(B)}$

Отсюда следует

Теорема умножения вероятностей. Вероятность совместного наступления двух событий равна произведению вероятности одного из них на условную вероятность другого, вычисленную в предположении, что первое событие уже наступило:

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

В частности для независимых событий

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

т.е. вероятность совместного наступления двух независимых событий равна произведению вероятностей этих событий.

Пример. B – выпадение четного числа на игральной кости; A – выпадение не менее 5 очков.

Тогда $p(A/B)=\frac{k}{m}=\frac{1}{3}$. Этот же результат можно получить из формулы $p(A/B)=\frac{p(AB)}{p(B)}=\frac{\frac{1}{6}}{\frac{3}{6}}=\frac{1}{6}:\frac{1}{2}=\frac{1}{3}$.

Аналогичная формула возникает, если события A и B поменять ролями:
$p(B/A)=\frac{p(AB)}{p(A)}$ и $p(AB)=p(A)\cdot p(B/A)$.

Свойства условных вероятностей
(аналогичны свойствам безусловной вероятности)**:**
1) $P_B(\Omega)=1$; 2) $P_B(\emptyset)=0$; 3) $0\le P_B(A)\le 1$; 4) если $A\subset C$, то $P_B(A)\le P_B(C)$; 5) $P_B(\overline{A})=1-P_B(A)$.

Если условная вероятность совпадает с безусловной $p(A/B)=p(A)$, то говорят, что события A и B являются независимыми. Другими словами, A и B – независимы, если вероятность события A не зависит от того, произошло или не произошло событие B.

Определение. **Событие** *A* называется независимым от события *B* (с $P(B)\neq 0$), если $P_B(A)=P(A)$, т.е. вероятность наступления события *A* не зависит от того, произошло событие *B* или нет.

Ясно, что если A не зависит от B, то и B не зависит от A (свойство взаимности).

Из определения независимости непосредственно следует, что $p(AB) = p(A) \cdot p(B)$ для случая А и В – независимы.

Формулу умножения можно обобщить на общий случай

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

Пример. Вероятность того, что потребитель увидит рекламу продукта по телевидению, равна 0,06. Вероятность того, что потребитель увидит рекламу того же продукта на рекламном стенде, равна 0,08. Предполагая, что оба события – независимы, определить вероятность того, что потребитель увидит: а) обе рекламы; б) хотя бы одну рекламу?

Решение. Пусть A – «Потребитель увидит рекламу по телевидению»; B – «Потребитель увидит рекламу на стенде»; C – «Потребитель увидит хотя бы одну рекламу». По условию $P(A) = 0{,}06$; $P(B) = 0{,}08$. События A и B совместные и независимые.

а) Потребитель увидит две рекламы. В наших обозначениях это событие $A \cdot B$,

$$P\left(A \cdot B\right) = P\left(A\right) \cdot P(B) = 0{,}06 \cdot 0{,}08 = 0{,}0048.$$

б) Событие C есть сумма событий A и B. Так как эти события ***совместны,*** то

$$P(C) = P(A + B) = P(A) + P(B) - P(AB).$$

$$P(C) = 0{,}06 + 0{,}08 - 0{,}0048 = 0{,}1352.$$

Эту же вероятность можно найти, используя свойство вероятностей противоположных событий

$$P\left(C\right) = P\left(A + B\right) = 1 - P\left(\overline{A} \cdot \overline{B}\right) = 1 - P\left(\overline{A}\right) \cdot P\left(\overline{B}\right);$$

$$P\left(\overline{A}\right) = 1 - P\left(A\right) = 1 - 0{,}06 = 0{,}94; \quad P\left(\overline{B}\right) = 1 - 0{,}08 = 0{,}92;$$

$$P(C) = 1 - 0,94 \cdot 0,92 = 1 - 0,8648 = 0,1352.$$

Замечание. $P\left(\sum_{i=1}^{n} A_i\right) = 1 - P\left(\overline{A}_1 \cdot \overline{A}_2 \cdot \ldots \cdot \overline{A}_n\right).$

В частности, если все n событий имеют одинаковую вероятность, равную p, то вероятность появления хотя бы одного из этих событий $P(A) = 1 - (1-p)^n$.

Пример. Вероятность хотя бы одного правильного ответа при опросе преподавателем четырех студентов равна 0,9984. Найти вероятность того, что наудачу выбранный студент правильно ответит на заданный вопрос.

Решение. Вероятность хотя бы одного правильного ответа при опросе четырех студентов определяется по формуле:

$$P = 1 - (1-p)^4,$$

где p – вероятность правильного ответа для одного наудачу выбранного студента. По условию $P = 0,9984$.

Решаем уравнение

$$1 - (1-p)^4 = 0,9984 \Rightarrow (1-p)^4 = 1 - 0,9984 \Rightarrow$$

$$(1-p)^4 = 0,0016 \Rightarrow (1-p)^2 = 0,04 \Rightarrow 1 - p = 0,2 \Rightarrow p = 0,8.$$

Пример. Сколько нужно произвести бросков монеты, чтобы с вероятностью не менее 0,9 выпал хотя бы один герб?

Решение. Вероятность выпадения герба при одном броске равна 0,5. Тогда вероятность выпадения хотя бы одного герба при n бросках равна $1 - (0,5)^n$. Тогда из неравенства $1 - (0,5)^n > 0,9$ следует, что $n > \log_2 10 \geq 4.$

При решении задач теоремы сложения и умножения обычно применяются вместе.

Пример. Два стрелка делают по одному выстрелу по мишени. Вероятности их попадания при одном выстреле равны соответственно 0,6 и 0,7. Найти вероятности следующих событий:

A – хотя бы одно попадание при двух выстрелах;

B – ровно одно попадание при двух выстрелах;

C – два попадания;

D – ни одного попадания.

Решение. Пусть событие H1 – попадание первого стрелка, H2 – попадание второго. Тогда

A = H1 + H2,

$B = H_1 \cdot \bar{H}_2 + \bar{H}_1 \cdot H_2$,

$C = H_1 \cdot H_2$, $D = \bar{H}_1 \cdot \bar{H}_2$.

События H1 и H2 совместны и независимы, поэтому теорема сложения применяется в общем виде, а теорема умножения – в виде (2.7). Следовательно,

P(A) = 0,6 + 0,7 – 0,42 = 0,88,

P(B) = 0,6·0,3 + 0,7·0,4 = 0,46

(так как события $H_1 \cdot \bar{H}_2$ и $\bar{H}_1 \cdot H_2$ несовместны),

P(C) = 0,6·0,7 = 0,42,

P(D) = 0,4·0,3 = 0,12.

Заметим, что события A и D являются противоположными, поэтому можно воспользоваться формулой

P(A) = 1 – P(D).

Формула полнлй вероятности и формула Байеса

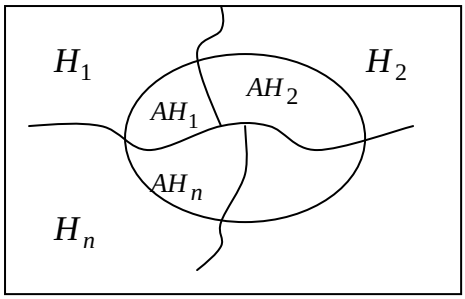
Пусть событие A может осуществиться вместе с одним из событий $H_1, H_2, \ldots, H_n$, причем последние образуют полную группу событий т.е.

1) $H_i \bigcap H_j = \varnothing$ – попарно несовместны,

2) $H_1 \bigcup \ldots \bigcup H_n = \Omega$ – их сумма есть достоверное событие.

Так как заранее неизвестно, с каким из событий $H_1, \ldots, H_n$ произойдет событие A, то H_i называют часто ***гипотезами.***

Эти события разбивают пространство Ω на непересекающиеся множества.



Тогда событие A можно представить в виде суммы несовместных событий:

$$A = A \bigcap \Omega = A \bigcap \left(\sum_{i=1}^{n} H_i \right) = \left(A \bigcap H_1 \right) \bigcup \left(A \bigcap H_2 \right) \bigcup \ldots \bigcup \left(A \bigcap H_n \right).$$

Здесь $\left(A \bigcap H_i \right)$ и $\left(A \bigcap H_j \right)$ $(i \neq j)$ попарно несовместны, т.к. гипотезы $H : \bigcap H_j = \varnothing$ несовместны. Тогда

$$p(A) = \sum_{i=1}^{n} p\left(A \bigcap H_i \right) = \sum_{i=1}^{n} p\left(H_i \right) p\left(A / H_i \right) \quad -$$

Формула полной вероятности. Пусть события $H_i,\ i=\overline{1,n}$ образуют полную группу событий ($P(H_i)>0$) и событие A может произойти с одним и только с одним из этих событий. Тогда вероятность события A равна

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

Т.о., если известны вероятности появления гипотез $p(H_i)$ и условная вероятность появления события A при реализации гипотезы H_i, $p(A/H_i)$, то (безусловная) вероятность $p(A)$ появления события A вообще в опыте определяется данной формулой.

Пример. Экспортно-импортная фирма собирается заключить контракт на поставку сельскохозяйственного оборудования в одну из развивающихся стран. Если основной конкурент фирмы не станет одновременно претендовать на заключение контракта, то вероятность получения контракта оценивается в 0,45; в противном случае – в 0,25. По оценкам экспертов компании вероятность того, что конкурент выдвинет свои предложения по заключению контракта, равна 0,40. Чему равна вероятность заключения контракта?

Решение. A = «фирма заключит контракт».

H_1 – «конкурент выдвинет свои предложения».

H_2 – « конкурент не выдвинет свои предложения».

По условию задачи $P(H_1)=0,4$, $P(H_2)=1-0,4=0,6$.

Условные вероятности по заключению контракта для фирмы $P(A/H_1)=0,25$, $P(A/H_2)=0,45$. По формуле

$$P(A)=P(H_1)\cdot P(A/H_1)+P(H_2)\cdot P(A/H_2).$$

$$P(A)=0,4\cdot 0,25+0,6\cdot 0,45=0,1+0,27=0,37$$

Пример. На экзамене студенту предварительно задается один из 20 вопросов. Если он правильно ответит, то с вероятностью 0,8 он может сдать экзамен, если не ответит, то в этом случае вероятность сдать экзамен ухудшается и равна 0,2. Студент освоил 70% всех вопросов (их всего 20). Какова вероятность для студента получить положительную оценку = сдать экзамен?

Решение. Пусть H_1 – попадется известный вопрос, H_2 – неизвестный, A – положительный исход экзамена.

$H_1+H_2=\Omega,\quad H_1\cap H_2=0\ \rightarrow$ полная группа.

$p(H_1)=0,7;\quad p(H_2)=0,3;\quad p(A/H_1)=0,7;\quad p(A/H_2)=0,2;$

$p(A)=p(H_1)\,p(A/H_1)+p(H_2)\,p(A/H_2)=0,7\cdot 0,8+0,3\cdot 0,2=$

$0,56+0,06=0,62.$

В предыдущей формуле были заранее, до опыта, известны вероятности $p(H_i)$ гипотез (до опытные вероятности называют еще *априорными* (из греческого)).

Однако, во многих приложениях встречается

следующая задача. Пусть до опыта имеется гипотезы $H_1,\ldots,H_n$. Предположим теперь, что **после опыта становится известной информация о его результатах**, но не полная. А именно, результаты наблюдений показывают, не какой конкретно исход (из $H_1,\ldots,H_n$) наступил, а что наступило некоторое событие А. Считая, что до опыта были известны априорные вероятности $p(H_1),\ldots,p(H_n)$ и условные

вероятности $p(A/H_i)$ требуется с учетом полученной информации о наступлении события А пересчитать вероятность наступления гипотезы H_k, $k=\overline{1,n}$, т.е. определить $p(H_k/A)$ (узнать, как изменились *апостериорные*, послеопытные, вероятности $p(H_k/A)$.

Эти новые вероятности гипотез даются формулой Байеса

$$p(H_k/A)=\frac{p(H_k)\,p(A/H_k)}{\sum_{i=1}^{n} p(H_i)\,p(A/H_i)}=\frac{p(H_k)\,p(A/H_k)}{p(A)}.$$

Для доказательства вспомним, что по определению условной вероятности $p(H_k/A)=\dfrac{p(AH_k)}{p(A)}$,

по формуле умножения вероятностей

$$p(AH_k)=p(H_k)\,p(A/H_k).$$

Тогда

$$p(H_k/A)=\frac{p(H_k)\,p(A/H_k)}{p(A)}.$$

Итак,

Формула Байеса. Если событие A произошло, то условные вероятности (апостериорные) гипотез H_i, $\left(i=\overline{1,n}\right)$ вычисляются по формуле, которая носит название формулы Байеса:

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

где $P(A)$ – вероятность события A, вычисленная по формуле полной вероятности.

Формула Байеса широко применяется в математической статистике.

Пример. Экономист-аналитик условно подразделяет экономическую ситуацию в стране на «хорошую», «посредственную» и «плохую» и оценивает их вероятности для данного момента времени в 0,15; 0,70 и 0,15 соответственно. Некоторый индекс экономического состояния возрастает с вероятностью 0,60, когда ситуация «хорошая»; с вероятностью 0,30, когда ситуация посредственная, и с вероятностью 0,10, когда ситуация «плохая». Пусть в настоящий момент индекс экономического состояния возрос. Чему равна вероятность того, что экономика страны на подъеме?

Решение. A = «индекс экономического состояния страны возрастет», H_1 = «экономическая ситуация в стране «хорошая»», H_2 = «экономическая ситуация в стране «посредственная»», H_3 = «экономическая ситуация в стране «плохая»». По условию: $P(H_1)=0{,}15$, $P(H_2)=0{,}70$, $P(H_3)=0{,}15$. Условные вероятности: $P(A/H_1)=0{,}60$, $P(A/H_2)=0{,}30$, $P(A/H_3)=0{,}10$. Требуется найти вероятность $P(H_1/A)$. Находим ее по формуле Байеса:

$$P(H_1/A)=\frac{P(H_1)\cdot P(A/H_1)}{P(H_1)P(A/H_1)+P(H_2)P(A/H_2)+P(H_3)P(A/H_3)}$$

$$P(H_1/A)=\frac{0{,}15\cdot 0{,}6}{0{,}15\cdot 0{,}6+0{,}7\cdot 0{,}3+0{,}15\cdot 0{,}1}=\frac{0{,}09}{0{,}09+0{,}21+0{,}015}$$

$$=\frac{0{,}09}{0{,}315}\approx 0{,}286.$$

Пример. 1. В канцелярии работают 4 секретарши, которые отправляют 40, 10, 30 и 20% исходящих бумаг. Вероятность неверной адресации бумаг равны 0,01; 0,04; 0,06; 0,01 соответственно. Известно, что документ неверно адресован. Найти вероятность того, что этот документ отправлен третьей секретаршей.

Решение.

H_i – документ отправлен i- секретаршей;

A- документ ошибочно адресован.

Из условия задачи следует

$$P(H_1)=0,4; P(H_2)=0,1; P(H_3)=0,3; P(H_4)=0,2;$$
$$P(A|H_1)=0,01; P(A|H_2)=0,04;$$
$$P(A|H_3)=0,06; P(A|H_4)=0,01;$$

Тогда

$$P(H_3|A)=\frac{P(H_3)P(A|H_3)}{P(A)}=....=0,391 \qquad \text{(Упр)}$$

Пример. В торговую фирму поступили телевизоры от трех поставщиков в соотношении 1:4:5. Практика показала, что телевизоры, поступающие от 1-го, 2-го и 3-го поставщиков, не потребуют ремонта в течение гарантийного срока соответственно в 98%, 88% и 92% случаев.

Найти вероятность того, что поступивший в торговую фирму телевизор не потребует ремонта в течение гарантийного срока.

Проданный телевизор потребовал ремонта в течение гарантийного срока. От какого поставщика вероятнее всего поступил этот телевизор.

Решение. Обозначим события:

H_i–телевизор поступил в торговую фирму от i-го поставщика ($i = 1,2,3$)

А – телевизор не потребует ремонта в течение гарантийного срока.

По условию:

$$P(H_1) = \frac{x}{x+4x+5x} = 0{,}1 \qquad P_{H_1}(A) = 0{,}98$$

$$P(H_2) = \frac{4x}{x+4x+5x} = 0{,}4 \qquad P_{H_2}(A) = 0{,}88$$

$$P(H_3) = \frac{5x}{x+4x+5x} = 0{,}5 \qquad P_{H_3}(A) = 0{,}92$$

Ответ на первый вопрос задачи найдем по формуле полной вероятности (2.9), а именно:

$$P(A) = 0{,}1 \cdot 0{,}98 + 0{,}4 \cdot 0{,}88 + 0{,}5 \cdot 0{,}92 = 0{,}91$$

Событие $\bar{A}$– телевизор потребует ремонта в течение гарантийного срока.

$$P(\bar{A}) = 1 - P(A) = 1 - 0{,}91 = 0{,}09$$

По условию $P_{H_1}(\bar{A}) = 1 - 0{,}98 = 0{,}02$

$$P_{H_2}(\bar{A}) = 1 - 0{,}88 = 0{,}12 \qquad P_{H_3}(\bar{A}) = 1 - 0{,}92 = 0{,}08$$

По формуле Байеса (2.11)

$$P_{\bar{A}}(H_1) = \frac{0{,}1 \cdot 0{,}02}{0{,}09} = 0{,}022 \qquad P_{\bar{A}}(H_2) = \frac{0{,}4 \cdot 0{,}12}{0{,}09} = 0{,}533$$

$$P_{\bar{A}}(H_3) = \frac{0{,}5 \cdot 0{,}08}{0{,}09} = 0{,}444$$

Интерпретация результата: После наступления события $\bar{A}$ вероятность гипотезы H_2 увеличилась с $P(H_2) = 0{,}4$ до максимальной $P_{\bar{A}}(H_2) = 0{,}533$, а гипотезы H_3– уменьшилась от максимальной $P(H_3) = 0{,}5$ до $P_{\bar{A}}(H_3) = 0{,}444$. Если ранее, до наступления события A, наиболее вероятной была гипотеза H_3, то теперь, в свете новой информации (наступления события A), наиболее вероятна гипотеза H_2– поступление данного телевизора от 2-го поставщика.

Задачи для самостоятельной работы

Пример 1. Во время воздушной атаки самолет противника выпустил 3 помехи для подавления системы ПВО. Оператор ПВО отличает самолет от помехи с вероятностью 0,9. Точка на экране оператора ПВО была принята за самолет. Что более вероятно: появление самолета или помехи?

Пример 2. Два стрелка независимо друг от друга стреляют по одной мишени, делая каждый по одному выстрелу. Вероятность попадания первого стрелка равна 0,8, второго — 0,4. После стрельбы обнаружили, что в мишени одна пробоина. Найти вероятность, что в мишень попал первый стрелок.

Пример 3. В урне 5 белых и 4 черных шара. Наудачу извлекают один шар, а затем другой. Найти вероятность того, что во втором случае был вынут белый шар (шары в урну не возвращаются).

Схема испытаний Бернулли

Ряд задач ТВ связан с экспериментом, в котором проводятся п**оследовательные независимые** испытания

Последовательные испытания называются **независимыми**, если вероятность осуществления любого исхода в n-м по счету испытании **не зависит** от реализации исходов предыдущих испытаний.

Простейшим классом повторных независимых испытаний является **последовательность независимых испытаний с двумя исходами** («успех» и «неуспех») **и с неизменными вероятностями «успеха» (*p*) и «неуспеха»** $(1-p=q)$ **в каждом испытании (схема испытаний Бернулли).**

Вероятность получить **ровно** m успехов в n независимых испытаниях вычисляется по **формуле Бернулли**: $P_{n,m}=C_n^m\cdot p^m(1-p)^{n-m}$.

Доказательство. Пусть A_i и $\bar{A}_i$ – соответственно появление и непоявление события A в i-м испытании ($i=1,2,\dots,n$), а B_m – событие, состоящее в том, что в n независимых испытаниях событие A появилось m раз. Представим событие B_m через элементарные события A_i. Например, при $n=3,\ m=2$ событие $B_2=A_1A_2\bar{A}_3+A_1\bar{A}_2A_3+\bar{A}_1A_2A_3$.

В общем виде

$$B_m=A_1A_2\dots A_m\hat{\bar{A}}_{m+1}\dots\bar{A}_n+A_1\bar{A}_2A_3\dots A_m\bar{A}_{m+1}\dots\bar{A}_{n-1}A_n+\dots+\bar{A}_1\bar{A}_2\dots\hat{\bar{A}}_{n-m}A_{n-m+1}\dots A_n,\quad (*)$$

Т.е. каждый вариант появления события B_m (каждый член суммы (*)) состоит из m событий A и $n-m$ событий $\bar{A}$ с различными индексами. Число всех комбинаций (слагаемых суммы (*)) равно числу способов выбора из n испытаний m, в которых событие A произошло, т.е. числу сочетаний C_n^m. Вероятность каждой такой комбинации по теореме умножения для независимых событий равна $p^m q^{n-m}$. В связи с тем, что комбинации между собой несовместны, то по теореме сложения вероятностей получим

$$P_{n,m} = P(B_m) = \underbrace{p^m q^{n-m} + \ldots + p^m q^{n-m}}_{C_n^m \text{ раз}} = C_n^m p^m q^{n-m}.$$

Пример. Изделия производства содержат 5 % брака. Найти вероятность того, что среди пяти взятых наугад изделий: а) нет ни одного испорченного; б) будут два испорченных.

Решение. а) По условию задачи $n = 5,\ p = 0{,}05$. Так как вероятность наступления события A (появление бракованной детали) постоянна для каждого испытания, то задача ***подходит под схему Бернулли***. По формуле

$$P_{5,0} = C_5^0 \cdot 0{,}05^0 \cdot 0{,}95^5 = 1 \cdot 1 \cdot 0{,}774 = 0{,}774.$$

б) $n = 5,\ m = 2,\ p = 0{,}05$: $P_{5,2} = C_5^2 \cdot 0{,}05^2 \cdot 0{,}95^3 = 0{,}021.$

Пример. Радиоактивная частица пролетает последовательно мимо 6 счетчиков. Каждый счетчик независимо друг от друга отмечает пролет частицы с вероятностью 0,8. Частица считается зарегистрированной (событие А) , если она отмечена по крайней мере 2 счетчиками. Найти вероятность регистрации частицы.

Решение. Задача ***подходит под схему Бернулли***.
Можно вычислять по формуле
$P(A) = P_{6,2} + P_{6,3} + P_{6,4} + P_{6,5} + P_{6,6}$
Удобнее с помощью отрицания (т.е через противоположные события)
$P(A) = 1 - P(\overline{A}) = 1 - (P_{6,0} + P_{6,1}) = 1 - C_6^0 p^0 q^6 - C_6^1 p^1 q^5 =$
$1 - 0,0016 = 0,9984$

Наивероятнейшее число наступлений события A

Определение. Число наступлений события A называется **наивероятнейшим**, если оно имеет наибольшую вероятность по сравнению с вероятностями наступления события A любое другое количество раз.

Наивероятнейшее число m_0 наступлений события A в n испытаниях заключено в интервале

$$np - q \le m_0 \le np + p.$$

Если $np - q$ – целое число, то наивероятнейших числа два $np - q$ и $np + p$.

Пример. В помещении четыре лампы. Вероятность работы в течение года для каждой лампы 0,8. Чему равно наивероятнейшее число ламп, которые будут работать в течение года?

Решение. **Из** неравенства $np - q \le m_0 \le np + p$ найдем m_0. По условию $n = 4,\ p = 0,8,\ q = 1 - 0,8 = 0,2$:

$$4 \cdot 0,8 - 0,2 \le m_0 \le 4 \cdot 0,8 + 0,8 \Leftrightarrow \qquad 3 \le m_0 \le 4.$$

Имеется два наивероятнейших числа $m_0 = 3$ или $m_0 = 4$.

Пример. Вероятность попадания в кольцо при штрафном броске для баскетболиста равна 0,8. Сколько надо произвести ему бросков, чтобы наивероятнейшее число попаданий было равно 20?

Решение. Известно, что $p = 0{,}8,\ m_0 = 20$. Тогда $q = 1 - 0{,}8 = 0{,}2$ и n найдем из системы неравенств

$$\begin{cases} n \cdot 0{,}8 - 0{,}2 \le 20 \\ n \cdot 0{,}8 + 0{,}8 \ge 20 \end{cases} \Leftrightarrow \begin{cases} n \le \dfrac{20{,}2}{0{,}8} \\ n \ge \dfrac{19{,}2}{0{,}8} \end{cases} \Leftrightarrow \quad 24 \le n \le 25{,}25.$$

Пример. Для того, чтобы проверить точность своих финансовых счетов, фирма приглашает аудиторов для проверки бухгалтерских проводок. Служащие фирмы при обработке счетов допускают примерно 5% ошибок. Аудитор наугад отобрал 20 документов. Найти наивероятнейшее число документов, в которых может быть ошибка.

Решение. Известно, что $n = 20,\ p = 0{,}05,\ q = 0{,}95$. Тогда найдем из системы неравенств $np - q \le m_0 \le np + p$. Имеем

$20 \cdot 0{,}05 - 0{,}95 \le m_0 \le 20 \cdot 0{,}05 + 0{,}05$ или

$1 - 0{,}95 \le m_0 \le 1 + 0{,}05$. Значит, $\boxed{m_0 = 1}$

Предельные теоремы для схемы Бернулли

Предположим, что мы хотим вычислить вероятность $P_n(m)$ появления события A при большом числе испытаний n, например, $P_{500}(300)$. По формуле Бернулли имеем: $\boxed{P_{500}(300) = C_{500}^{300} p^{300} q^{200}}$. Ясно, что в этом случае непосредственное вычисление по формуле Бернулли технически сложно, тем более, если учесть, что сами p и $q-$ числа дробные. Поэтому возникает естественное желание иметь более простые, пусть даже и приближенные, формулы для вычисления $P_n(m)$ при больших n.

Такие формулы, называемые асимптотическими, существуют, среди которых наиболее известны теорема Пуассона, локальная и интегральная теоремы Лапласа.

Теорема (Пуассона). Предположим, что произведение np является постоянной величиной, когда n неограниченно возрастает. Обозначим $\lambda = np$. Тогда для любого фиксированного m и любого постоянного λ:

$$\lim_{\substack{n\to\infty \\ np=\lambda}} P_{m,n} = \frac{\lambda^m}{m!} e^{-\lambda}.$$

Доказательство. Запишем формулу Бернулли

$$P_{n,m} = C_n^m \cdot p^m (1-p)^{n-m} = \frac{n(n-1)...(n-m+1)}{m!} p^m (1-p)^{n-m} = \otimes$$

$= ($обозначения $p=\frac{\lambda}{n})$

$$= \frac{\lambda^m}{m!} \cdot \frac{n-1}{n} \cdot \frac{n-2}{n} \cdots \frac{n-m+1}{n} \cdot (1-\frac{\lambda}{n})^n (1-\frac{\lambda}{n})^{-m} =$$

=(Известно, что $\lim\limits_{n\to\infty}(1-\frac{\lambda}{n})^n = e^{-\lambda}$. Кроме того

$\frac{n-1}{n} \to 1, ..., \frac{n-m+1}{n} \to 1,$ $(1-\frac{\lambda}{n})^{-m} \to 1$ при $n \to \infty$)=

$\otimes \to \frac{\lambda^m}{m!} \cdot e^{-\lambda}$ что и завершает доказательство.

Следствие. В случае, **когда *n* велико, а *p* мало** ($p < 0{,}1$; $npq \le 9$) вместо формулы Бернулли применяют приближенную формулу Пуассона

$$P_{m,n} \approx \frac{\lambda^m e^{-\lambda}}{m!}, \text{ где } \lambda = np.$$

Пример. Прядильщица обслуживает 1000 веретен. Вероятность обрыва нити на одном веретене в течение одной минуты равна 0,004. Найти вероятность того, что в течение одной минуты обрыв произойдет на пяти веретенах.

Решение. Для определения вероятности $P_{5;1000}$ применим приближенную формулу Пуассона:

$$\lambda = np = 0{,}004 \cdot 1000 = 4; \quad P_{5;1000} \approx \frac{4^5 \cdot e^{-4}}{5!} \approx 0{,}1563.$$

(Значение функции Пуассона найдено по таблице 3 для $m=5$ и $\lambda=4$).

Пример. На факультете насчитывается 1825 студентов. Какова вероятность того, что 1 сентября является днем рождения одновременно четырех студентов факультета?

Решение. Вероятность того, что день рождения студента 1 сентября, равна $p=1/365\approx 0,027<0,1$. Число $n=1825>100$ – велико и $\lambda=np=1825\cdot 1/365=5\le 10$. По формуле Пуассона: $P_{1825}(4)\approx\dfrac{5^4e^{-5}}{4!}\approx 0,1755$.

Теоремы Муавра-Лапласа

Теорема (Муавра-Лапласа (локальная)). Если вероятность наступления события A в каждом из n независимых испытаниях равна p и отлична от нуля и единицы, а число испытаний достаточно велико, то вероятность $P_{m,n}$ того, что в n испытаниях события A наступит m раз, приближенно равна (чем больше n, тем точнее) $\boxed{P_{n,m}\approx\dfrac{1}{\sqrt{npq}}\cdot\varphi(x)}$,

где $x=\dfrac{m-np}{\sqrt{npq}}$, $\varphi(x)=\dfrac{1}{\sqrt{2\pi}}\cdot e^{-x^2/2}$ — функция Гаусса или «малая» функция Лапласа

Таблица значений функции $\varphi(x)$ приведена в приложениях книг по ТВ.

Пользуясь этой таблицей, необходимо иметь в виду очевидные свойства функции $\varphi(x)$:

1. Функция $\varphi(x)$ является четной, т.е. $\varphi(-x)=\varphi(x)$
2. Функция $\varphi(x)$–монотонно убывающая при положительных значениях x, причем при $x\to\infty,\ \varphi(x)\to 0$.
3. Практически можно считать, что уже при $x>4\ \ \varphi(x)\approx 0$

Упр. Свойства функции $\varphi(x)=\frac{1}{\sqrt{2\pi}}\cdot e^{-x^2/2}$

Пример. Вероятность найти белый гриб среди прочих равна $\frac{1}{4}$. Какова вероятность того, что среди 300 грибов белых будет 75?

Решение. По условию задачи $p=\frac{1}{4}, m=75, n=300,$ $q=1-p=1-\frac{1}{4}=\frac{3}{4}$. Находим $x=\frac{m-np}{\sqrt{npq}}=\frac{75-300\cdot\frac{1}{4}}{\sqrt{300\cdot\frac{1}{4}\cdot\frac{3}{4}}}=0.$

По таблице находим $\varphi(x)=0{,}3989$.

$$P_{75;300}=\frac{\varphi(x)}{\sqrt{npq}}=\frac{0{,}3989}{\sqrt{\frac{900}{16}}}=\frac{4\cdot 0{,}3989}{30}\approx 0{,}053.$$

Замечание. Пусть в условиях примера 5 необходимо найти вероятность того, что белых грибов будет, например, от 70 до 150. В этом случае по теореме сложения вероятность искомого события

$$P_{300}(70\le m\le 150)=P_{300}(70)+P_{300}(71)+P_{300}(72)+\ldots+P_{300}(150).$$

В принципе вычислить каждое слагаемое можно по локальной формуле Муавра-Лапласа, но большое количество слагаемых делает расчет весьма громоздким. В таких случаях используется следующая теорема.

Теорема (Муавра-Лапласа (интегральная)).

Если вероятность наступления события A в каждом из n независимых испытаниях равна p и отлична от нуля и единицы, а число испытаний достаточно велико, то вероятность того, что в n испытаниях число успехов m находится между m_1 и m_2 приближенно равна (чем больше n, тем точнее)

$$P(m_1 \le m \le m_2) \approx \frac{1}{2}(\Phi(x_2) - \Phi(x_1)),$$

где $q = 1 - p$, $x_i = \dfrac{m_i - np}{\sqrt{npq}}, i = 1,2$

$$\Phi(x) = \frac{2}{\sqrt{2\pi}} \cdot \int_0^x e^{\frac{-t^2}{2}}\, dt$$

функция Лапласа (большая)
(см. значения функции в Приложениях книг)

Отметим свойства функции $\Phi(x)$:

1. Функция $\Phi(x)$ является нечетной, т.е. $\Phi(-x) = -\Phi(x)$
2. Функция $\Phi(x)$–монотонно возрастающая при положительных значениях x, причем при $x \to +\infty, \Phi(x) \to 1$.
3. Практически можно считать, что уже при $x > 4$ $\Phi(x) \approx 1$.

Упр. Свойства функции $\Phi(x)=\frac{2}{\sqrt{2\pi}}\cdot\int_0^x e^{\frac{-t^2}{2}}dt$

Пример. В партии из 768 арбузов каждый арбуз оказывается неспелым с вероятностью $\frac{1}{4}$. Найти вероятность того, что количество спелых арбузов будет в пределах от 564 до 600.

Решение. Имеем $n=768,\ p=0{,}75,\ m_1=564,\ m_2=600.$

По интегральной теореме Лапласа

$$P(564\le m\le 600)\approx$$

$$\approx\frac{1}{2}\left(\Phi\left(\frac{600-768\cdot 0{,}75}{\sqrt{768\cdot 0{,}25\cdot 0{,}75}}\right)-\Phi\left(\frac{564-768\cdot 0{,}75}{\sqrt{768\cdot 0{,}25\cdot 0{,}75}}\right)\right)=$$

$$=\frac{1}{2}\left(\Phi\left(\frac{600-576}{12}\right)-\Phi\left(\frac{564-576}{12}\right)\right)=\frac{1}{2}\left(\Phi(2)+\Phi(1)\right)\approx$$

$$\approx\frac{1}{2}(0{,}9545+0{,}6827)=0{,}8186.$$

Пример. Город ежедневно посещает 1000 туристов, которые днем идут обедать. Каждый из них выбирает для обеда один из двух городских ресторанов с равными вероятностями и независимо друг от друга. Владелец одного из ресторанов желает, чтобы с вероятностью приблизительно 0,99 все пришедшие в его ресторан туристы могли там одновременно пообедать. Сколько мест должно быть для этого в его ресторане?

Решение. Пусть A = «турист пообедал у владельца». Наступление события A будем считать «успехом»,

$p = P(A) = 0{,}5$, $n = 1000$. Нас интересует такое наименьшее число k, что вероятность наступления не менее чем k «успехов» в последовательности из $n = 1000$ независимых испытаний с вероятностью успеха p = 0,5 равна 1–0,99=0,01. Это как раз вероятность переполнения ресторана. Таким образом, нас интересует такое наименьшее число k, что $P(k \le \mathrm{m} \le 1000) \approx 0{,}01$. Применим интегральную теорему Муавра-Лапласа:

$$P(k \le m \le 1000) \approx 0{,}01 \approx \frac{1}{2}\left(\Phi\left(\frac{1000-500}{\sqrt{250}}\right) - \Phi\left(\frac{k-500}{\sqrt{250}}\right)\right) \approx$$

$$\approx \frac{1}{2}\left(\Phi\left(\frac{100}{\sqrt{10}}\right) - \Phi\left(\frac{k-500}{5\sqrt{10}}\right)\right) \approx \frac{1}{2}\left(1 - \Phi\left(\frac{k-500}{5\sqrt{10}}\right)\right).$$

Откуда следует, что $\Phi\left(\frac{k-500}{5\sqrt{10}}\right) \approx 0{,}98$.

Используя таблицу для $\Phi(x)$, находим $\frac{k-500}{5\sqrt{10}} \approx 2{,}33$,

Отсюда $k = 2{,}33 \cdot 5\sqrt{10} + 500 \approx 536{,}8$. **Ответ:** 537 мест.

Пример. В страховой компании 10 000 клиентов. Страховой взнос каждого клиента 500 рублей. При наступлении страхового случая, вероятность которого по оценкам экспертов равна $p = 0{,}005$. Страховая компания обязана выплатить клиенту страховую сумму в размере 50 000 рублей. На какую прибыль может рассчитывать компания с надежностью 0,95?

Решение. Прибыль равна $\Pi = 500 \cdot 10000 - 50000 \cdot n_0$ рублей, где n_0 число клиентов, для которых наступил страховой случай. Применим интегральную теорему Муавра-Лапласа:

$$P_{10000}(0 \le m \le n_0) = \frac{1}{2}(\Phi(x_2) - \Phi(x_1) = 0{,}95, \qquad (***)$$

где m есть число клиентов, которым будет выплачена страховка,

$$x_1 = \frac{0 - np}{\sqrt{npq}} = -\sqrt{\frac{np}{q}} = -\sqrt{\frac{10000 \cdot 0{,}005}{0{,}995}} = -7{,}09; x_2 = \frac{n_0 - np}{\sqrt{npq}}$$

Откуда следует, что

$$n_0 = np + x_2\sqrt{npq} = 10000 \cdot 0{,}005 + x_2\sqrt{49{,}75} = 50 + x_2\sqrt{49{,}75}$$

Из (***) следует

$$\Phi(x_2) = 1{,}9 + \Phi(x_1) = 1{,}9 + \Phi(-7{,}09) = 1{,}9 + (-1) = 0{,}9.$$

Используя таблицу для $\Phi(x)$, находим из уравнения $\Phi(x_2) = 0{,}9$, значение $x_2 = 1{,}645$. Теперь $n_0 = 50 + x_2\sqrt{49{,}75} = 50 + 1{,}645\sqrt{49{,}75} = 61{,}6$. Следовательно, прибыль равна

$$\Pi = 500 \cdot 10000 - 50000 \cdot n_0 = 500 \cdot 10000 - 50000 \cdot 61{,}6 = 1920000$$

Ответ: 1920 тыс. рублей.

Из интегральной теоремы Лапласа можно получить формулу **(Упр)**

$$P\left(\left|\frac{m}{n} - p\right| < \varepsilon\right) \approx \Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right).$$

Пример. 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. По условию $p = 0{,}5,\ q = 0{,}5,\ \varepsilon = 0{,}02$;

$$P\left(\left|\frac{m}{n} - 0{,}5\right| \le 0{,}02\right) = 0{,}7698.$$

Воспользуемся формулой $P\left(\left|\frac{m}{n} - p\right| \le \varepsilon\right) \approx \Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right)$:

$$\Phi\left(0{,}02\sqrt{\frac{n}{0{,}5 \cdot 0{,}5}}\right) = 0{,}7698 \quad \Rightarrow \quad 0{,}04\sqrt{n} = 1{,}2 \quad \Rightarrow$$

$$\Rightarrow \quad \sqrt{n} = 30 \Rightarrow \quad n = 900$$

Ответ: $n = 900$.

Случайные величины

7

Случайной называют величину, которая в результате испытания примет одно и только одно из возможных значений, *наперед неизвестное* и зависящее от случайных причин, которые заранее нельзя учесть.

Случайные величины (СВ) обозначают буквами X, Y, Z, а их возможные значения – x, y, z…..

Примеры:

- число очков, выпавших при броске игральной кости;
- число появлений герба при 10 бросках монеты;
- число выстрелов до первого попадания в цель;
- расстояние от центра мишени до пробоины.

Можно заметить, что множество возможных значений для перечисленных случайных величин имеет разный вид: для первых трех величин множество значений из отдельных (дискретных), изолированных друг от друга значений, а для четвертой оно представляет собой непрерывную область. По этому показателю случайные величины подразделяются на две группы: дискретные и непрерывные.

Случайная величина называется дискретной (ДСВ), если множество $\{x_1, x_2, ..., x_n, ...\}$ ее возможных значений конечно или счетно (т.е. если все ее значения можно занумеровать).

Случайная величина называется непрерывной (НСВ), если множество ее возможных значений целиком заполняет некоторый конечный или бесконечный интервал или системы интервалов на числовой оси.

Дискретная случайная величина

Дискретной называют случайную величину, которая принимает отдельные, изолированные друг от друга значения с определенными вероятностями.

Число возможных значений дискретной случайной величины может быть конечным или бесконечным (счетным).

Как охарактеризовать СВ?

Наиболее полным, исчерпывающим описанием случайной величины является ее закон распределения.

Законом распределения случайной величины называется всякое соотношение, устанавливающее связь между возможными значениями случайной величины и вероятностями, с которыми она принимает эти значения. В этом случае про случайную величину говорят, что она распределена по данному закону распределения или подчинена этому закону распределения

Дискретная случайная величина может быть задана ***рядом распределения*** – это таблица, в которой перечислены все возможные значения дискретной случайной величины и соответствующие им вероятности:

X	x_1	x_2	…	x_n
P	p_1	p_2	…	p_n

$$p_i = P(X = x_i),\ i = \overline{1,n}.$$

События $X = x_1, X = x_2, \ldots, X_n = x_n$ образуют полную группу, следовательно, сумма вероятностей этих событий равна единице: $p_1 + p_2 + p_3 + \ldots p_n = 1.$

Пример 1. Два стрелка делают по одному выстрелу по мишени. Вероятности их попадания при одном выстреле равны соответственно 0,6 и 0,7. Составить ряд распределения случайной величины X – числа попаданий после двух выстрелов.

Решение. Очевидно, что X может принимать три значения: 0, 1 и 2. Найдем их вероятности:

Пусть события A_1 и A_2 – попадание по мишени соответственно первого и второго стрелка. Тогда

$$P(X=0)=P(\bar{A}_1\bar{A}_2)=0,4\cdot 0,3=0,12$$

$$P(X=1)=P(\bar{A}_1A_2+A_1\bar{A}_2)=0,4\cdot 0,7+0,6\cdot 0,3=0,46$$

$$P(X=2)=P(A_1A_2)=0,6\cdot 0,7=0,42$$

Следовательно, ряд распределения имеет вид:

x_i	0	1	2
p_i	0,12	0,46	0,42

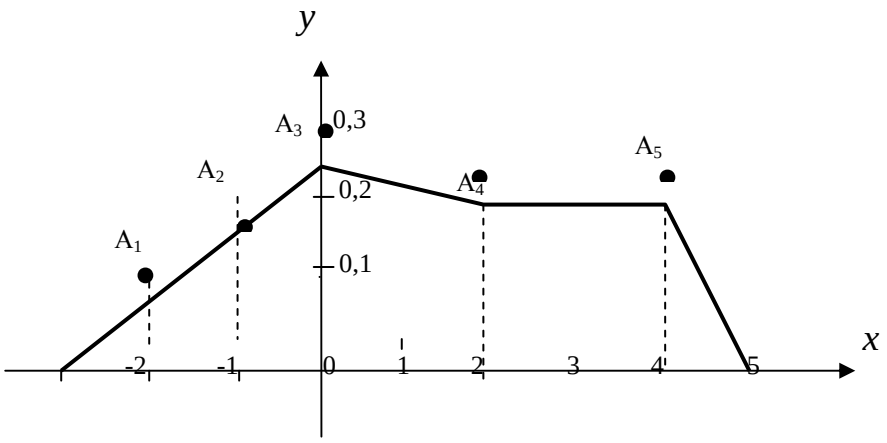
Ряд распределения дискретной случайной величины можно изобразить графически в виде ***полигона*** или ***многоугольника распределения*** вероятностей. Для этого по горизонтальной оси в выбранном масштабе нужно отложить значения случайной величины, а по вертикальной вероятности этих значений. Тогда точки с координатами (x_i, p_i) будут изображать полигон распределения вероятностей, соединив же эти точки отрезками прямой, получим многоугольник распределения вероятностей.

Пример 2. Пусть X – дискретная случайная величина, заданная рядом распределения

x_i	-2	-1	0	2	4
p_i	0,1	0,2	0,3	0,2	0,2

Построить полигон и многоугольник распределения вероятностей.

Решение. На оси X откладываем значения x_i, равные – 2, – 1, 0, 2, 4, а по вертикальной оси вероятности этих значений:



Точки A_1, A_2, A_3, A_4, A_5 изображают полигон распределения, а ломаная $A_1\,A_2\,A_3\,A_4\,A_5$ – многоугольник распределения вероятностей.

Пример 3. В лотерее разыгрывается: автомобиль стоимостью 5000 ден.ед., 4 телевизора стоимостью 250 ден. ед., 5 видеомагнитофонов стоимостью 200 ден.ед. Всего продается 1000 билетов по 7 ден.ед. Составить закон распределения чистого выигрыша, полученного участником лотереи, купившим один билет.

Решение. Возможные значения случайной величины X–чистого выигрыша на один билет равны $0-7=-7$ ден.ед. (если билет не выиграл), $200-7=193$, 250-7=243, 5000-7=4993 ден.ед.(если на билет выпал выигрыш соответственно видеомагнитофона, телевизора или автомобиля). Учитывая, что из 1000 билетов число

невыигравших составляет 990, а указанных выигрышей соответственно 5, 4 и 1, используя классическое определение вероятности, получим:

$$P(X=-7)=\frac{990}{1000}=0,990$$

$$P(X=193)=\frac{5}{1000}=0,005$$

$$P(X=243)=\frac{4}{1000}=0,004$$

$$P(X=4993)=\frac{1}{1000}=0,001$$

Ряд распределения имеет вид:

x_i	-7	193	243	4993
p_i	0,990	0,005	0,004	0,001

Математические операции над ДСВ.

Пусть X — ДСВ, заданная своим рядом распределения

X	x_1	x_2	…	x_n
P	p_1	p_2	…	p_n

Произведением kX случайной величины X на постоянное число k называется случайная величина, которая с теми же вероятностями, что и X принимает новые значения, равные kx_i, т.е. закон распределения для новой случайной величины kX имеет вид:

kX	kx_1	kx_2	…	kx_n
P	p_1	p_2	…	p_n

Пусть Y — еще одна дискретная СВ:

Y	y_1	y_2	…	y_m
P	p_1	p_2	…	p_m

Сумма $Z = X + Y$ двух случайных величин — эта новая случайная величина, которая принимает все значения вида $z_k = x_i + y_j, i = 1,...,n, j = 1,...,m$ с вероятностями

$$p_k = \sum_{i,j: x_i + y_j = z_k} p_{ij} \quad \text{где} \quad p_{ij} = P(\{X = x_i\} \cdot \{Y = y_j\})$$

Произведением $Z = X \cdot Y$ называется новая случайная величина, принимающая все значения вида $z_k = x_i + y_j, i = 1,...,n, j = 1,...,m$ с вероятностями

$$p_k = \sum_{i,j: x_i + y_j = z_k} p_{ij} \quad \text{где} \quad p_{ij} = P(\{X = x_i\} \cdot \{Y = y_j\})$$

В частности, ***квадрат*** случайной величины X^2 — эта новая случайная величина, которая с теми же вероятностями, что и X принимает значения x_i^2.

Пример.

x_i	- 1	0	1	2
p_i	0,1	0,2	0,25	0,45

x_i^2	0	1	4
p_i	0,2	0,35	0,45

Две случайные величины X, Y называются **независимыми**, если вероятность совместного появления событий равна произведению вероятностей

$$P(\{X = a\} \cdot \{Y = b\}) = P(\{X = a\}) \cdot P(\{Y = b\}).$$

Числовые характеристики случайной величины

Ряд распределения полностью описывает ДСВ. Но такая подробность в описании зачастую не требуется (утомительна, громоздка, неинформативна….). Важно знать некие обобщенные числовые характеристики.

Математическое ожидание *M (X)*

Пусть случайная величина X может принимать только значения $x_1, x_2, \ldots, x_n$, вероятности которых соответственно равны $p_1, p_2, \ldots, p_n$.

Тогда математическое ожидание $M(X)$ случайной величины X определяется равенством

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

Из определения следует, что математическое ожидание дискретной случайной величины ***есть неслучайная*** (постоянная) величина.

Свойства математического ожидания

1. Математическое ожидание постоянной величины равно самой постоянной: $M(C) = C$.
2. Постоянный множитель можно выносить за знак математического ожидания: $M(CX) = CM(X)$.
3. Математическое ожидание алгебраической суммы конечного числа случайных величин равно алгебраической сумме их математических ожиданий:
$$M(X \pm Y) = M(X) \pm M(Y).$$
4. Математическое ожидание произведения конечного числа ***независимых*** случайных величин равно произведению их математических ожиданий: $M(XY) = M(X)\,M(Y)$
5. Математическое ожидание отклонения случайной величины от ее математического ожидания равно нулю: $M(X - M(X)) = 0$.

Действительно,

$$M(X - M(X)) = M(X) - M(M(X)) = M(X) - M(X) = 0$$

Дисперсия случайной величины

На практике часто требуется оценить ***рассеяние*** возможных значений случайной величины вокруг ее среднего значения.

Дисперсией $D(X)$ случайной величины X называется математическое ожидание квадрата ее отклонения от ее математического ожидания: $D(X)=M\left[X-M(X)\right]^2$.

Дисперсия – это мера рассеяния случайной величины около ее математического ожидания.

Если X – дискретная случайная величина, то дисперсию вычисляют по следующим формула $D(X)=\sum_{i=1}^{n}(x_i-a)^2 p_i$, (где $a = M(X)$); или $D(X)=M\left(X^2\right)-(M(X))^2$.

Доказательство. Используя то, что $M(X)$ – постоянная величина, и свойства математического ожидания, преобразуем формулу (5.10) к виду:
$D(X) = M(X – M(X))^2 = M(X^2 - 2X\cdot M(X) + M^2(X)) = M(X^2) – 2M(X)\cdot M(X) + M^2(X) = M(X^2) – 2M^2(X) + M^2(X) = M(X^2) – M^2(X)$, что и требовалось доказать.

Свойства дисперсии случайной величины

1. Дисперсия постоянной величины есть нулю: $D(C)=0$.
2. Постоянный множитель можно выносить за знак дисперсии, возводя его в квадрат: $D(CX)=C^2\cdot D(X)$.
3. Дисперсия суммы двух ***независимых*** случайных величин равна сумме дисперсий этих величин:
$$D(X+Y)=D(X)+D(Y).$$

4. Дисперсия разности двух ***независимых*** случайных величин равна сумме их дисперсий:

$$D(X-Y)=D(X)+D(Y).$$

Средним квадратическим отклонением σ случайной величины X называется арифметическое значение корня квадратного из ее дисперсии: $\sigma=\sqrt{D(X)}$.

Функция распределения случайной величины

До сих пор в качестве исчерпывающего описания ДСВ мы рассматривали ее закон распределения, представляющий собой ряд распределения. Однако такое описание случайной величины X не является единственным, а главное, не универсально. Например, оно не применимо для непрерывной случайной величины (НСВ), т.к. во-первых, нельзя перечислить все бесконечное несчетное множество ее значений; во-вторых, как мы увидим дальше, вероятности каждого отдельно взятого значения НСВ равны нулю.

Как дискретная так и непрерывная случайная величина может быть задана ***функцией распределения.***

Функцией распределения случайной величины X называется функция $F(x)$, выражающая для каждого x вероятность того, что случайная величина X примет значение меньшее x:

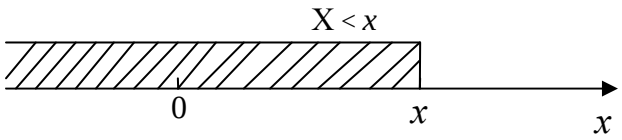
$$F(x)=P(X<x)$$

Эта функция каждому действительному числу x (текущая переменная) ставит в соответствие другое действительное число $F(x)$

$$x \to F(x)=P(X<x)$$

Если значения случайной величины – точки на числовой оси, то ***геометрически*** функция распределения

интерпретируется как вероятность того, что случайная величина X попадает левее заданной точки x:



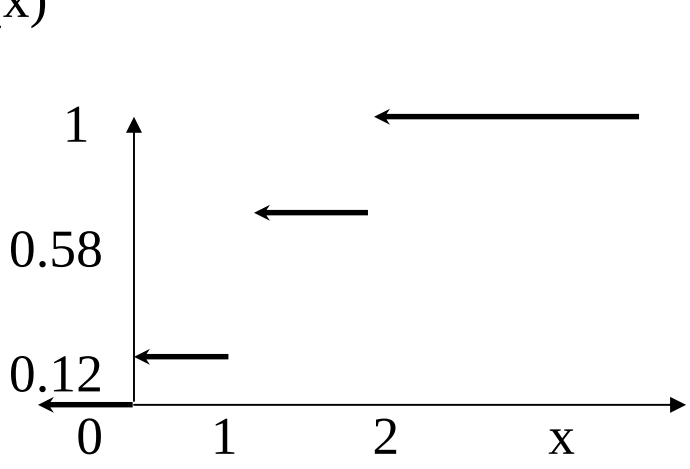
Пример 4. Найдем $F(x)$ для примера 1, где ряд распределения имеет вид:

x_i	0	1	2
p_i	0,12	0,46	0,42

Тогда, например для $1 < x \le 2$ имеем $F(x) = P(X = 0) + P(x = 1) = 0,12 + 0,46 = 0,58$. Следовательно,

$$F(x) = \begin{cases} 0, & x \le 0 \\ 0,12, & 0 < x \le 1 \\ 0,12 + 0,46 = 0,58, & 1 < x \le 2 \\ 0,58 + 0,42 = 1, & x > 2 \end{cases}$$

Соответственно график функции распределения имеет ступенчатый вид:



$F(x)$ обладает свойствами:

1. Функция распределения случайной величины есть неотрицательная функция, заключенная между нулем и единицей: $0 \le F(x) \le 1$.

2. Функция распределения есть неубывающая функция на всей числовой оси.

Это следует из того, что

$F(x_2)=P(X<x_2)=P((X<x_1)+(x_1\le X<x_2))=$
$P(X<x_1)+P(x_1\le X<x_2)=F(x_1)+P(x_1\le X<x_2).$

Т.к. вероятность $P(x_1 \le X < x_2) \ge 0$, то из последнего равенства вытекает $F(x_2) \ge F(x_1)$.

3. Функция $F(x)$ в точке x_0 непрерывна слева, т.е.

$\lim\limits_{x \to x_0 - 0} F(x) = F(x_0)$ или $F(x_0 - 0) = F(x_0)$

4. На минус бесконечности функция распределения равна нулю, на плюс бесконечности равна 1, т.е.

$$F(-\infty) = \lim_{x \to -\infty} F(x) = 0, \qquad F(+\infty) = \lim_{x \to +\infty} F(x) = 1.$$

5. Вероятность попадания случайной величины в интервал $[x_1, x_2)$ (включая x_1) равна приращению ее функции распределения на этом интервале, т.е.

$$P(x_1 \le X < x_2) = F(x_2) - F(x_1).$$

6. $P(X \ge x) = 1 - F(x)$

7. Функция распределения для дискретной случайной величины имеет вид

$$F(x) = \sum_{i: x_i < x} P(X = x_i)$$

(т.е. для заданного x суммирование осуществляется по всем тем индексам i, для которых верно неравенство $x_i < x$).

Непрерывные случайные величины. Плотность вероятности

Случайная величина X называется ***непрерывной***, если ее функция распределения $F(x)$ непрерывна в любой точке и дифференцируема всюду, кроме, быть может, отдельных точек.

Примеры непрерывных случайных величин: диаметр детали, которую токарь обтачивает до заданного размера, рост человека, дальность полета снаряда и др.

Теорема. Вероятность любого отдельно взятого значения непрерывной случайной величины равна нулю:

$$P(X = x_1) = 0.$$

Доказательство. Используя свойство непрерывности функции $F(x)$, имеем:

$$P(X = x) = \lim_{\Delta x \to 0} P(x \le X < x + \Delta x) =$$

$$= \lim_{\Delta x \to 0} [F(x + \Delta x) - F(x) = F(x) - F(x) = 0$$

Следствие.

Если X – непрерывная случайная величина, то вероятность попадания случайной величины в интервал (x_1, x_2) не зависит от того, является этот интервал открытым или закрытым, т.е.

$$P(x_1 < X < x_2) = P(x_1 \le X < x_2) =$$

$$P(x_1 < X \le x_2) = P(x_1 \le X \le x_2)$$

Для непрерывной случайной величины

$$\boxed{P(x_1 < X < x_2) = F(x_2) - F(x_1)}.$$

Задание непрерывной случайной величины с помощью функции распределения не является единственно возможным. Для непрерывных СВ существует неотрицательная функция $p(x)$, удовлетворяющая при любых x равенству:

$$p(x) = \lim_{\Delta x \to 0} \frac{P(x < X < x + \Delta x)}{\Delta x} =$$

$$\lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} = F'(x)$$

Плотностью вероятности *(плотностью распределения* или просто *плотностью) p (x)* непрерывной случайной величины *X* называется производная ее функции распределения:

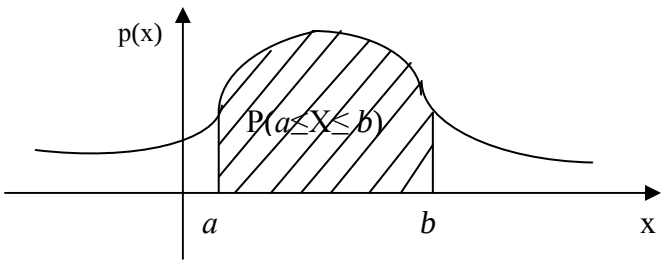
$$p(x) = F'(x).$$

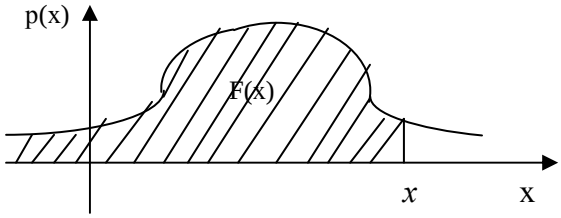
Плотность вероятности *p(x)*, как и функция распределения *F(x)*, является одной из форм закона распределения, но в отличие от функции распределения она существует только для *непрерывных* случайных величин.

Свойства плотности вероятности непрерывной случайной величины:

1. $p(x) \geq 0$;

2. $P(a \leq X \leq b) = \int_a^b p(x)\,dx$, (рис. 1);





3. $F(x)=\int_{-\infty}^{x} p(x)\,dx$, (рис. 2);

4. $\int_{-\infty}^{+\infty} p(x)\,dx=1.$

Геометрически свойства плотности вероятности означают, что ее график – кривая распределения – лежит не ниже оси абсцисс, и полная площадь фигуры, ограниченной кривой распределения и осью абсцисс, равна единице.

Числовые характеристики непрерывных случайных величин

Математическое ожидание непрерывной случайной величины X, возможные значения которой принадлежат всей оси Ox, есть величина $M(X)=\int_{-\infty}^{+\infty} xp(x)\,dx$

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

Предполагается, что интеграл сходится абсолютно. В частности, если все возможные значения принадлежат интервалу $(a;b)$, то $M(X)=\int_{a}^{b} xp(x)\,dx$.

Дисперсия непрерывной случайной величины X, возможные значения которые принадлежат всей оси Ox, определяется равенством $D(X)=\int\limits_{-\infty}^{+\infty}[x-M(X)]^2\,p(x)\,dx$, или равносильным равенством

$$D(X)=\int\limits_{-\infty}^{+\infty}x^2\,p(x)\,dx-[M(X)]^2.$$

В частности, если все возможные значения X принадлежат $(a;b)$, то $D(X)=\int\limits_{a}^{b}[x-M(X)]^2\,p(x)\,dx$, или

$$\boxed{D(X)=\int\limits_{a}^{b}x^2\,p(x)\,dx-[M(X)]^2}.$$

Все свойства математического ожидания и дисперсии для дискретных случайных величин справедливы и для непрерывных величин.

Среднее квадратическое отклонение непрерывной случайной величины есть $\boxed{\sigma(X)=\sqrt{D(X)}}$.

Интерпретация математического ожидания и дисперсии в финансовом анализе

Пусть, например, известно распределение доходности X некоторого актива (например, акции), т.е. известны значения доходности x_i и соответствующие им вероятности p_i за рассматриваемый промежуток времени. Тогда, очевидно, математическое ожидание $M(X)$ выражает среднюю (прогнозную) доходность актива, а дисперсия $D(X)$ или среднее квадратическое отклонение $\sigma(X)$– меру отклонения, колеблемости

доходности от ожидаемого среднего значения, т.е. риск данного актива.

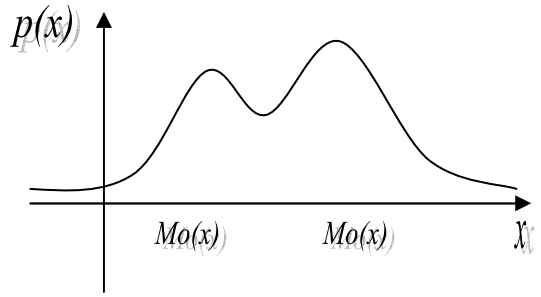
Замечание. Обратим внимание на то, что сама величина X – случайная, а ее числовые характеристики (математическое ожидание, дисперсия, среднее квадратическое отклонение и др.), призванные в сжатой форме выразить наиболее существенные черты распределения, есть величины *неслучайные* (постоянные).

Мода и медиана. Квантили. Моменты СВ

Кроме математического ожидания и дисперсии в теории вероятностей применяется еще ряд числовых характеристик, отражающих те или иные особенности распределения.

Модой $M_0(X)$ непрерывной случайной величины X называется ее наиболее вероятное значение x (где плотность вероятности $p(x)$ достигает максимума).

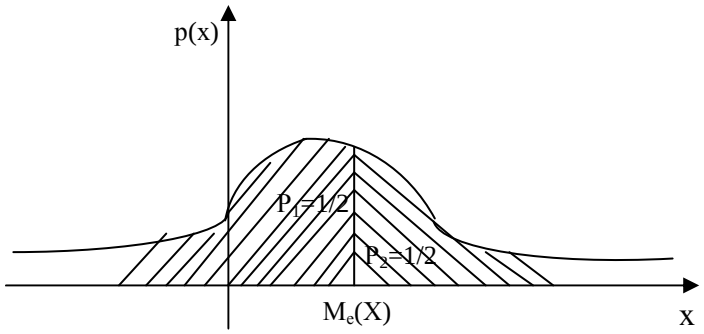
Если вероятность p_i или плотность $p(x)$ достигает максимума не в одной, а в нескольких точках, распределение называется ***полимодальным***



Медианой $M_e(X)$ непрерывной случайной величины X называется такое ее значение, для которого

$$\boxed{P\left(X < M_e(X)\right) = P\left(X > M_e(X)\right) = \frac{1}{2}}.$$

Геометрически, вертикальная прямая $x = M_e(X)$, проходящая через точку с абсциссой, равной $M_e(X)$, делит площадь фигуры под кривой распределения на две равные части (рис.). Очевидно, что $F\left(M_e(X)\right) = 1/2$.



Квантилем уровня q (q***-квантилем***) называется такое значение x_q случайной величины, при котором функция ее распределения принимает значение, равное q, т.е. $\boxed{F(x_q) = P(x < x_q) = q}$

Введенное выше понятие медианы СВ есть квантиль уровня 0,5. Квантили $x_{0,25}$ и $x_{0,75}$ получили название соответственно *верхнего* и *нижнего квартилей*. В литературе также встречаются термины: *децили*, под которыми понимаются квантили $x_{0,1}, x_{0,2}, ..., x_{0,9}$ и *процентили* – квантили $x_{0,01}, x_{0,02}, ..., x_{0,99}$.

Среди числовых характеристик СВ особое значение имеют ***моменты***–начальные и центральные.

Начальный теоретический момент порядка *k* непрерывной случайной величины X определяется равенством $\boxed{v_k = \int\limits_{-\infty}^{+\infty} x^k \, p(x)\, dx}$.

Центральный теоретический момент порядка *k* непрерывной случайной величины X определяется равенством $\boxed{\mu_k = \int_{-\infty}^{+\infty} [x - M(X)]^k \, p(x)\, dx}$.

Если все возможные значения X принадлежат интервалу $(a;b)$, то $v_k = \int_a^b x^k p(x)\, dx$, $\mu_k = \int_a^b [x - M(X)]^k \, p(x)\, dx$.

Очевидно, что

$$v_0 = 1;\ \mu_0 = 1;\ v_1 = M(X);\ \mu_1 = 0;\ \mu_2 = D(X).$$

Центральные моменты выражаются через начальные моменты по формулам:

$\mu_2 = v_2 - v_1^2, \mu_3 = v_3 - 3v_1v_2 + 2v_1^3,$

$\mu_4 = v_4 - 4v_1v_3 + 6v_1^2v_2 - 3v_1^4.$

Например,

$$\mu_3 = M(X - M(X))^3 = M(X^3 - 3X^2M(X) + 3XM(X)^2 - M(X)^3) =$$

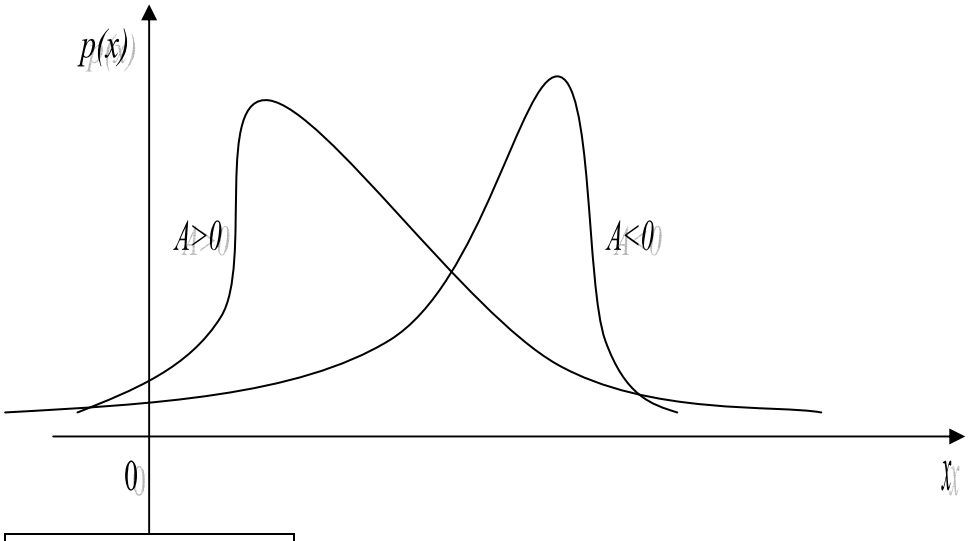
$$= M(X^3 - 3X^2v_1 + 3Xv_1^2 - v_1^3) = M(X^3) - 3M(X^2)v_1 + 3M(X)v_1^2 - v_1^3 =$$

$$= v_3 - 3v_1v_2 + 3v_1^3 - v_1^3 = v_3 - 3v_1v_2 + 2v_1^3$$

Математическое ожидание *M (X)*, или первый начальный момент, характеризует среднее значение распределения случайной величины *X*; второй центральный момент или дисперсия $D(X)$ – степень рассеяния распределения *X* относительно *M (X)*.

Третий центральный момент служит для характеристики ***ассиметрии (скошенности)*** распределения. Величина $A = \frac{\mu_3}{\sigma^3}$ называется **коэффициентом асимметрии** случайной величины.

$A = 0$, если распределение симметрично относительно математического ожидания.

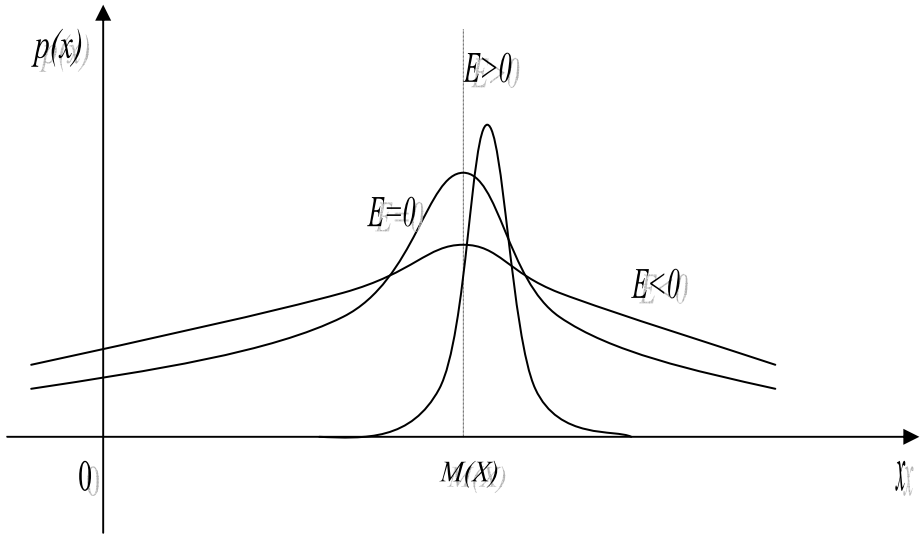


Четвертый центральный момент характеризует крутость распределения.

Эксцессом случайной величины называется число

$$E = \frac{\mu_4}{\sigma^4} - 3.$$

Число 3 вычитается из соотношения $\frac{\mu_4}{\sigma^4}$, т.к. для наиболее часто встречающегося нормального распределения, о котором речь пойдет ниже, отношение $\frac{\mu_4}{\sigma^4} = 3$. Кривые, более островершинные, чем нормальная, обладают положительным эксцессом, более плосковершинные – отрицательным эксцессом.


p(x)
E>0
E=0
E<0
0
M(X)
x

Описаны основные законы распределения дискретных и непрерывных случайных величин, которые часто применяются при математическом моделировании социально-экономических явлений.

Биномиальный закон распределения

Если вероятность появления события A в каждом испытании постоянна и равна p, то ***число появлений события A – дискретная случайная величина*** X, принимающая значения $0,1,2,\ldots,m,\ldots,n$ с вероятностями

$$P_n(m)=C_n^m p^m q^{n-m},\quad 0<p<1,\ q=1-p,\ m=0,1,\ldots,n.$$

Другими словами, ряд распределения биномиального закона имеет вид:

x_i	0	1	2	…	m	…	n
p_i	q^n	$C_n^1 p^1 q^{n-1}$	$C_n^2 p^2 q^{n-2}$	…	$C_n^m p^m q^{n-m}$	…	p^n

Очевидно, что $\sum_{i=0}^{n} p_i = 1$ так как это есть сумма всех членов разложения бинома Ньютона (отсюда и название закона – ***биномиальный***):

$$C_n^0 q^n + C_n^1 p^1 q^{n-1} \ldots + C_n^m p^m q^{n-m} + \ldots + C_n^{n-1} p^{n-1} q^1 + C_n^n p^n =$$

$$(q+p)^n = 1^n = 1.$$

Математическое ожидание и дисперсия случайной величины X, распределенной по биномиальному закону, вычисляются соответственно по формулам:

$$M(X)=np,\ D(X)=npq$$

Доказательство. СВ X –число m наступлений события A в n независимых испытаниях–можно представить в виде суммы n независимых величин

$X = X_1 + X_2 + ... + X_n = \sum_{k=1}^{n} X_k$, каждая из которых имеет один и тот же закон распределения:

x_i	0	1
p_i	q	p

Случайная величина X_k, которую называют индикатором события A, выражает число наступлений события A в k-м испытании $k = 1, 2, ..., n$, т.е. при наступлении события A $X_k = 1$ с вероятностью p, при ненаступлении A $X_k = 0$ с вероятностью q. Найдем числовые характеристики индикатора события A по формулам

$$M(X_k) = x_1 p_1 + x_2 p_2 = 1 \cdot p + 0 \cdot q = p$$

$$D(X_k) = x_1^2 p_1 + x_2^2 p_2 - (M(X_k))^2 = 1^2 \cdot p + 0^2 \cdot q - p^2 = p(1-p) = pq$$

Таким образом, математическое ожидание и дисперсия рассматриваемой СВ X:

$$M(X) = M(X_1 + X_2 + ... + X_n) = \underbrace{p + p + ... + p}_{n\ раз} = np$$

$$D(X) = D(X_1 + X_2 + ... + X_n) = \underbrace{pq + pq + ... + pq}_{n\ раз} = npq.$$

Отметим, что при нахождении дисперсии суммы СВ учтена их независимость. Теорема доказана.

Распределение Пуассона

Дискретная величина X имеет закон распределения Пуассона, если она принимает значения $0, 1, 2, \ldots, m, \ldots$ (счетное множество значений) с вероятностями

$$P_n(X = m) = \frac{\lambda^m e^{-\lambda}}{m!},$$

где λ – некоторая положительная величина, называемая *параметром* закона Пуассона.

Ряд распределения закона Пуассона имеет вид:

x_i	0	1	2	…	m	…
p_i	$e^{-\lambda}$	$\lambda e^{-\lambda}$	$\frac{\lambda^2 e^{-\lambda}}{2!}$	…	$\frac{\lambda^m e^{-\lambda}}{m!}$	…

Определение закона Пуассона корректно, так как сумма всех вероятностей равна 1:

$$\sum_{m=0}^{\infty} p(X = m) = e^{-\lambda} + \lambda e^{-\lambda} + \frac{\lambda^2 e^{-\lambda}}{2!} + \frac{\lambda^m e^{-\lambda}}{m!} + ... =$$

$$= e^{-\lambda} \sum_{m=0}^{\infty} \frac{\lambda^m}{m!} = e^{-\lambda} \cdot e^{\lambda} = 1$$

(разложение в ряд Тейлора функции e^x при $x = \lambda$).

Математическое ожидание и дисперсия случайной величины X, распределенной по закону Пуассона, вычисляются соответственно по формулам:

$$M(X) = \lambda \qquad D(X) = \lambda$$

Доказательство.

$$M(X) = \sum_{i=1}^{\infty} x_i p_i = \sum_{m=1}^{\infty} m \frac{\lambda^m e^{-\lambda}}{m!} = \sum_{m=1}^{\infty} \frac{\lambda^m e^{-\lambda}}{(m-1)!} = e^{-\lambda} \lambda \sum_{m=1}^{\infty} \frac{\lambda^{m-1}}{(m-1)!} = e^{-\lambda} \cdot \lambda \cdot e$$

Дисперсию найдем по формуле (5.18). Для этого найдем

$$M(X^2)=\sum_{i=1}^{n}x_i^{\,2}p_i=\sum_{m=1}^{\infty}m^2\frac{\lambda^m e^{-\lambda}}{m!}=\sum_{m=1}^{\infty}m\frac{\lambda^m e^{-\lambda}}{(m-1)!}=e^{-\lambda}\sum_{m=1}^{\infty}\frac{(m-1+1)\lambda^m}{(m-1)!}$$

$$=e^{-\lambda}\lambda^2\sum_{m=2}^{\infty}\frac{\lambda^{m-2}}{(m-2)!}+e^{-\lambda}\lambda\sum_{m=1}^{\infty}\frac{\lambda^{m-1}}{(m-1)!}=e^{-\lambda}\lambda^2e^{\lambda}+e^{-\lambda}\lambda e^{\lambda}=\lambda^2+\lambda$$

Теперь $D(X)=M(X^2)-(M(X))^2=\lambda^2+\lambda-\lambda^2=\lambda$, что и требовалось доказать.

Сумма двух независимых случайных величин, подчиняющихся распределению Пуассона с параметрами λ_1 и λ_2, также имеет распределение Пуассона с параметром $\lambda_1+\lambda_2$.

Замечание. ***Ф***ормула Пуассона выражает биномиальное распределение при большом числе опытов и малой вероятности события. Поэтому закон Пуассона часто называют *законом редких явлений*.

В начале прошлого столетия в связи с задачами биологии и телефонной связи возникла простая, но весьма полезная схема, получившая наименование процессов гибели и размножения. Например, закону Пуассона подчиняется число α-частиц, достигающих в течение времени t некоторого участка пространства, число клеток с измененными под действием рентгеновского излучения хромосомами, число ошибочных телефонных вызовов в течение суток и т.д.

УПР

Геометрическое распределение

ДСВ X, принимающую только целые положительные значения (1, 2,…, m,…), последовательность которых

бесконечна, но счетна, имеет геометрическое распределение, если вероятность того, что она примет значение m, выражается формулой:

$$P(X = m) = pq^{m-1}$$

Ряд распределения геометрического закона имеет вид:

x_i	1	2	…	m	…
p_i	p	pq	…	pq^{m-1}	…

Определение геометрического закона корректно, так как

$$\sum_{i=1}^{\infty} p_i = p + pq + pq^2 + \ldots + pq^{m-1} + \ldots = \frac{p}{1-q} = \frac{p}{p} = 1,$$

Здесь использована формула $S = \frac{b_1}{1-q}$ – суммы бесконечной убывающей «геометрической прогрессии».

Геометрическое распределение есть испытание по схеме Бернулли ***до первого положительного исхода***

Математическое ожидание и дисперсия случайной величины X, распределенной по геометрическому закону, вычисляются соответственно по формулам:

$$M(X) = \frac{1}{p} \qquad D(X) = \frac{q}{p^2}$$

Пример. Определить математическое ожидание и дисперсию случайной величины X – числа бросков монеты до первого появления герба. Эта величина может принимать бесконечное число значений (множество возможных значений есть множество натуральных чисел).

Решение. Ряд ее распределения имеет вид:

x_i	1	2	…	n	…
p_i	$1/2$	$1/2^2$	…	$1/2^n$	…

Тогда $M(X)=\dfrac{1}{1/2}=2.$ $D(X)=\dfrac{1/2}{1/4}=2.$

Гипергеометрическое распределение

Пусть имеется N элементов, из которых M элементов обладают некоторым признаком A. Извлекаются случайным образом без возвращения n элементов.

X – дискретная случайная величина, означающая число элементов обладающих признаком A, среди отобранных n элементов.

Вероятность, что $\{X=m\}$, где $m=0,1,2,...,\min\{n,M\}$, определяется по формуле

$$P(X=m)=\frac{C_M^m C_{N-M}^{n-m}}{C_N^n}.$$

По-другому, вероятность $P(X=m)$ означает вероятность выбрать m объектов, обладающих заданным свойством из множества n объектов, случайно извлеченных (без возврата) из совокупности N объектов, среди которых M объектов имеют заданное свойство ($M \le N$).

Математическое ожидание и дисперсия случайной величины, распределенной по гипергеометрическому закону, определяются формулами:

$$M(X) = n\frac{M}{N},$$

$$D(X) = n\frac{M}{N-1}\left(1-\frac{M}{N}\right)\left(1-\frac{n}{N}\right).$$

Пример. В лотерее «Спортлото 6 из 36» денежные призы получают участники, угадавшие 3, 4, 5 и 6 видов спорта из отобранных случайно 6 видов из 36 (размер приза увеличивается с увеличением числа угаданных видов спорта). Определить математическое ожидание и дисперсию случайной величины X – числа угаданных видов спорта среди случайно отобранных шести. Какова вероятность получения денежного приза?

Решение. Число угаданных видов спорта в лотерее «6 из 36» есть случайная величина, имеющая гипергеометрическое распределение с параметрами

$$n = 6,\ M = 6,\ N = 36.$$

Ряд распределения СВ X имеет вид (см. формулы)

x_i	0	1	2	3	4	5	6
p_i	$\frac{28275}{92752}$	$\frac{40716}{92752}$	$\frac{19575}{92752}$	$\frac{1450}{34782}$	$\frac{2175}{649264}$	$\frac{15}{162316}$	$\frac{1}{1947792}$

Вероятность получения денежного приза

$$P(3 \le X \le 6) = \sum_{i=3}^{6} P(X = i) = \frac{1450}{34782} + \frac{2175}{649264} + \frac{15}{162316} + \frac{1}{1947792}$$

$$\approx 0,043$$

Далее найдем:

$$M(X)=n\frac{M}{N}=\frac{6\cdot 6}{36}=1$$

$$D(X)=n\frac{M}{N-1}\left(1-\frac{M}{N}\right)\left(1-\frac{n}{N}\right)=\frac{36}{35}\cdot\frac{30}{36}\cdot\frac{30}{36}=\frac{5}{7}.$$

Таким образом, среднее число угаданных видов спорта из 6 всего 1, а вероятность выигрыша только около 4%.

Равномерный закон распределения

Непрерывная случайная величина X имеет **равномерный закон** распределения на отрезке $[a,b]$, если ее плотность вероятности $p(x)$ постоянна на этом отрезке и равна нулю вне его, т.е.

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

Функция распределения случайной величины X, распределенной по равномерному закону, есть

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

Действительно,

при $x\le a$

$$F(x)=\int_{-\infty}^{x}p(\tau)d\tau=\int_{-\infty}^{x}0\,d\tau=0$$

при $a<x\le b$

$$F(x)=\int_{-\infty}^{x}p(\tau)d\tau=\int_{-\infty}^{a}0\,d\tau+\int_{a}^{x}\frac{1}{b-a}d\tau=\frac{1}{b-a}\tau\Big|_{a}^{x}=\frac{x-a}{b-a}$$

при $x > b$

$$F(x) = \int_{-\infty}^{x} p(\tau)d\tau = \int_{-\infty}^{a} 0d\tau + \int_{a}^{b} \frac{1}{b-a}d\tau + \int_{b}^{x} 0d\tau = \frac{1}{b-a}\tau\Big|_{a}^{b} = \frac{b-a}{b-a} = 1$$

Математическое ожидание $\boxed{M(X) = \frac{a+b}{2}}$,

Действительно,

$$M(X) = \int_{-\infty}^{+\infty} \tau p(\tau)d\tau = \int_{a}^{b} \tau \frac{1}{b-a}d\tau = \frac{1}{b-a}\frac{\tau^2}{2}\Big|_{a}^{b} = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}$$

Дисперсия $\boxed{D(X) = \frac{(b-a)^2}{12}}$,

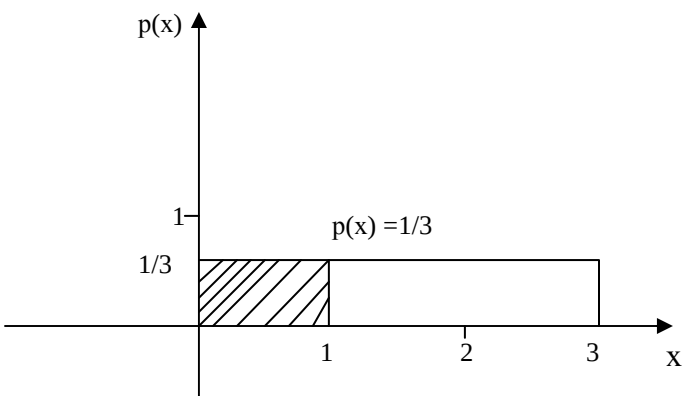
$$D(X) = \int_{-\infty}^{+\infty} x^2 p(x)dx - M^2(X) = \int_{a}^{b} x^2 \frac{1}{b-a}dx - \left(\frac{b+a}{2}\right)^2 = \frac{1}{b-a}\frac{x^3}{3}\Big|_{a}^{b} - \left(\frac{b+a}{2}\right)^2 =$$

$$= \frac{b^3-a^3}{3(b-a)} - \left(\frac{b+a}{2}\right)^2 = \frac{b^2+ab+a^2}{3} - \frac{b^2+2ab+a^2}{4} = \frac{b^2-2ab+a^2}{12} = \frac{(b-a)^2}{12}$$

Средне- квадратическое отклонение $\sigma(X) = \frac{b-a}{2\sqrt{3}}$ **(УПР)**

Пример Поезда метрополитена идут регулярно с интервалом 3 мин. Пассажир выходит на платформу в случайный момент времени. Какова вероятность того, что ждать пассажиру придется не больше минуты. Найти математическое ожидание и среднее квадратическое отклонение случайной величины X – времени ожидания поезда.

Решение. Случайная величина X – время ожидания поезда на временном (в минутах) отрезке $[0, 3]$ имеет



равномерный закон распределения $p(x)=\frac{1}{3}$. Поэтому вероятность того, что пассажиру придется ждать не более минуты, равна $\frac{1}{3}$ от равной единице площади прямоугольника (см. рис.), т.е.

$$P(x\le 1)=\int_0^1 \frac{1}{3}\,dx=\frac{1}{3}x\Big|_0^1=\frac{1}{3}. \qquad M(X)=\frac{0+3}{2}=1{,}5\,(\text{мин.}),$$

$$D(X)=\frac{(3-0)^2}{12}=\frac{3}{4},\ \sigma(X)=\sqrt{D(X)}=\frac{\sqrt{3}}{2}\approx 0{,}86\ (\text{мин.}).$$

Показательный (экспоненциальный) закон распределения

Непрерывная случайная величина X имеет **показательный (экспоненциальный)** закон распределения с параметром λ, если ее плотность вероятности имеет вид: $p(x)=\begin{cases}\lambda e^{-\lambda x} & \text{при } x\ge 0,\\ 0 & \text{при } x<0.\end{cases}$

Функция распределения случайной величины, распределенной по показательному закону, равна

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

Действительно,

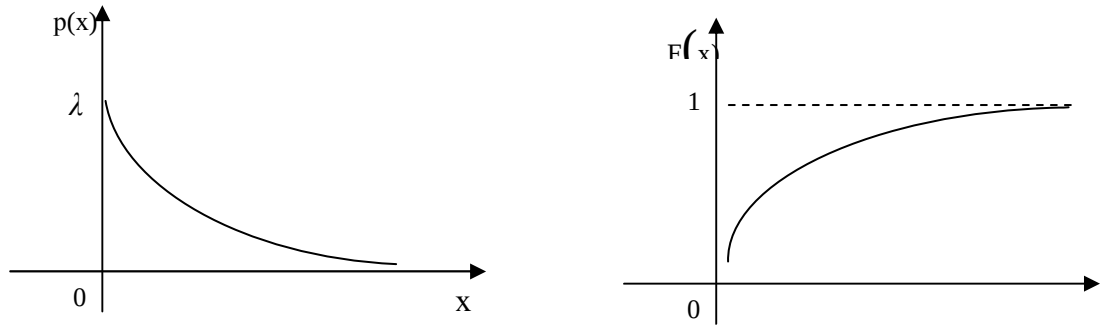
при $x\le 0$ $F(x)=\int_{-\infty}^{x} p(\tau)d\tau=\int_{-\infty}^{x} 0\,d\tau=0$

при $x > 0$

$$F(x) = \int_{-\infty}^{x} p(\tau)d\tau = \int_{-\infty}^{0} 0d\tau + \int_{0}^{x} \lambda e^{-\lambda\tau} d\tau = -\int_{0}^{x} e^{-\lambda\tau} d(-\lambda\tau) =$$

$$= -\int_{0}^{x} d(e^{-\lambda\tau}) = -e^{-\lambda\tau} \Big|_{0}^{x} = -e^{-\lambda x} + 1$$

Кривая плотности $p(x)$ и функции распределения $F(x)$:



Для случайной величины, распределенной по показательному закону

$$M(X) = \sigma(X) = \frac{1}{\lambda}; \quad D(X) = \frac{1}{\lambda^2} \qquad \sigma(X) = \frac{1}{\lambda}$$

Доказательство:

$$M(X) = \int_{-\infty}^{+\infty} xp(x)dx = \int_{-\infty}^{0} x \cdot 0dx + \int_{0}^{+\infty} x(\lambda e^{-\lambda x})dx =$$

$$= -\int_{0}^{+\infty} xd(e^{-\lambda x}) = -\lim_{b\to+\infty} \int_{0}^{b} xd(e^{-\lambda x}) =$$

$$= -\lim_{b\to+\infty}\int_0^b x d(e^{-\lambda x}) = -\lim_{b\to+\infty}\left(xe^{-\lambda x}\Big|_0^b - \int_0^b e^{-\lambda x}dx \right) =$$

$$= -\lim_{b\to+\infty}\left(be^{-b\lambda} - 0e^{-0\lambda} + \frac{1}{\lambda}\int_0^b d(e^{-\lambda x}) \right) =$$

$$= -\lim_{b\to+\infty}\left(be^{-b\lambda} - 0e^{-0\lambda} + \frac{1}{\lambda}e^{-\lambda x}\Big|_0^b \right) = -\lim_{b\to+\infty}\left(be^{-b\lambda} + \frac{1}{\lambda}e^{-b\lambda} - \frac{1}{\lambda}e^{-0\lambda} \right) =$$

$$-\lim_{b\to+\infty}\left(\frac{b}{e^{b\lambda}} + \frac{1}{\lambda e^{b\lambda}} - \frac{1}{\lambda} \right) = \frac{1}{\lambda} - \lim_{b\to+\infty}\frac{b}{e^{b\lambda}} = \frac{1}{\lambda} - \lim_{b\to+\infty}\frac{1}{\lambda e^{b\lambda}} = \frac{1}{\lambda}.$$

$$M(X^2) = \int_{-\infty}^{+\infty} x^2 p(x)dx = ...(Упр)... = \frac{2}{\lambda^2}.$$

$$D(X) = M(X^2) - M^2(X) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

Вероятность попадания в интервал $(a;b)$ непрерывной случайной величины X, распределенной по показательному закону:

$$\boxed{P(a < X < b) = e^{-\lambda a} - e^{-\lambda b}}.$$

Замечание. Показательный закон распределения вероятностей встречается во многих задачах, связанных с простейшим потоком событий. Под ***потоком событий*** понимают последовательность событий, наступающих одно за другим в случайные моменты. Например, поток вызовов на телефонной станции, поток заявок в системе массового обслуживания и др.

Часто длительность времени безотказной работы элемента имеет показательное распределение, функция распределения которого

$$F(t)=P(T<t)=1-e^{-\lambda t} \quad (\lambda>0)$$

определяет ***вероятность отказа*** элемента за время длительностью t. Здесь T – длительность времени безотказной работы элемента, λ – интенсивность отказов (среднее число отказов в единицу времени).

Функция надежности

$$R(t)=e^{-\lambda t}$$

определяет вероятность безотказной работы элемента за время длительностью t.

Пример. Установлено, что время ремонта магнитофонов есть случайная величина X, распределенная по показательному закону. Определить вероятность того, что на ремонт магнитофона потребуется не менее 15 дней, если среднее время ремонта магнитофонов составляет 12 дней. Найти плотность вероятности, функцию распределения и среднее квадратическое отклонение случайной величины X.

Решение. По условию математическое ожидание $M(X)=\frac{1}{\lambda}=12$, откуда параметр $\lambda=\frac{1}{12}$ и тогда плотность вероятности и функция распределения имеют вид: $p(x)=\frac{1}{12}e^{-\frac{1}{12}x}$; $F(x)=1-e^{-\frac{1}{12}x}$ $(x\geq 0)$. Искомую вероятность $P(X\geq 15)$ можно было найти, используя функцию распределения:

$$P(X \geq 15) = 1 - P(X < 15) = 1 - F(15) = 1 - \left(1 - e^{-\frac{15}{12}}\right) = e^{-\frac{15}{12}} = 0,2865$$

Среднее квадратическое отклонение $\sigma(X) = M(X) = 12$ дней.

Пример. Испытывают три элемента, которые работают независимо один от другого. Длительность времени безотказной работы элементов распределена по показательному закону: для первого элемента $F_1(t) = 1 - e^{-0,1t}$; для второго $F_2(t) = 1 - e^{-0,2t}$; для третьего элемента $F_3(t) = 1 - e^{-0,3t}$. Найти вероятности того, что в интервале времени $(0;5)$ ч. откажут:

а) только один элемент;

б) только два элемента;

в) все три элемента.

Решение. Вероятность отказа первого элемента

$$P_1 = F_1(5) = 1 - e^{-0,1 \cdot 5} = 1 - e^{-0,5} = 1 - 0,5957 = 0,4043.$$

Вероятность отказа второго элемента

$$P_2 = F_2(5) = 1 - e^{-0,2 \cdot 5} = 1 - e^{-1} = 1 - 0,3779 = 0,6321.$$

Вероятность отказа третьего элемента

$$P_3 = F_3(5) = 1 - e^{-0,3 \cdot 5} = 1 - e^{-1,5} = 1 - 0,2231 = 0,7769.$$

Искомая вероятность

а) $P = p_1 q_2 q_3 + q_1 p_2 q_3 + q_1 q_2 p_3 = 0,034 + 0,084 + 0,1749 = 0,2929.$

б) $P = p_1 p_2 q_3 + p_1 q_2 p_3 + q_1 p_2 p_3 = 0,057 + 0,1187 + 0,2925 = 0,4682$

в) $P = p_1 p_2 p_3 = 0,1985.$

Нормальный закон распределения

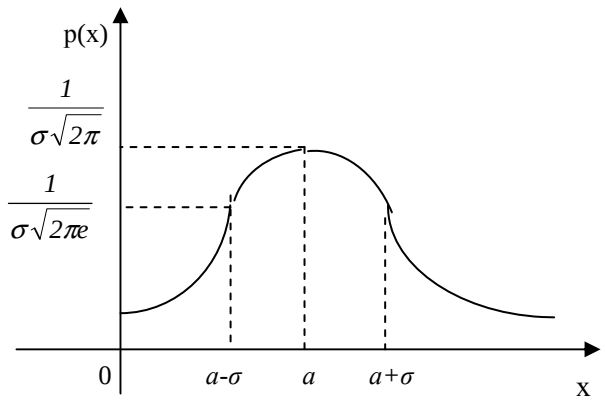
В ТВ и МС важнейшую роль играет так называемое ***нормальное или гауссовское*** распределение. Значимость нормального распределения определяется тем, что оно служит хорошим приближением для большого числа наборов случайных величин, получаемых при наблюдениях и экспериментах. Нормальное распределение почти всегда имеет место, когда наблюдаемые случайные величины формируются под влиянием большого числа случайных факторов, ни один из которых существенно не превосходит остальные (см. далее «закон больших чисел»).

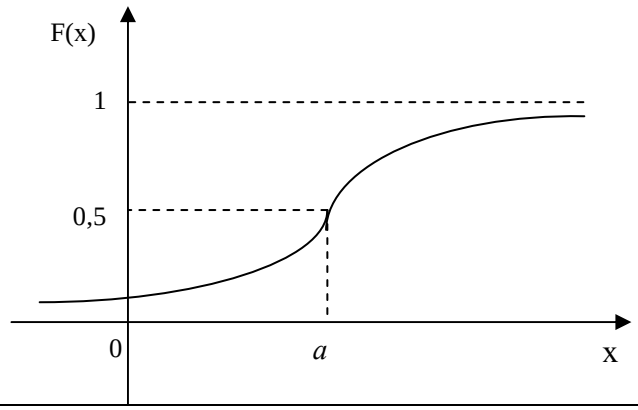
Непрерывная случайная величина X имеет **нормальный закон распределения (закон Гаусса)** с параметрами a и σ^2, если ее плотность вероятности имеет вид:

$$p(x)=\frac{1}{\sigma\sqrt{2\pi}}\cdot e^{-\frac{(x-a)^2}{2\sigma^2}},-\infty<x<+\infty.$$

Кривую нормального закона распределения называют **нормальной** или **гауссовой кривой**.

На рис. приведены ***нормальная кривая p(x)*** с параметрами a и σ^2, т.е. $N\left(a;\sigma^2\right)$ и ***график функции распределения*** случайной величины X, имеющей нормальный закон:





Нормальная кривая симметрична относительно прямой $x = a$, имеет максимум в точке $x = a$, равный $\frac{1}{\sigma\sqrt{2\pi}}$, и две точки перегиба $x = a \pm \sigma$ с ординатой $\frac{1}{\sigma\sqrt{2\pi e}}$.

Действительно, найдем точки экстремума и точки перегиба.

$$p'(x) = -\frac{2(x-a)}{2\sigma^2}\frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-a)^2}{2\sigma^2}}, \quad -\infty < x < +\infty$$

$$p''(x)=-\frac{1}{\sqrt{2\pi}\sigma^3}\cdot e^{-\frac{(x-a)^2}{2\sigma^2}}+\frac{(x-a)^2}{\sqrt{2\pi}\sigma^5}\cdot e^{-\frac{(x-a)^2}{2\sigma^2}}=$$

$$=-\frac{1}{\sqrt{2\pi}\sigma^3}\cdot e^{-\frac{(x-a)^2}{2\sigma^2}}\left(1-\frac{(x-a)^2}{\sigma^2}\right)$$

Т.к. первая производная обращается в 0 при $x=a$ и меняет знак при переходе через эту точку с «+» на «-», то в точке $x=a$ функция принимает максимальное значение, равное $p_{\max}(a)=\frac{1}{\sigma\sqrt{2\pi}}$.

Т.к. вторая производная обращается в 0 при $x=a\pm\sigma$ и меняет знак при переходе через эти точки, то в точках $\left(a+\sigma,\frac{1}{\sigma\sqrt{2\pi e}}\right)$ и $\left(a-\sigma,\frac{1}{\sigma\sqrt{2\pi e}}\right)$ функция меняет направление выпуклости.

Для случайной величины, распределенной по нормальному закону, $M(X)=a,\quad D(X)=\sigma^2$.

Функция распределения случайной величины X, распределенной по нормальному закону, выражается через функцию Лапласа $\Phi(x)$ по формуле

$$F(x)=\frac{1}{2}+\frac{1}{2}\Phi\left(\frac{x-a}{\sigma}\right),\quad \text{где}\quad \Phi(x)=\frac{2}{\sqrt{2\pi}}\int_0^x e^{-t^{2/2}}dt.$$

Вероятность попадания значений нормальной случайной величины X в интервал $[\alpha,\beta]$ определяется формулой

$$P(\alpha\le x\le\beta)=\frac{1}{2}\left[\Phi\left(\frac{\beta-a}{\sigma}\right)-\Phi\left(\frac{\alpha-a}{\sigma}\right)\right].$$

Вероятность того, что отклонение случайной величины X, распределенной по нормальному закону, от математического ожидания a не превысит величину $\varepsilon > 0$ (по абсолютной величине), равна

$$P\left(\left(\left|x-a\right|\right) \le \varepsilon\right) = \Phi\left(\frac{\varepsilon}{\sigma}\right).$$

«Правило трех сигм»: Если случайная величина X имеет нормальный закон распределения с параметрами a и σ^2, т.е. $N\left(a;\sigma^2\right)$, то практически достоверно, что ее значения заключены в интервале $(a-3\sigma;\ a+3\sigma)$:

$$P\left(\left(\left|x-a\right|\right) \le 3\sigma\right) = \Phi(3) = 0{,}9973.$$

Асимметрия нормального распределения $A = 0$;
эксцесс нормального распределения $E = 0$.

Пример. Определить закон распределения случайной величины X, если ее плотность распределения вероятностей задана функцией: $p(x) = \frac{1}{6\sqrt{2\pi}} \cdot e^{-\frac{(x-1)^2}{72}}$.

Найти математическое ожидание, дисперсию и функцию распределения случайной величины X.

Решение. Сравнивая данную функцию $p(x)$ с функцией плотности вероятности для случайной величины, распределенной по нормальному закону, заключаем, что

случайная величина X распределена по нормальному закону с параметрами $a = 1$ и $\sigma = 6$.

Тогда $M(X) = 1,\ \sigma(X) = 6,\ D(X) = 36.$

Функция распределения случайной величины X имеет вид:

$$F(x) = \frac{1}{2} + \frac{1}{2}\Phi\left(\frac{x-1}{6}\right).$$

Пример. Текущая цена акции может быть смоделирована с помощью нормального закона распределения с математическим ожиданием 15 ден. ед. и средним квадратическим отклонением 0,2 ден. ед.

Найти вероятность того, что цена акции:

а) не выше 15,3 ден. ед.;

б) не ниже 15,4 ден. ед.;

в) от 14,9 до 15,3 ден. ед.

г) с помощью правила «трех сигм» найти границы, в которых будет находиться текущая цена акции.

Решение. Так как $a = 15$ и $\sigma = 0{,}2$, то

$$P(X \le 15{,}3) = F(15{,}3) = \frac{1}{2} + \frac{1}{2}\Phi\left(\frac{15{,}3-15}{0{,}2}\right) = \frac{1}{2} + \frac{1}{2}\Phi(1{,}5) =$$

$$= \frac{1}{2} + \frac{1}{2}\cdot 0{,}8664 = 0{,}9332.$$

$$P(X \ge 15{,}4) = 1 - F(15{,}4) = 1 - \left(\frac{1}{2} + \frac{1}{2}\Phi\left(\frac{15{,}4-15}{0{,}2}\right)\right) =$$

$$\frac{1}{2} - \frac{1}{2}\Phi(2) = \frac{1}{2} - \frac{1}{2}\cdot 0{,}9545 = 0{,}0228.$$

$$P(14{,}9 \le x \le 15{,}3) = \frac{1}{2}\left[\Phi\left(\frac{15{,}3-15}{0{,}2}\right) - \Phi\left(\frac{14{,}9-15}{0{,}2}\right)\right] =$$

$$= \frac{1}{2}[\Phi(1{,}5) + \Phi(0{,}5)] = \frac{1}{2}(0{,}8664 + 0{,}3829) = 0{,}6246.$$

По правилу трех сигм $P(|X-15| \le 0{,}6) = 0{,}9973$ и, следовательно, $15 - 0{,}6 \le X \le 15 + 0{,}6$.

Окончательно $\boxed{14{,}4 \le X \le 15{,}6}$.

Пример. Автомат изготавливает детали, которые считаются годными, если отклонение X от контрольного размера по модулю не превышает 0,8 мм. Каково наиболее вероятное число годных деталей из 150, если случайная величина X распределена нормально с $\sigma = 0{,}4$мм?

Решение. Найдем вероятность отклонения при $\sigma = 0{,}4$ и $\varepsilon = 0{,}8$: $P(|X-a| \le 0{,}8) = \Phi\left(\frac{0{,}8}{0{,}4}\right) = \Phi(2) = 0{,}9545.$

Считая приближенно $p = 0{,}95$ и $q = 0{,}05$, в соответствии с формулой $np - q \le m_0 \le np + p$, где m_0 – наивероятнейшее число, находим при $n = 150$:

$$150 \cdot 0{,}95 - 0{,}05 \le m_0 \le 150 \cdot 0{,}95 + 0{,}95.$$

Откуда $\boxed{m_0 = 143}$.

Пример Рост взрослых мужчин является случайной величиной, распределенной по нормальному закону. Пусть математическое ожидание ее равно 175 см, а среднее квадратическое отклонение – 6 см. Определить

вероятность того, что хотя бы один из наудачу выбранных пяти мужчин будет иметь рост от 170 до 180

Решение. Найдем вероятность того, что рост мужчины будет принадлежать интервалу $(170;180)$:

$$P(170 < x < 180) = \frac{1}{2}\left[\Phi\left(\frac{180-175}{6}\right) - \Phi\left(\frac{170-175}{6}\right)\right] =$$

$$= \frac{1}{2}[\Phi(0{,}83) + \Phi(0{,}83)] = \Phi(0{,}83) = 0{,}5935 \approx 0{,}6.$$

Тогда вероятность того, что рост мужчины не будет принадлежать интервалу (170; 180) $q = 1 - 0{,}6 = 0{,}4$.

Вероятность того, что хотя бы один из 5 мужчин будет иметь рост от 170 до 180 см равна:

$$\boxed{P = 1 - q^5 = 1 - 0{,}4^5 = 0{,}9898}$$

Закон больших чисел

Изучение статистических закономерностей позволило установить, что при некоторых условиях суммарное поведение большого количества случайных величин почти утрачивает случайный характер и становится закономерным (иначе говоря, случайные отклонения от некоторого среднего поведения взаимно погашаются). В частности, если влияние на сумму отдельных слагаемых является равномерно малым, закон распределения суммы *приближается к нормальному*. Математическая формулировка этого утверждения дается в группе теорем, называемой ***законом больших чисел***.

Неравенства Маркова и Чебышева

Неравенства Маркова и Чебышева, используемые для доказательства дальнейших теорем, справедливы как для непрерывных, так и для дискретных случайных величин.

Неравенство Маркова. *Если случайная величина X принимает только неотрицательные значения и имеет математическое ожидание, то для любого положительного числа A верны неравенства*

$$P(X > A) \le \frac{M(X)}{A} \quad \text{или} \quad P(X \le A) \ge 1 - \frac{M(X)}{A}$$

Доказательство (для ДСВ). Пусть X задается рядом распределения:

X	x_1	x_2	…	x_n
p	p_1	p_2	…	p_n

Рассмотрим три возможных случая расположения числа A и значений x_i:

1. Пусть $A < x_1$, тогда событие $x > A$ является достоверным и $P(x > A) = 1$. По свойству 1 математического ожидания следует $x_1 \le M(X) \le x_n$, откуда очевидно $A < M(X)$ и $\frac{M(X)}{A} > 1$. Поэтому требуемое неравенство справедливо.
2. Пусть $A > x_n$, тогда событие $x > A$ является невозможным и $P(x > A) = 0$. По свойству 1 математического ожидания следует $x_1 \le M(X) \le x_n$, откуда очевидно $A > M(X)$ и $0 \le \frac{M(X)}{A} < 1$. Поэтому неравенство и в этом случае справедливо.
3. Пусть часть значений $x_1, x_2, ..., x_k$ будут не более числа A, а другая часть $x_{k+1}, x_{k+2}, ..., x_n$ будут больше числа A. Математическое ожидание ДСВ X вычисляется по формуле:

$$M(X) = x_1 p_1 + x_2 p_2 + ... + x_k p_k + x_{k+1} p_{k+1} + ... + x_n p_n$$

Отбрасывая первые k неотрицательных слагаемых, получим

$$M(X) \ge x_{k+1} p_{k+1} + ... + x_n p_n .$$

Заменяя в последнем неравенстве значения $x_{k+1}, x_{k+2}, ..., x_n$ меньшим числом A, получим более сильное неравенство

$$M(X) \ge A\left(p_{k+1} + ... + p_n\right) \text{ или } p_{k+1} + ... + p_n \le \frac{M(X)}{A}.$$

Сумма вероятностей в левой части полученного неравенства представляет собой сумму вероятностей

событий $X = x_{k+1}, ..., X = x_n$, т.е. вероятность события $X > A$. Поэтому $P(x > A) \le \frac{M(X)}{A}$.

Отметим, что события $X > A$ и $X \le A$ противоположные, поэтому заменяя $P(x > A)$ в уже доказанном первом неравенстве на $1 - P(x \le A)$, придем к другой форме неравенства Маркова. *Теорема доказана.*

Пример Сумма всех вкладов в банк составляет 2 млн.руб., а вероятность того, что случайно взятый вклад не превысит 10 тыс. руб., равна 0,6. Оценить число вкладчиков банка.

Решение. Пусть X – размер случайно взятого вклада, а n - число всех вкладов. Тогда из условия следует, что средний размер вклада есть $M(X) = \frac{2000000}{n}$ рублей. Используя неравенство Маркова в виде $P(X \le 10000) \ge 1 - \frac{M(X)}{10000} = 1 - \frac{2000000}{10000 \cdot n}$. Так как $P(X \le 10000) = 0,6$ по условию задачи, то имеем неравенство $1 - \frac{2000\,000}{10\,000 \cdot n} \le 0,6$. Откуда $n \le 500$, т.е. число вкладчиков не более 500.

Пример Оценить вероятность того, что в течение ближайшего дня потребность в воде в населенном пункте превысит 150 000 л, если среднесуточная потребность в ней составляет 50 000 л.

Решение. Используя неравенство Маркова в виде $P(X > A) \le \frac{M(X)}{A}$, получим $P(X > 150\,000) \le \frac{50\,000}{150\,000} = \frac{1}{3}$.

Пример Среднее число солнечных дней в году для данной местности равно 90. Оценить вероятность того, что в течение года в этой местности будет не более 240 солнечных дней

Решение. Согласно неравенству $P(X \le A) \ge 1 - \frac{M(X)}{A}$, имеем $P(X \le 240) \ge 1 - \frac{90}{240} = 1 - 0{,}375 = 0{,}625$.

Неравенство Чебышева. *Для любой случайной величины X, имеющей математическое ожидание и дисперсию, справедливы неравенства*

$$P(|X - M(X)| > \varepsilon) \le \frac{D(X)}{\varepsilon^2},$$

$$P(|X - M(X)| \le \varepsilon) \ge 1 - \frac{D(X)}{\varepsilon^2}, \quad \text{где } \varepsilon > 0.$$

Доказательство. Применим неравенство Маркова к случайной величине $\tilde{X} = (X - M(X))^2$, выбрав в качестве положительного числа $A = \varepsilon^2$:

$$P((X - M(X))^2 > \varepsilon^2) \le \frac{M(X - M(X))^2}{\varepsilon^2} \qquad (*)$$

Т.к. неравенство $(X - M(X))^2 > \varepsilon^2$ равносильно неравенству $|X - M(X)| > \varepsilon$, $M(X - M(X))^2 = D(X)$, то из неравенства (*) получаем требуемое неравенство. Учитывая, что события $|X - M(X)| > \varepsilon$ и $|X - M(X)| \le \varepsilon$ противоположные, из первого неравенства Чебышева получаем другое его представление.

Пример Оценить вероятность того, что отклонение любой случайной величины от ее математического

ожидания по абсолютной величине будет не более трех средних квадратических отклонений.

Решение Воспользуемся неравенством Чебышева, учитывая, что $\varepsilon = 3\sigma$, $D(X) = \sigma^2$:

$$P\left(\left|X - M(X)\right| < 3\sigma\right) \geq 1 - \frac{\sigma^2}{9\sigma^2} = \frac{8}{9} = 0{,}889.$$

(сравни с правилом «3 сигм» для нормального закона)

Пример Среднесуточное потребление электроэнергии в населенном пункте равно 20 000 квт-ч, а среднеквадратичное отклонение – 200 квт-ч. Какого потребления электроэнергии в этом населенном пункте можно ожидать в ближайшие сутки с вероятностью, не меньшей 0,96?

Решение Воспользуемся неравенством Чебышева $P\left(\left|X - M(X)\right| < \varepsilon\right) \geq 1 - \frac{D(X)}{\varepsilon^2}$. Подставим в правую часть неравенства вместо $D(X)$ величину $200^2 = 40\,000$, сделаем ее большей или равной 0,96:

$$1 - \frac{40\,000}{\varepsilon^2} \geq 0{,}96 \quad \Leftrightarrow \quad \frac{40\,000}{\varepsilon^2} \leq 0{,}04 \quad \Leftrightarrow \quad \varepsilon^2 \geq \frac{40\,000}{0{,}04}, \quad \varepsilon \geq 1000$$

Следовательно, в этом населенном пункте можно ожидать с вероятностью не меньшей 0,96 потребление электроэнергии $20\,000 \pm 1000$, т.е.

$X \in [19\,000; 21\,000]$.

Теоремы Чебышева и Бернулли

Теорема Чебышева. *Если X_1, X_2,…, X_n – попарно независимые случайные величины, у каждой из которой есть математическое ожидание и дисперсия, причем*

дисперсии ограничены одной и той же постоянной, т.е. $D(X_i) \le C$, то при неограниченном увеличении числа n и для сколь угодно малого числа ε имеет место равенство:

$$\lim_{n\to\infty} P\left(\left|\frac{X_1+X_2+\ldots+X_n}{n}-\frac{M(X_1)+M(X_2)+\ldots+M(X_n)}{n}\right|<\varepsilon\right)=1.$$

Доказательство. Рассмотрим новую случайную величину $\tilde{X}=\dfrac{X_1+X_2+\ldots+X_n}{n}$ и найдем ее математическое ожидание. Используя свойства математического ожидания, получим, что $M\left(\dfrac{X_1+X_2+\ldots+X_n}{n}\right)=\dfrac{M(X_1)+M(X_2)+\ldots+M(X_n)}{n}$.

Применим к $\tilde{X}$ неравенство Чебышева:

$$P\left(\left|\frac{X_1+X_2+\ldots+X_n}{n}-\frac{M(X_1)+M(X_2)+\ldots+M(X_n)}{n}\right|<\varepsilon\right)\ge$$

$$\ge 1-\frac{D\left(\dfrac{X_1+X_2+\ldots+X_n}{n}\right)}{\varepsilon^2}.$$

Так как рассматриваемые случайные величины независимы, то, учитывая условие теоремы, имеем:

$$D\left(\frac{X_1+X_2+\ldots+X_n}{n}\right)=\frac{D(X_1)+D(X_2)+\ldots+D(X_n)}{n^2}\le\frac{Cn}{n^2}=\frac{C}{n}.$$

Используя этот результат, представим предыдущее неравенство в виде:

$$P\left(\left|\frac{X_1+X_2+\ldots+X_n}{n}-\frac{M(X_1)+M(X_2)+\ldots+M(X_n)}{n}\right|<\varepsilon\right)\ge 1-\frac{C}{n\varepsilon^2}.$$

Перейдем к пределу при $n\to\infty$:

$$\lim_{n\to\infty} P\left(\left|\frac{X_1+X_2+...+X_n}{n}-\frac{M(X_1)+M(X_2)+...+M(X_n)}{n}\right|<\varepsilon\right)\geq 1$$

Поскольку вероятность не может быть больше 1, можно утверждать, что

$$\lim_{n\to\infty} P(\left|\frac{X_1+X_2+...+X_n}{n}-\frac{M(X_1)+M(X_2)+...+M(X_n)}{n}\right|<\varepsilon)=1.$$

Теорема доказана.

Замечание. Формулу в теореме можно записать в виде:

$$\boxed{\frac{\sum_{i=1}^{n} X_i}{n} \underset{n\to\infty}{\xrightarrow{P}} \frac{\sum_{i=1}^{n} M(X_i)}{n}}.$$

Эта формула отражает тот факт, что при выполнении условий теоремы Чебышева, средняя арифметическая случайных величин *сходится по вероятности* к средней арифметической их математических ожиданий.

В отличие от записи $\frac{\sum_{i=1}^{n} X_i}{n} \underset{n\to\infty}{\to} \frac{\sum_{i=1}^{n} M(X_i)}{n}$, которая обозначает, что начиная с некоторого n для сколь угодно малого числа ε неравенство $\left|\frac{X_1+X_2+...+X_n}{n}-\frac{M(X_1)+M(X_2)+...+M(X_n)}{n}\right|<\varepsilon$ будет верно всегда, из теоремы (формулы) не следует такого же категоричного утверждения. Возможно, что в отдельных случаях требуемое неравенство выполняться не будет, однако, с увеличением числа n вероятность такого неравенства стремится к 1, что означает практическую достоверность выполнения этого неравенства при $n\to\infty$.

Следствие. Если $X_1, X_2, \ldots, X_n$ – ***попарно независимые*** случайные величины с ***равномерно ограниченными*** дисперсиями, имеющие ***одинаковое математическое*** ожидание, равное a, то для любого сколь угодно малого $\varepsilon > 0$ вероятность неравенства $\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - a\right| < \varepsilon$ будет как угодно близка к 1, если число случайных величин достаточно велико. Иначе говоря,

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - a\right| < \varepsilon\right) = 1.$$

Вывод: среднее арифметическое достаточно большого числа *случайных величин* принимает значения, близкие к сумме их математических ожиданий, то есть как угодно мало отличается от *неслучайной величины*. Например, если проводится серия измерений какой-либо физической величины, причем:

а) результат каждого измерения не зависит от результатов остальных, то есть все результаты представляют собой попарно независимые случайные величины;

б) измерения производятся без систематических ошибок (их математические ожидания равны между собой и равны истинному значению a измеряемой величины);

в) обеспечена определенная точность измерений, следовательно, дисперсии рассматриваемых случайных величин равномерно ограничены;

то при достаточно большом числе измерений их среднее арифметическое окажется сколь угодно близким к истинному значению измеряемой величины.

Теорема Чебышева и ее следствие имеют большое практическое значение.

Например, страховой компании необходимо установить размер страхового взноса, который должен уплачивать страхователь; при этом страховая компания обязуется выплатить при наступлении страхового случая определенную страховую сумму. Рассматривая частоту (убытки страхователя) при наступлении страхового случая как величину случайную и обладая известной статистикой таких случаев, можно определить среднее число (средние убытки) при наступлении страховых случаев, которое на основании теоремы Чебышева с большой степенью уверенности можно считать величиной почти неслучайной. Тогда на основании этих данных и предполагаемой страховой суммы определяется размер страхового взноса. Без учета действия закона больших чисел (теоремы Чебышева) возможны существенные убытки страховой компании (при занижении размера страхового взноса) или потеря привлекательности страховых услуг (при завышении размера взноса).

Пример За значение некоторой величины принимают среднеарифметическое достаточно большого числа ее измерений. Предполагая, что среднеквадратичное отклонение возможных результатов каждого измерения не превосходит 5 мм, оценить вероятность того, что при 1000 измерений неизвестной величины отклонение принятого значения от истинного по абсолютной величине не превзойдет 0,5 мм.

Решение. Воспользуемся неравенством

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{n}\sum_{i=1}^{n}M\left(X_i\right)\right|<\varepsilon\right)\geq 1-\frac{C}{n\varepsilon^2}.$$

По условию $n=1000$, $\varepsilon=0{,}5$, $C=5^2=25$. Итак, искомая вероятность

$$P\left(\left|\frac{1}{1000}\sum_{i=1}^{1000}X_i-\frac{1}{1000}\sum_{i=1}^{1000}M\left(X_i\right)\right|<0,5\right)\geq 1-\frac{25}{1000\cdot 0,25}=0,9.$$

Теорема Бернулли. *Частость события в n повторных независимых испытаниях, в каждом из которых оно может произойти с одной и той же вероятностью p, при неограниченном увеличении числа n сходится по вероятности к вероятности p этого события в отдельном испытании:*

$$\boxed{\lim_{n\to\infty}P\left(\left|\frac{m}{n}-p\right|<\varepsilon\right)=1.} \qquad \textit{или} \qquad \boxed{\frac{m}{n}\underset{n\to\infty}{\overset{P}{\to}}p}$$

Доказательство.

Введем случайные величины X_1, X_2, …, X_n, где X_i *(индикатор события A* см. в лекции 6 биномиальный закон распределения*)*– число появлений A в i-м опыте. При этом X_i могут принимать только два значения: 1(с вероятностью p) и 0 (с вероятностью $q = 1 – p$). Кроме того, рассматриваемые случайные величины попарно независимы и их дисперсии равномерно ограничены ($D(X_i) = pq$, $p + q = 1$, откуда $pq \leq ¼$). Следовательно, к ним можно применить теорему Чебышева при $M_i = p$

$$\lim_{n\to\infty}P\left(\left|\frac{X_1+X_2+\ldots+X_n}{n}-p\right|<\varepsilon\right)=1.$$

Но $\frac{X_1+X_2+\ldots+X_n}{n}=\frac{m}{n}$, так как X_i принимает значение, равное 1, при появлении A в данном опыте, и значение, равное 0, если A не произошло. Таким образом, $\lim_{n\to\infty}P\left(\left|\frac{m}{n}-p\right|<\varepsilon\right)=1$, что и требовалось доказать.

Замечание. *Для индикаторов события A справедлива часто применяемая на практике оценка*

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

Пример. При контрольной проверке изготавливаемых приборов было установлено, что в среднем 15 шт. из 100 оказывается с теми или иными дефектами. Оценить вероятность того, что доля приборов с дефектами среди 400 изготовленных будет по абсолютной величине отличаться от математического ожидания этой доли не более чем на 0,05.

Решение. Воспользуемся последним замечанием. По условию $n=400$, $\varepsilon=0{,}05$. В качестве p возьмем величину, полученную при проверке для доли брака $p=\frac{15}{100}=0{,}15$.

Итак, $P\left(\left|\frac{m}{n}-p\right|<\varepsilon\right)\geq 1-\frac{0{,}15\cdot 0{,}85}{400\cdot 0{,}05^2}=0{,}8725$.

Пример Вероятность того, что изделие является качественным, равна 0,9. Сколько следует проверить изделий, чтобы с вероятностью не меньшей 0,95 можно было утверждать, что абсолютная величина отклонения доли качественных изделий от 0,9 не превысит 0,01?

Решение. Воспользуемся еще раз замечанием. По условию $p=0{,}9$, $q=1-0{,}9=0{,}1$, $\varepsilon=0{,}01$. Подставим в правую часть вышеприведенного неравенства эти значения:

$$1-\frac{0{,}9\cdot 0{,}1}{n\cdot 0{,}0001}\geq 0{,}95 \quad\Leftrightarrow\quad \frac{900}{n}\leq 0{,}05 \quad\Leftrightarrow\quad n\geq 18\,000.$$

Ответ: $n\geq 18\,000$.

Центральная предельная теорема

Закон больших чисел устанавливает факт приближения большого числа ***случайных величин к определенным постоянным***. Однако суммарное действие случайных величин не ограничивается только такими закономерностями. Оказывается, что совокупное действие случайных величин приводит к ***определенному*** закону распределения, а именно — **к нормальному закону**.

Теорема Ляпунова. *Если $X_1, X_2, \ldots, X_n$ –**независимые** случайные величины, у каждой из которой есть математическое ожидание $M(X_i) = a_i$ и дисперсия $D(X_i) = \sigma_i^2$, абсолютный центральный момент третьего порядка $M(|X_i - a_i|^3) = m_i$, причем*

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} m_i}{\left(\sum_{i=1}^{n} \sigma^2{}_i\right)^{3/2}} = 0$$

то закон распределения суммы $Y_n = X_1 + X_2 + \ldots + X_n$ при $n \to \infty$ неограниченно приближается к нормальному закону с математическим ожиданием $\sum_{i=1}^{n} a_i$ и дисперсией $\sum_{i=1}^{n} \sigma^2{}_i$.

Следствие. *Если $X_1, X_2, \ldots, X_n$ –**независимые** случайные величины, у которых существуют равные математические ожидания $M(X_i) = a$, дисперсии*

$D(X_i)=\sigma_i^2$, абсолютный центральный момент третьего порядка $M(|X_i-a_i|^3)=m_i$, то закон распределения суммы $Y_n=X_1+X_2+...+X_n$ при $n\to\infty$ неограниченно приближается к нормальному .

В частности, если все случайные величины X_i одинаково распределены, то закон распределения их суммы при $n\to\infty$ неограниченно приближается к нормальному.

Системы нескольких случайных величин

Наряду с одномерными случайными величинами, возможные значения которых определяются одним числом, в теории вероятностей рассматриваются и многомерные случайные величины, когда некоторое испытание характеризуется не одной величиной, а некоторой системой случайных величин: $X_1, X_2, \ldots X_n$. Геометрической иллюстрацией этого понятия служат точки n-мерного пространства, каждая координата которых является случайной величиной (дискретной или непрерывной). Поэтому многомерные случайные величины называют еще случайными векторами $\vec{X} = (X_1, X_2, \ldots X_n)$. Каждое возможное значение такой величины представляет собой упорядоченный набор нескольких чисел $\vec{x} = (x_1, x_2, \ldots x_n)$. Вектор $\vec{x} = (x_1, x_2, \ldots x_n)$ называется реализацией случайного вектора $\vec{X} = (X_1, X_2, \ldots X_n)$.

Двумерные случайные величины

Дискретные двумерные случайные величины

Закон распределения дискретной двумерной случайной величины (X, Y) имеет вид таблицы с двойным входом, задающей перечень возможных значений каждой компоненты $X = x_i$ и $Y = y_j$ и вероятности произведения событий $p_{ij}=p(x_i, y_j)=P((X = x_i)\cdot(Y = y_j))$, с которыми двумерная случайная величина (X, Y) принимает значение (x_i, y_j):

Y	X					
	x_1	x_2	…	x_i	…	x_n
y_1	$p(x_1, y_1)$	$p(x_2, y_1)$	…	$p(x_i, y_1)$	…	$p(x_n, y_1)$
…	…	…	…	…	…	…
y_j	$p(x_1, y_j)$	$p(x_2, y_j)$	…	$p(x_i, y_j)$	…	$p(x_n, y_j)$
…	…	…	…	…	…	…
y_m	$p(x_1,y_m)$	$p(x_2, y_m)$	…	$p(x_i, y_m)$	…	$p(x_n, y_m)$

Так как события $[(X = x_i)(Y = y_j)]$ ($i=1,2,\ldots,n$; $j=1,2,\ldots,m$), состоящие в том, что случайная величина X примет значение x_i, а случайная величина Y примет значение y_j, несовместны и единственно возможны, т.е. образуют полную группу, то сумма их вероятностей равна 1, т.е.

$$\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} = 1$$

Зная закон распределения двумерной случайной величины, можно найти законы распределения ее составляющих. Действительно, событие $X=x_1$ представляет собой сумму несовместных событий $(X=x_1,\ Y=y_1)$, $(X=x_1,\ Y=y_2),\ldots,$ $(X=x_1,\ Y=y_m)$, поэтому $p(X=x_1)=p(x_1,\ y_1) + p(x_1,\ y_2) +\ldots+ p(x_1,\ y_m)$ (в правой части находится сумма вероятностей, стоящих в столбце, соответствующем $X=x_1$). Так же можно найти

вероятности остальных возможных значений X. Для определения вероятностей возможных значений Y нужно сложить вероятности, стоящие в строке таблицы, соответствующей $Y=y_j$.

Пример 1. Дан закон распределения двумерной случайной величины:

Y	X		
	-2	3	6
-0,8	0,1	0,3	0,1
-0,5	0,15	0,25	0,1

Найти законы распределения составляющих.

Решение. Складывая стоящие в таблице вероятности «по столбцам», получим ряд распределения для X:

X	-2	3	6
p	0,25	0,55	0,2

Складывая те же вероятности «по строкам», найдем ряд распределения для Y:

Y	-0,8	-0,5
p	0,5	0,5

Функция распределения двумерной случайной величины

Функцией распределения $F(x, y)$ двумерной случайной величины (X, Y) называется вероятность совместного выполнения неравенств $X<x$ и $Y<y$:

$$F(x, y)=P(X<x, Y<y).$$

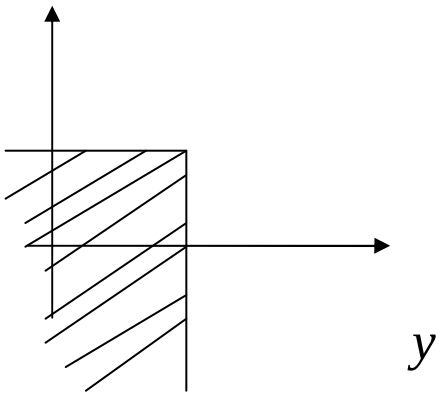


Рис.1.

Геометрически функция распределения $F(x, y)$ означает вероятность попадания случайной точка (X, Y) в область, заштрихованную на рис.1, если вершина прямого угла располагается в точке (x, y). Правая и верхняя границы в квадрант не включаются – это означает, что функция распределения *непрерывна слева* по каждому из аргументов.

Замечание. Определение функции распределения справедливо как для непрерывной, так и для дискретной двумерной случайной величины. В случае дискретной двумерной случайной величины ее функция распределения определяется по формуле:

$$F(x, y) = \sum_{\substack{i \\ x_i < x}} \sum_{\substack{j \\ y_j < y}} p_{ij}$$

Свойства функции распределения:

1) $0 \le F(x, y) \le 1$ (так как $F(x, y)$ является вероятностью).

2) $F(x, y)$ есть неубывающая функция по каждому аргументу:

$$F(x_2, y) \geq F(x_1, y), \text{ если } x_2 > x_1;$$

$$F(x, y_2) \geq F(x, y_1), \text{ если } y_2 > y_1.$$

Доказательство.

$$F(x_2,y)=P(X<x_2,Y<y)=P(X<x_1,Y<y)+p(x_1\leq X<x_2,Y<y)\geq$$

$$\geq P(X < x_1, Y < y) = F(x_1, y).$$

Аналогично доказывается и второе утверждение.

3) Имеют место предельные соотношения:

a) $F(-\infty, y) = 0$; b) $F(x, -\infty) = 0$; c) $F(-\infty, -\infty) = 0$;

d) $F(+\infty, +\infty) = 1$.

Доказательство. События a), b) и c) невозможны (так как невозможно событие $X<-\infty$ или $Y<-\infty$), а событие d) достоверно, откуда следует справедливость приведенных равенств.

4) При $y=+\infty$ функция распределения двумерной случайной величины становится функцией распределения составляющей X:

$$F(x,+\infty) = F_1(x).$$

При $x=+\infty$ функция распределения двумерной случайной величины становится функцией распределения составляющей Y:

$$F(+\infty, y)=F_2(y).$$

Доказательство. Так как событие $Y<+\infty$ достоверно, то $F(x,+\infty)=P(X<x)=F_1(x)$. Аналогично доказывается второе утверждение.

5. Справедлива формула

$$P[(x_1 < X < x_2)(y_1 < Y < y_2)] = F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1)$$

Двумерная случайная величина (X, Y) называется *непрерывной*, если ее функция распределения $F(x, y)$ – непрерывная функция, дифференцируемая по каждому из аргументов, и существует вторая смешанная производная: $\dfrac{\partial^2 F(x,y)}{\partial x \partial y}$

Плотность вероятности двумерной случайной величины

Плотностью совместного распределения вероятностей (двумерной плотностью вероятности) непрерывной двумерной случайной величины называется смешанная частная производная 2-го порядка от функции распределения:

$$\boxed{f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y}}.$$

Замечание. Двумерная плотность вероятности представляет собой предел отношения вероятности попадания случайной точки в прямоугольник со сторонами Δx и Δy к площади этого прямоугольника при $\Delta x \to 0,\ \Delta y \to 0$.

Свойства двумерной плотности вероятности:

1) $f(x, y) \geq 0$ (см. предыдущее замечание: вероятность попадания точки в прямоугольник неотрицательна, площадь этого прямоугольника положительна, следовательно, предел их отношения неотрицателен).

2) $F(x,y) = \int\limits_{-\infty}^{y}\int\limits_{-\infty}^{x} f(x,y)dxdy$

(следует из определения двумерной плотности вероятности).

3) $\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty} f(x,y)dxdy = 1$

(поскольку это вероятность того, что точка попадет на плоскость O*xy*, то есть достоверного события).

4) Вероятность попадания случайной точки в произвольную область: $\boxed{P((X,Y) \subset D) = \iint\limits_{D} f(x,y)dxdy.}$

Пусть в плоскости O*xy* задана произвольная область *D*. Найдем вероятность того, что точка, координаты которой представляют собой систему двух случайных величин (двумерную случайную величину) с плотностью распределения $f(x, y)$, попадет в область *D*. Разобьем эту область прямыми, параллельными осям координат, на прямоугольники со сторонами Δx и Δy. Вероятность попадания в каждый такой прямоугольник равна $f(\xi_i,\eta_i)\Delta x\Delta y$, где (ξ_i,η_i) – координаты точки, принадлежащей прямоугольнику. Тогда вероятность

попадания точки в область D есть предел интегральной суммы $\sum_{i=1}^{n} f(\xi_i,\eta_i)\Delta x\Delta y$, то есть

$$P((X,Y)\subset D)=\iint_D f(x,y)dxdy.$$

Функции распределения и плотности вероятностей одномерных составляющих двумерной случайной величины

Нахождение функции распределения каждой составляющей по известной двумерной функции плотности вероятности :

$$F_1(x)=F(x,+\infty)=\int_{-\infty}^{+\infty}\int_{-\infty}^{x} f(x,y)dxdy$$

$$F_2(y)=F(+\infty,y)=\int_{-\infty}^{y}\int_{-\infty}^{+\infty} f(x,y)dxdy$$

Тогда для определения одномерной плотности распределения имеем

$$f_1(x)=\frac{dF_1(x)}{dx}=\frac{dF(x,\infty)}{dx}=\frac{d\left(\int_{-\infty}^{x}\int_{-\infty}^{\infty} f(x,y)\right)}{dx}=\int_{-\infty}^{\infty} f(x,y)dy.$$

Аналогично найдем $$f_2(y)=\int_{-\infty}^{\infty} f(x,y)dx.$$

Условные законы распределения

Рассмотрим дискретную двумерную случайную величину и найдем закон распределения составляющей X при условии, что Y примет определенное значение (например, $Y=y_1$). Для этого воспользуемся формулой Байеса, считая гипотезами события $X=x_1$, $X=x_2$,…, $X=x_n$, а событием A – событие $Y=y_1$. При такой постановке задачи нам требуется найти условные вероятности гипотез при условии, что A произошло. Следовательно,

$$p(x_i \mid y_1) = \frac{P((X = x_i)(Y = y_1))}{P(Y = y_1)}.$$

Таким же образом можно найти вероятности возможных значений X при условии, что Y принимает любое другое свое возможное значение:

$$p(x_i \mid y_j) = \frac{P((X = x_i)(Y = y_j))}{P(Y = y_j)}.$$

Аналогично находят условные законы распределения составляющей Y:

$$p(y_j \mid x_i) = \frac{P((X = x_i)(Y = y_j))}{P(X = x_i)}.$$

Пример 2. Найдем закон распределения X при условии Y=-0,8 и закон распределения Y при условии X=3 для случайной величины, рассмотренной в примере 1.

Решение.

$$p(x_1 \mid y_1) = \frac{0,1}{0,5} = \frac{1}{5} = 0,2; \qquad p(x_2 \mid y_1) = \frac{0,3}{0,5} = \frac{3}{5} = 0,6;$$

$$p(x_3 \mid y_1) = \frac{0,1}{0,5} = \frac{1}{5} = 0,2.$$

$$p(y_1 \mid x_2) = \frac{0,3}{0,55} = \frac{6}{11}; \quad p(y_2 \mid x_2) = \frac{0,25}{0,55} = \frac{5}{11}.$$

Условной плотностью вероятности $f(x \mid y) = f_y(x)$ распределения составляющих X при данном значении $Y=y$ называется

$$f(x \mid y) = f_y(x) = \frac{f(x,y)}{f_2(y)} = \frac{f(x,y)}{\int\limits_{-\infty}^{\infty} f(x,y)dx}.$$

Аналогично определяется условная плотность вероятности Y при $X=x$:

$$f(y \mid x) = f_x(y) = \frac{f(x,y)}{f_1(x)} = \frac{f(x,y)}{\int\limits_{-\infty}^{\infty} f(x,y)dy}.$$

Числовые характеристики двумерных случайных величин

Такие характеристики, как начальные и центральные моменты, можно ввести и для системы двух случайных величин.

Начальным моментом порядка k, s двумерной случайной величины (X, Y) называется математическое ожидание произведения X^k на Y^s:

$$\alpha_{k,s} = M(X^k Y^s).$$

Для дискретных случайных величин

$$\alpha_{k,s} = \sum_i \sum_j x_i^k y_j^s p_{ij},$$

Для непрерывных случайных величин

$$\alpha_{k,s} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x^k y^s f(x,y)dxdy.$$

Центральным моментом порядка k, s двумерной случайной величины (X, Y) называется математическое ожидание произведения $(X - M(X))^k$ на $(Y - M(Y))^s$:

$$\mu_{k,s}=M(((X - M(X))^k(Y - M(Y))^s).$$

Для дискретных случайных величин

$$\mu_{k,s} = \sum_i \sum_j (x_i - M(X))^k (y_j - M(Y))^s p_{ij}$$

Для непрерывных случайных величин

$$\mu_{k,s} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - M(X))^k (y - M(Y))^s f(x,y)dxdy \quad (8.23)$$

При этом

$M(X) = \alpha_{1,0}$, $M(Y) = \alpha_{0,1}$, $D(X) = \mu_{2,0}$, $D(Y) = \mu_{0,2}$.

Наряду с числовыми характеристиками $M(X)$, $M(Y)$, $D(X)$, $D(Y)$ одномерных составляющих рассматриваются также числовые характеристики условных распределений: условные математические ожидания $M_y(X)$, $M_x(Y)$, $D_y(X)$, $D_x(Y)$.

Например,

$$M_y(X)=\int_{-\infty}^{+\infty} xf_y(x)dx, \qquad D_y(X)=\int_{-\infty}^{+\infty}(x-M(X))^2 f_y(x)dx$$

Корреляционный момент и коэффициент корреляции

Корреляционным моментом (или ковариацией) системы двух случайных величин называется центральный момент порядка 1,1:

$$K_{xy}=\mu_{1,1}=M(((X-M(X))(Y-M(Y))).$$

Для дискретных случайных величин

$$K_{xy}=\sum_i\sum_j (x_i-M(X))(y_j-M(Y))\,p_{ij}$$

Для непрерывных случайных величин

$$K_{xy}=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-M(X))(y-M(Y))\,f(x,y)dxdy$$

Ковариация может быть вычислена по формуле:

$$K_{xy}=M(XY)-M(X)\cdot M(Y)$$

Действительно,

$$K_{xy}=M(XY-Y\cdot M(X)-X\cdot M(Y)+M(X)M(Y))$$

Тогда по свойству математического ожидания:

$$K_{xy}=M(XY)-M(Y\cdot M(X))-M(X\cdot M(Y))+M(M(X)M(Y))=$$
$$=M(XY)-M(Y)\cdot M(X)-M(X)\cdot M(Y)+M(X)M(Y)=$$
$$=M(XY)-M(X)\cdot M(Y)$$

Корреляционный момент описывает связь между составляющими двумерной случайной величины.

Случайные величины X и Y называются *некоррелированными*, если $K_{xy}=0$.

Убедимся, что для независимых X и Y $K_{xy}=0$. Действительно, в этом случае $f(x,y)=f_1(x)f_2(y)$, тогда

$$K_{xy}=\int_{-\infty}^{\infty}(x-M(X))\,f_1(x)dx\int_{-\infty}^{\infty}(y-M(Y))\,f_2(y)dy=\mu_1(x)\mu_1(y)=0.$$

Итак, две ***независимые*** случайные величины являются и ***некоррелированными***.

Однако понятия коррелированности и зависимости не эквивалентны, а именно, величины могут быть зависимыми, но при этом некоррелированными. Дело в том, что корреляционный момент характеризует не всякую зависимость, а только *линейную*. В частности, если $Y=aX+b$, то $K_{xy}=\pm\sigma_x\sigma_y$.

Безразмерной характеристикой коррелированности двух случайных величин является *коэффициент корреляции*

$$r_{xy}=\frac{K_{xy}}{\sigma_x\sigma_y}\ .$$

Теорема 1. Возможные значения коэффициента корреляции удовлетворяют неравенству $|r_{xy}| \le 1$.

Доказательство. Докажем сначала, что $|K_{xy}| \le \sigma_x \sigma_y$. Действительно, если рассмотреть случайную величину $Z_1 = \sigma_y X - \sigma_x Y$ и найти ее дисперсию, то получим: $D(Z_1) = 2\sigma_x^2 \sigma_y^2 - 2\sigma_x \sigma_y K_{xy}$. Так как дисперсия всегда неотрицательна, то $2\sigma_x^2 \sigma_y^2 - 2\sigma_x \sigma_y K_{xy} \ge 0$, откуда $|K_{xy}| \le \sigma_x \sigma_y$. Отсюда $\left|\frac{K_{xy}}{\sigma_x \sigma_y}\right| = |r_{xy}| \le 1$, что и требовалось доказать.

Равномерное распределение на плоскости

Система двух случайных величин называется *равномерно распределенной на плоскости*, если ее плотность вероятности $f(x, y)=C=\text{const}$ внутри некоторой области и равна 0 вне ее. Пусть данная область – прямоугольник вида $a \le x \le b,\ c \le y \le d$.

Тогда

$$f(x,y) = \begin{cases} \frac{1}{S_{пр}} = \frac{1}{(b-a)(d-c)} & \textit{внутри прямоугольника}, \\ 0 & \textit{вне его}. \end{cases}$$

Действительно

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)dxdy = \int_a^b\int_c^d C dxdy = C(b-a)(d-c) = 1 \Rightarrow C = \frac{1}{(b-a)(d-c)}$$

Найдем двумерную функцию распределения:

$$F(x,y)=\frac{1}{(b-a)(d-c)}\int_c^y\int_a^x dxdy=\frac{(x-a)(y-c)}{(b-a)(d-c)} \quad \text{при} \quad a<x<b,$$

$c<y<d$,

$F(x, y)=0$ при $x \le a$ или $y \le c$,

$F(x, y)=1$ при $x \ge b, y \ge d$.

Функции распределения составляющих, вычисленные по формулам, приведенным в свойстве 4 функции распределения, имеют вид:

$$F_1(x)=\frac{x-a}{b-a}, \quad F_2(y)=\frac{y-c}{d-c}.$$

КОНТРОЛЬНЫЕ ВОПРОСЫ
по курсу «Теория вероятностей»

1. Предметы и методы теории вероятностей и математической статистики.
2. Элементы комбинаторики, размещения, перестановки, сочетания.
3. Случайные события. Операции над событиями.
4. Классическая формула вероятности. Статистическая вероятность. Геометрические вероятности.
5. Теорема сложения вероятностей.
6. Условная вероятность. Теорема умножения вероятностей.
7. Формула полной вероятности. Формула Байеса.

8. Формула Бернулли.
9. Формула Пуассона.
10. Локальная и интегральная теоремы Лапласа.
11. Дискретные случайные величины.
12. Непрерывные случайные величины.
13. Функция распределения вероятностей.
14. Плотность распределения вероятностей.
15. Математическое ожидание случайной величины и его свойства.
16. Дисперсия случайной величины и ее свойства.
17. Моменты случайных величин.
18. Биномиальное распределение и его характеристики.
19. Распределение Пуассона.
20. Геометрическое и гипергеометрическое распределения.
21. Равномерное распределение в интервале.
22. Показательное распределение.
23. Нормальный закон распределения.
24. Неравенство Маркова. Неравенство Чебышева.
25. Теорема Чебышева. Теорема Бернулли.
26. Системы нескольких случайных величин. Дискретные двумерные случайные величины.
27. Системы нескольких случайных величин. Функция распределения и плотность вероятностей двумерной случайной величины.
28. Числовые характеристики двумерных случайных величин.
29. Корреляционный момент и коэффициент корреляции двумерных случайных величин.

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Задачи математической статистики.

Теория вероятностей (ТВ) — математическая наука, изучающая закономерности случайных явлений. Под случайными явлениями понимаются явления с неопределенным исходом, происходящие при неоднократном воспроизведении некоторого комплекса условий.

Математическая статистика (МС) — раздел математики, изучающий методы сбора, систематизации и обработки результатов наблюдений с целью выявления статистических закономерностей. МС опирается на ТВ. ТВ изучает закономерности случайных явлений на основе абстрактной (теоретической) вероятностной модели действительности. МС оперирует непосредственно с результатами наблюдений над случайным явлением. Основной задачей МС является разработка методов нахождения законов и числовых характеристик случайных величин по результатам экспериментов или наблюдений. Используя результаты, полученные теорией вероятностей, МС по наблюдаемым значениям (выборке) оценивает вероятности этих событий либо осуществляет проверку предположений (гипотез) относительно этих вероятностей.

Таким образом, задачи МС:

1. указать способы сбора и группировки статистических сведений, полученных в результате наблюдений или в результате специально поставленных экспериментов;
2. разработать методы анализа статистических данных в зависимости от целей исследования.

Ко второй задаче относятся:

а) оценка неизвестной вероятности события; оценка неизвестной функции распределения; оценка параметров распределения, вид которого известен; оценка зависимости

случайной величины от одной или нескольких случайных величин и т.д.

б) проверка статистических гипотез о виде неизвестного распределения или о величине параметров распределения, вид которого известен.

Современную МС определяют как ***науку о принятии решений в условиях неопределенности***.

Генеральная и выборочная совокупности.

Пусть требуется изучить совокупность однородных объектов относительно некоторого *качественного* или *количественного* признака, характеризующего эти объекты. Например, в партии изделий *качественным* признаком может служить *стандартность* деталей, а *количественным* – контролируемый *размер* деталей.

Предположим, что имеется некоторое множество $x_1, x_2, \ldots, x_N$ однородных предметов, определенный *признак* которых исследуется.

Вся подлежащая изучению совокупность предметов называется ***генеральной совокупностью***.

В статистике различают два вида наблюдений: ***сплошное***, когда изучаются все объекты совокупности (перепись населения, например) и несплошное , т.е ***выборочное***, когда изучается только часть объектов (социологическое обследование, например).

Предположим далее, что исследовать данный признак у всех предметов этой совокупности **не представляется возможным** (либо их очень много, либо они физически уничтожаются, либо по другим причинам). В этом случае используют *выборочный метод*, согласно которому из данной генеральной совокупности *случайным образом* выбираются n элементов $x_1, x_2, \ldots, x_n$. Та часть объектов, которая отобрана для непосредственного изучения из генеральной

совокупности, называется ***выборочной совокупностью*** или ***выборкой***.

Задача МС состоит в исследовании свойств выборки и обобщении этих свойств на всю генеральную совокупность. Полученный при этом результат называют ***статистическим***.

Преимущества статистического метода:

1. экономия затрат и ресурсов
2. является единственно возможным в случае бесконечной генеральной совокупности или если исследование связано с уничтожением объектов (крэш-тесты, например)
3. снизить ошибки регистрации

Недостатки: Ошибки исследования, порожденные ограниченным объемом изученных объектов генеральной совокупности.

Размахом выборки R называют разность между максимальным $x_{\max}$ и минимальным $x_{\min}$ значениями элементов выборки: $R = x_{\max} - x_{\min}$.

Объёмом совокупности (выборочной или генеральной) называют *число* (n или N соответственно) *объектов* этой совокупности. Если, например, из 1000 деталей для обследования отобрано 85, то объём генеральной совокупности N=1000, объём выборки n=85.

Требования к выборке. Для того, чтобы результаты обследования выборки отражали основные черты изучаемого признака, необходимо, чтобы объём выборки не был слишком малым.

Выборка называется ***репрезентативной*** (представительной), если она достаточно хорошо представляет количественные соотношения генеральной совокупности.

Например, о распределении жителей г. Минска по росту нельзя судить по результатам обследования одной квартиры. Ясно, что данные, относящиеся к одному высотному дому или группе домов, более показательны, репрезентативны.

Одним из способов по обеспечению репрезентативности выборки является **случайность** в ее отборе.

Случайность отбора элементов в выборку достигается соблюдением принципа равной возможности всем элементам генеральной совокупности быть отобранным в выборку.
На практике это достигается, например, тем, что извлечение производится путем лотереи (жеребьевки), или с помощью датчика случайных чисел.
Различают 2 способа образования случайной выборки:

- Повторный отбор, когда каждый элемент случайно отобранный и обследованный, возвращается в общую генеральную совокупность и *может быть повторно отобран*
- Бесповторный отбор, когда уже однажды отобранный элемент не возвращается.

Вариационный ряд и его основные числовые характеристики

Пусть генеральная совокупность описывается неким общим признаком, например, измеряется некоторая величина ξ. На измерения могут влиять как систематические ошибки (погрешность прибора), так и случайные ошибки (внешние условия и т.п.), получаемые в результате воздействия различных (случайных) факторов. *Таким образом, ξ можно интерпретировать как случайную величину (СВ).*
Предположим далее, что возможные значения СВ ξ известны:

$$x_1, x_2, \cdots, x_n \qquad (1.1)$$

Эти значения можно считать генеральной совокупностью. Если же известны и вероятности появления значений $x_1, x_2, \cdots, x_n$, то нам известен и закон распределения СВ ξ(*теоретический закон распределения* или *распределение генеральной совокупности*). На практике она, как правило, неизвестна.

В этом случае производят измерения (случайные) величины ξ, и в результате получают значение $\tilde{x}_1$ этой СВ. Однако судить о значениях СВ ξ по одному измеренному значению $\tilde{x}_1$ неубедительно. Поэтому на практике производят n *независимых совокупностей* случайных испытаний (измерений) данной СВ. В результате получают k реализовавшихся значений

$$\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_k \qquad (1.2)$$

данной СВ, которые называют *выборочными значениями* (данной СВ). Совокупность (1.2) можно интерпретировать как *выборку объёма* k. Выборочные значения $\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_k$ называют *вариантами*. Рассмотрение и осмысление этих данных (при большом k) *затруднительно*.

Первый шаг к осмыслению полученного статистического материала – это его упорядочивание (или ранжирование) по возрастанию или убыванию вариант ряда.

Совокупность (1.2), ранжированная в порядке возрастания (убывания), называется ***вариационным рядом***.

Выберем из (1.2) **только различные**, расположенные в возрастающем порядке значения $x_1^*, x_2^*, ,\cdots, x_r^*$.

Пусть признак x_1^* наблюдался n_1 раз, x_2^* – n_2 раза,…, x_r^* – n_r раз. Тогда $\sum_{i=1}^{r} n_i = n$, где n – объём выборки, наблюдаемые величины $x_1^*, x_2^*, \cdots \cdots, x_r^*$ – варианты.

Числа наблюдений n_i, $i=1,2,\cdots,p$, называют ***частотами***, а отношение n_i/n частот к объёму выборки – ***относительными частотами*** ω_i (частотами или долями):

$$\omega_i=\frac{n_i}{n},\quad i=\overline{1,p};\qquad \sum_{i=1}^{r}\omega_i=\sum_{i=1}^{r}\frac{n_i}{n}=\frac{1}{n}\underbrace{\sum_{i=1}^{r}n_i}_{=n}=1. \tag{1.3}$$

В теории вероятностей под распределением понимают соответствие между возможными значениями случайных величин и их вероятностями

ξ_i	x_1	x_2	…	x_r
p_i	p_1	p_2	…	p_r

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

В математической статистике под распределением понимают соответствие между наблюдаемыми вариантами и их частотами или относительными частотами.

Статистическим рядом или статистическим распределением выборки называют совокупность пар (x_i,n_i), $i=\overline{1,k}$, где x_1, $x_2,\ldots,x_k$ – различные элементы выборки, а $n_1,n_2,\ldots,n_k$ – частота выборочных значений

$$x_1,\ x_2,\ldots,x_k,\ \sum_{i=1}^{k}n_i=n.$$

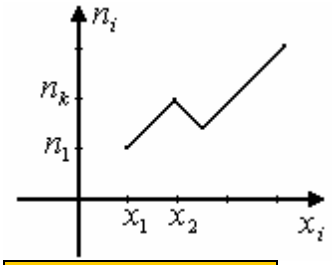
Статистический ряд записывается в виде таблицы

x_i	x_1	x_2	…	x_k
n_i	n_1	n_2	…	n_k

$$\sum_{i=1}^{k}n_i=n$$

Для наглядности принято использовать ***полигон частот*** как форму *графического представления статистических распределений*. ***Полигоном частот*** (относительных частот) выборки называется ломаная с вершинами в точках (x_i,n_i), $i=\overline{1,k}$, $(x_i,n_i/n)$, $i=\overline{1,k}$ (по оси координат

откладываются выборочные значения x_i, по оси ординат – соответствующие частоты n_i или относительные частоты ω_i):



Пример 1. Выборка, полученная в результате статистического наблюдения (ед. измерения опускаем) –
7, 17, 14, 17, 10, 7, 7, 14, 7, 14;
Ранжированный вариационный ряд –
x_j : $\underbrace{7,7,7,7}_{n_1=4}, \underbrace{10,}_{n_2=1} \underbrace{14,14,14}_{n_3=3}, \underbrace{17,17}_{n_4=2}$, где $j=1,2,\ldots,n$, n = 10;
Статистическое распределение ($i=1,2,\ldots,k$, k = 4):

x_i	7	10	14	17
n_i	4	1	3	2.

При большом объеме выборки ее элементы объединяют в группы (разряды, интервалы), представляя результаты опытов в виде *интервального статистического ряда*. Для этого весь диапазон значений случайной величины ξ (от $x_{\min}$ до $x_{\max}$) разбивают на k интервалов одинаковой длины h (обычно k меняется от 5 до 20).

Число интервалов рекомендуют брать согласно формуле Стерджеса $\boxed{k=1+3,93\cdot\ln n}$

Затем подсчитывают частоты n_i (или относительные частоты ω_i) значений выборки, попавших в выделенные интервалы.

Величина n_i/h называется *плотностью частоты*, а ω_i/h – *плотностью относительной частоты*.

Пусть x_i^* – середина i-го интервала, n_i – число элементов выборки, попавших в i-й интервал (при этом элемент, совпавший с верхней границей интервала, относится к

последующему интервалу). Таким образом, *получим группированный статистический ряд*, в верхней строке которого записаны середины соответствующих интервалов x_i^* (иногда пишут сами интервалы), а в нижней — частоты:

x_i^*	x_1^*	x_2^*	…	x_k^*
n_i	n_1	n_2	…	n_k

$$\sum_{i=1}^{k} n_i = n$$

Пример 2. Выборка, полученная в результате статистического наблюдения

3,14; 1,41; 2,87; 3,62; 2,71; 3,95;

– ранжированный вариационный ряд –

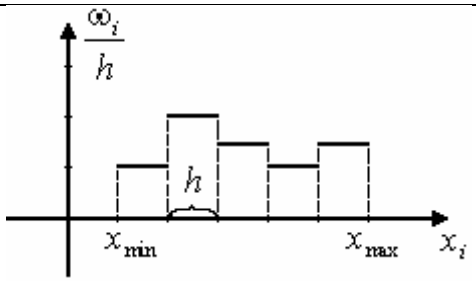
x_j: 1,41; 2,71; 2,87; 3,14; 3,62; 3,95; где $j = 1, 2, \ldots, n$, $n = 6$;

– соответствующее интервальное статистическое распределение ($i = 1, 2, \ldots, k$, $k = 3$):

x_i	1–2	2–3	3–4
n_i	1	2	3.

Для *графического представления интервальных статистических распределений* принято использовать ***гистограмму относительных частот.***

Гистограммой относительных частот интервального статистического ряда называется ступенчатая фигура, составленная из прямоугольников, построенных на интервалах группировки длины h и высоты ω_i / h так, что площадь каждого прямоугольника равна относительной частоте ω_i.



Для построения гистограммы относительных частот на оси абсцисс откладывают частичные интервалы, а над ними

проводят отрезки длиной ω_i / h параллельно оси ординат. Очевидно, площадь i-го частичного прямоугольника равна ω_i – относительной частоте вариант, попавших в i-ый интервал. Следовательно, ***площадь гистограммы относительных частот равна сумме всех относительных частот (т.е. равна 1), а площадь гистограммы частот равна объему выборки*** n**.**

Пример 3. Имеется распределение 80 предприятий по числу работающих на них (чел.):

x_i	150	250	350	450	550	650	750
n_i	1	3	7	30	19	15	5 .

Построить графическое представление.

Решение. Признак X – число работающих (чел.) на предприятии. В данной задаче признак X является дискретным. Поскольку различных значений признака сравнительно немного – $k = 7$, применять интервальный ряд для представления статистического распределения нецелесообразно (в прикладной статистике в подобных задачах часто используют именно интервальный ряд). Ряд распределения – дискретный. Построим полигон распределения частот

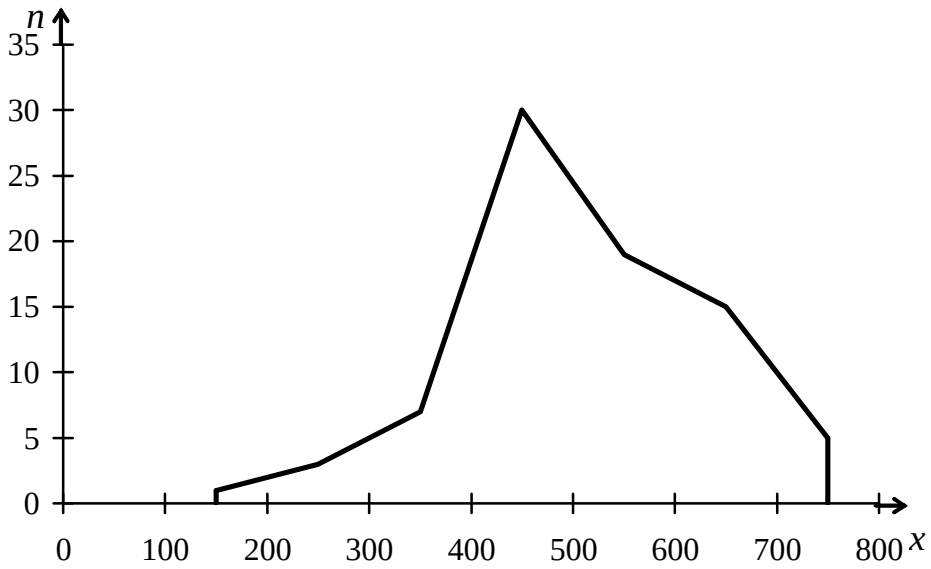


Рис. 1

Пример 4. Дано распределение 100 рабочих по затратам времени на обработку одной детали (мин):

x_{i-1}–x_i	22–24	24–26	26–28	28–30	30–32	32–34
n_i	2	12	34	40	10	2

Решение. Признак X – затраты времени на обработку одной детали (мин). Признак X – непрерывный, ряд распределения – интервальный. Построим гистограмму частот (рис. 2), предварительно определив $h = (x_k - x_0)/k = (34-22)/6 = 2$ $(k = 6)$ и плотность частоты n_i/h:

x_{i-1}–x_i	22–24	24–26	26–28	28–30	30–32	32–34
n_i/h	1	6	17	20	5	1 .

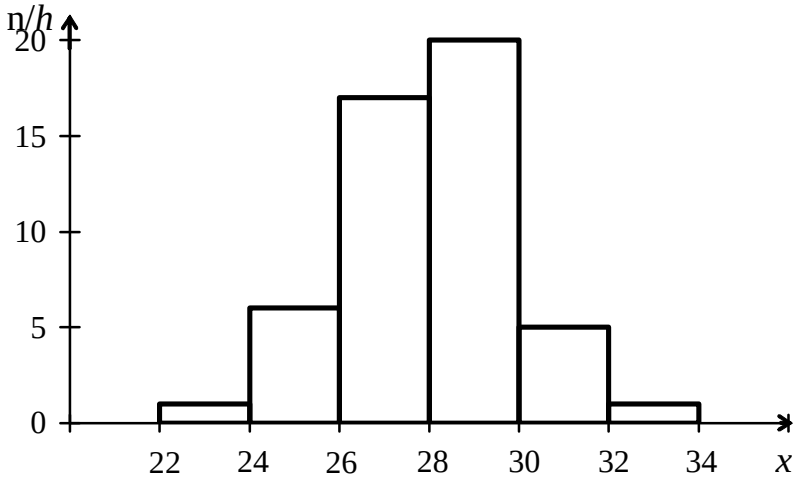


Рис. 2

Эмпирическая функция распределения

Пусть известно статистическое распределение (или статистический ряд) количественного признака ξ;

n_x – число наблюдений, при которых наблюдалось значение признака, меньшее x, т.е. $\xi < x$;

n – общее число наблюдений (объём выборки).

Тогда относительная частота события $\xi < x$ есть n_x/n. При изменении x меняется и n_x/n, т.е. относительная частота

n_x/n является функцией x. Так как эта функция находится эмпирическим (т.е. опытным) путём, то её называют ***эмпирической.***

Эмпирической функцией распределения (функцией распределения выборки) называется функция

$$F^*(x)=\frac{n_x}{n}, \qquad (1.4)$$

определяющая для каждого значения $x\in R$ относительную частоту события $\xi<x$.

В (1.4) n_x – число вариант, меньших x т.е. ($n_x=\sum_{x_i<x} n_i$, x_i – варианты, n – объём выборки).

Поэтому для расчетов удобна формула вида:

$$F^*(x)=\sum_{x_i<x}\frac{n_{x_i}}{n} \qquad (1.5)$$

Тогда, например, $F^*(x_3)$ означает $F^*(x_3)=n_{x_3}/n$, где n_{x_3} – число вариант, меньших x_3.

Функцию распределения $F(x)$ *генеральной совокупности* ***называют теоретической функцией распределения***.

Различие между эмпирической $F^*(x)$ и теоретической $F(x)$ функциями распределения состоит в том, что $F(x)$ определяет вероятность события $\xi<x$, а $F^*(x)$ – относительную частоту того же события.

Функция $F^*(x)$ обладает всеми свойствами $F(x)$.

Свойства эмпирической функции распределения $F^*(x)$*:*

1. значения $F^*(x)$ принадлежат [0,1]; $F^*(x)\in[0;1]$;
2. $F^*(x)$ – неубывающая функция;
3. если x_1 – наименьшая варианта, а x_k – наибольшая варианта, то $F^*(x)=0$ для $x\le x_1$, $F^*(x)=1$ для $x>x_k$;
4. $F^*(x)$ непрерывная слева функция.

Эмпирическая функция распределения выборки $F^*(x)$ ***служит для оценки теоретической функции распределения*** $F(x)$ ***генеральной совокупности.***

Пример 5. Построить эмпирическую функцию распределения по данному распределению выборки

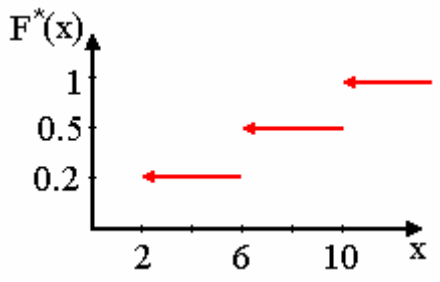
Варианты x_i	2	6	10
Частоты n_i	12	18	30

$n=\sum_{i=1}^{3} n_i = 60.$

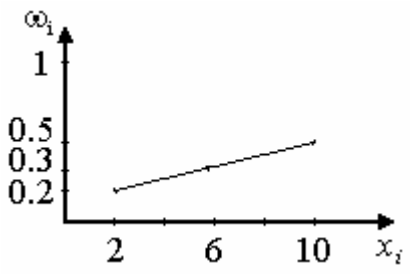
Решение. Здесь $x_1 = 2$ – наименьшая варианта, следовательно, $F^*(x)=0$ для $x \le 2$; $x_3 = 10$ – наибольшая варианта, тогда $F^*(x)=1$ при $x>10$. Для $2<x\le 6$ имеем $F^*(x)=n_x/n=\sum_{x_i<x} n_i/n$=12/60=0,2, а для $6<x\le 10$ следует $F^*(x)$=(12+18)/60=0,5.

Приведём аналитический вид полученной эмпирической функции распределения $F^*(x)$, её график и полигон частот:

$$F^*(x)=\begin{cases} 0, & x\le 2 \\ 0{,}2, & 2<x\le 6 \\ 0{,}5, & 6<x\le 10 \\ 1, & x>10 \end{cases},$$



Полигон относительных частот имеет вид

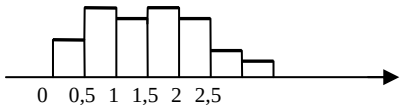


где координаты ω_i его вершин (x_i, ω_i) определяются по формулам

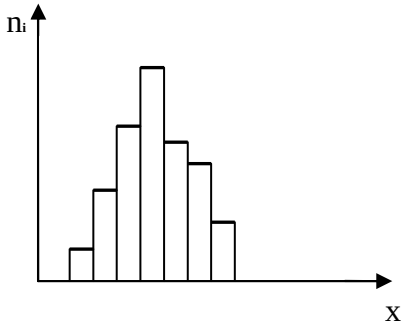
$$\omega_1 = \frac{12}{60} = 0,2, \quad \omega_2 = \frac{18}{60} = 0,3, \quad \omega_3 = \frac{30}{60} = 0,5, \quad \Sigma\omega_i = 1.$$

Точечное оценивание параметров распределения.

Анализ полигона, гистограммы, эмпирической функции распределения даёт возможность сделать допущение о законе распределения случайных величин. По виду полученной гистограммы можно строить гипотезы об истинном характере распределения СВ ξ. Например, получив гистограмму вида:



можно заключить, что СВ ξ на отрезке [0,5;2,5] распределена равномерно. Из гистограммы вида



естественно предположить, что распределение СВ ξ является нормальным. На практике, однако, редко встречается такое положение, когда изучаемый закон распределения СВ ξ неизвестен полностью. Чаще всего из каких-либо теоретических соображений вид закона распределения ясен заранее и требуется найти только некоторые *параметры*, от которых он зависит. Например, если известно, что закон распределения СВ ξ нормальный (с плотностью распределения $f_\xi(x)=\frac{2}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}$), то задача сводится к нахождению значений двух параметров: a и σ. В некоторых задачах и сам вид закона распределения несущественен, а требуется найти только его числовые характеристики. Во всех подобных случаях можно обойтись сравнительно небольшим числом наблюдений – порядка одного или нескольких десятков.

При *изучении числовых характеристик* СВ ξ мы рассматриваем математическое ожидание M_ξ, дисперсию D_ξ, среднее квадратичное отклонение σ_ξ. Эти числовые (т.е. точечные) характеристики играют большую роль в теории вероятностей. Аналогичные числовые характеристики существуют и для статистических распределений. Каждой числовой характеристике СВ ξ соответствует её статистическая аналогия.

Пусть закон распределения СВ ξ содержит некоторый параметр θ. Численное значение θ не указано, хотя оно и является вполне определенным числом. В связи с этим возникает следующая задача.

Исходя из набора наблюдаемых значений (выборки) $x_1, x_2, \ldots, x_n$ *случайной величины* ξ*, полученного в результате* n *независимых испытаний, оценить значение параметра* θ.

Оценкой (или ***статистикой***) θ_n^* неизвестного параметра θ теоретического распределения называют функцию $f(x_1, x_2, \ldots, x_n)$ от наблюдаемых (выборочных) значений случайных величин $x_1, x_2, \ldots, x_n$, обладающую свойством статистической устойчивости. Так как $x_1, x_2, \ldots, x_n$ рассматриваются как независимые случайные величины, то и оценка θ_n^* является случайной величиной, зависящей от закона распределения СВХ и числа n. Сам же оцениваемый параметр есть величина неслучайная (детерминированная).

Оценки параметров разделяются на *точечные* и *интервальные*.

Точечной называют статистическую оценку, определяемую одним числом $\theta_n^* = f(x_1, x_2, \ldots, x_n)$ (далее будем обозначать просто θ^*), где $x_1, x_2, \ldots, x_n$ – результаты n наблюдений (выборки) над количественным признаком ξ.

К оценке θ^* естественно предъявить ряд *требований:*

1. Желательно, чтобы, пользуясь величиной θ^* вместо θ, не делалось систематических ошибок ни в сторону занижения, ни в сторону завышения, т.е. чтобы выполнялось равенство

$$M(\theta^*) = \theta. \qquad (2.1)$$

Оценка, удовлетворяющая условию (2.1), называется ***несмещённой***.

Несмещённой называют точечную оценку, математическое ожидание которой равно оцениваемому параметру при любом объёме выборки. Требование несмещённости оценки особенно важно при малом числе испытаний.

Смещённой называют оценку, математическое ожидание которой не равно оцениваемому параметру θ.

2. Желательно, чтобы с увеличением числа n опытов значения случайной величины θ^* концентрировались около θ всё более тесно, т.е.

$$\theta^* \xrightarrow{P} \theta \quad при \quad n \to \infty \text{ или } \lim_{n\to\infty} P(\left|\theta^* - \theta\right| < \varepsilon) = 1 \qquad (2.2)$$

Оценку, обладающую свойством (2.2), называют ***состоятельной***.

Если оценка θ^* параметра θ является несмещенной, а ее дисперсия

$$D(\theta^*) \to 0 \quad при \quad n \to \infty, \qquad (2.3)$$

то оценка θ^* является и состоятельной. Это непосредственно вытекает из неравенства Чебышева $P(\left|\theta^* - \theta\right| < \varepsilon) \geq 1 - \dfrac{D(\theta^*)}{\varepsilon^2}$

3. Если θ_1^* и θ_2^* – различные несмещённые оценки параметра θ, то оценка θ_1^* называется более эффективной, чем оценка θ_2^*, если

$$D_{\theta_1^*} < D_{\theta_2^*}. \qquad (2.4)$$

Поэтому разумно самой эффективной оценкой назвать оценку, на которой достигается $\min D$.

Эффективной называют статистическую оценку, которая при заданном объёме выборки n имеет наименьшую возможную дисперсию.

Оценки математического ожидания и дисперсии

Замечание. Статистический (вариационный) ряд содержит достаточно полную информацию об изучаемом признаке и его изменчивости (вариации). Однако обилие числовых данных всего ряда усложняет его использование. Поэтому целесообразно ввести некие обобщенные (сводные или средние) характеристики.

Расчет статистических характеристик — второй (после ранжирования) этап обработки данных наблюдения.

Пусть изучается дискретная генеральная совокупность относительно некоторого количественного признака ξ.

Генеральной средней $\overline{x}$ называют среднее арифметическое значений признака генеральной совокупности.

Если все значения $x_1, x_2, \ldots, x_N$ признака генеральной совокупности объёма N различны, то

$$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N}. \qquad (2.5)$$

Если же значения признака $x_1, x_2, \ldots, x_k$ имеют соответственно частоты $N_1, N_2, \ldots, N_k$, причём $\sum_{i=1}^{k} N_i = N$, то

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i N_i}{N}. \qquad (2.6)$$

т.е. ***генеральная средняя есть средняя взвешенная значений признака с весами, равными соответствующим частотам.***

Свойства средней арифметической ряда аналогичны свойствам математического ожидания для ДСВ (**УПР**)

Пусть для изучения *генеральной совокупности* относительного количественного признака ξ извлечена *выборка* объема n. Наиболее распространёнными оценками в математической статистике являются *выборочное среднее* $\overline{x}_в$ – оценка математического ожидания M_ξ, *выборочная дисперсия* $D_в$ – оценка дисперсии D_ξ, *выборочное среднеквадратичное отклонение* S – оценка среднеквадратичного отклонения σ.

Если все значения $x_1, x_2, \ldots, x_n$ признака ξ выборки объема n различны, то

$$\overline{x}_в = \frac{\sum_{i=1}^{n} x_i}{n}, \qquad (2.7)$$

т.е. ***выборочная средняя*** $\overline{x}_в$ есть среднее арифметическое значение признака выборочной совокупности. Если же значения признака $x_1, x_2, \ldots, x_k$ имеют соответственно частоты $n_1, n_2, \ldots, n_k$, причем $\sum_{i=1}^{k} n_i = n,$, то выборочным средним является величина

$$\overline{x}_в = \frac{\sum_{i=1}^{k} n_i x_i}{n}, \qquad (2.8)$$

т.е. *выборочная средняя есть средняя взвешенная значений признака с весами, равными соответствующим частотам.*

Покажем, что в качестве точечной оценки M^* для M_ξ – математического ожидания СВ ξ – может служить выборочное среднее $\overline{x}_в$, т. е. $M^* = \overline{x}_в$

Действительно, т.к. СВ $x_1, x_2, \ldots, x_n$ имеют один и тот же закон распределения, совпадающий с законом распределения СВ ξ, то

$$M(M^*) \overset{(2.7)}{=} M\frac{\sum_{i=1}^{n} x_i}{n} = \frac{1}{n}M\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n} M(x_i) = \frac{1}{n}\cdot nM_\xi = M_\xi,$$

т.е. оценка M^* для математического ожидания СВ ξ согласно (2.1) является несмещенной.

Рассмотрим дисперсию DM^*:

$$DM^* = D\overline{x}_в = D\frac{\sum_{i=1}^{n} x_i}{n} = \frac{1}{n^2}D\sum_{i=1}^{n} x_i = \frac{1}{n^2}\sum_{i=1}^{n} D(x_i) = \frac{1}{n^2}\cdot nD_\xi = \frac{D_\xi}{n},$$

где D_ξ – дисперсия СВ ξ. Так как $DM^* \to 0$ при $n \to \infty$, то из последнего равенства в силу (2.3) следует, что оценка M^* является состоятельной и несмещённой.

Заметим, что введенные выше средние не характеризуют степень изменчивости (вариации) изучаемого значения признака. Для этих целей вводится понятие ***дисперсии*** вариационного ряда.

По определению: $D_\xi = M[\xi - M_\xi]^2$.

Так как D_ξ есть математическое ожидание СВ вида $[\xi - M\xi]^2$, то естественной *оценкой* для D_ξ представляются выражения:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N} \qquad (2.9)$$

$$\sigma^2 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 N_i}{N}, \quad \sum_{i=1}^{k} N_i = N \qquad (2.9^*)$$

$$D_в = \frac{\sum_{i=1}^{n}(x_i - \overline{x}_в)^2}{n}. \qquad (2.10)$$

$$D_в = \frac{\sum_{i=1}^{k} n_i (x_i - \overline{x}_в)^2}{n}, \qquad \sum_{i=1}^{k} n_i = n. \qquad (2.10^*)$$

Статистическую оценку $D_в$ целесообразно использовать для оценки σ^2 дисперсии генеральной совокупности. Ее называют *выборочной дисперсией*.

Выборочная дисперсия $D_в$ ***есть среднее арифметическое квадратов отклонений наблюдаемых значений признака*** ξ ***от их выборочного среднего***.

Преобразуем формулу (2.10):

$$D_в = \frac{1}{n} \cdot \sum_{i=1}^{n}(x_i - \overline{x}_в)^2 = \frac{1}{n} \cdot \sum_{i=1}^{n}[(x_i - M_\xi) - (\overline{x}_в - M_\xi)]^2 =$$

$$= \frac{1}{n} \cdot \sum_{i=1}^{n}(x_i - M_\xi)^2 - 2 \cdot \frac{1}{n} \cdot (\overline{x}_в - M_\xi) \cdot \sum_{i=1}^{n}(x_i - M_\xi) + \frac{1}{n} \cdot n \cdot (\overline{x}_в - M_\xi)^2 =$$

$$= \frac{1}{n} \cdot \sum_{i=1}^{n}\left(x_i - M_\xi\right)^2 - 2 \cdot (\overline{x}_в - M_\xi)^2 + (\overline{x}_в - M_\xi)^2$$

$$= \frac{1}{n} \cdot \sum_{i=1}^{n}\left(x_i - M_\xi\right)^2 - (\overline{x}_в - M_\xi)^2.$$

Найдем математическое ожидание оценки $D_в$**:**

$$MD_в = \frac{1}{n}\cdot\sum_{i=1}^{n} M\left(x_i - M_\xi\right)^2 - M(\overline{x}_в - M_\xi)^2 = \frac{1}{n}\cdot n\cdot D_\xi - D_в =$$

$$= D_\xi - \frac{D_\xi}{n} = D_\xi - \frac{D_\xi}{n} = \frac{n-1}{n} D_\xi. \qquad (2.11)$$

Полученная оценка (2.11) является смещенной, т.к. $MD_в \neq D_\xi$, а именно: $MD_в = \frac{n-1}{n} D_\xi$.

Несмещенную оценку для D_ξ можно получить, если положить

$$D_\xi \approx \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_в)^2 = S^2. \qquad (2.12)$$

Действительно, тогда из (2.1) следует

$$MS^2 = \frac{1}{n-1} M\sum_{i=1}^{n}(x_i - \overline{x}_в)^2 = \frac{n}{n-1}\cdot M\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_в)^2 = \frac{n}{n-1} MD_в = \frac{n}{n-1}$$

Данная оценка является *несмещенной оценкой* дисперсии. Её называют *исправленной дисперсией* и обозначают

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_в)^2. \qquad (2.13)$$

Если значения $x_1, x_2, \ldots, x_k$ встречаются с частотами $n_1, n_2, \ldots, n_k$, то *исправленная выборочная дисперсия* имеет вид

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i (x_i - \overline{x}_в)^2 . \qquad (2.13^*)$$

Для оценки среднего квадратичного отклонения генеральной совокупности используют *исправленное среднее квадратичное отклонение*

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x}_в)^2}{n-1}}, \qquad (2.15)$$

которое, однако, ***не является несмещенной оценкой***.

Из сопоставления оценок дисперсии (2.10) и (2.13), (2.10*) и (2.13*) видно, что они отличаются лишь знаменателями. Очевидно, что при достаточно большом объеме выборки n выборочная дисперсия (2.10), (2.10*) и исправленная дисперсия (2.13), (2.13*) различаются незначительно. Исправленная дисперсия используется на практике при объеме выборки $n < 30$.

Для вычислений $D_в$ часто используется формула:

$$D_в = \overline{x_в^2} - (\overline{x}_в)^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n} \right)^2 = \frac{\sum_{i=1}^{k} n_i \cdot x_i^2}{n} - \left(\frac{\sum_{i=1}^{k} n_i x_i}{n} \right)^2 \quad (2.16)$$

К показателям вариации относят также $R = x_{\max} - x_{\min}$ – *размах вариации* и $v = \frac{\sigma_в}{\overline{x}_в} \cdot 100\ \%\ (\overline{x}_в \neq 0)$ – *коэффициент вариации*.

Коэффициент вариации – безразмерная характеристика. На практике считают, что если $v < 33\ \%$, то совокупность однородная.

Пример 1. Найти дисперсию и исправленную дисперсию по данному распределению выборки:

x_i	1	2	3	4
n_i	20	15	10	5

n=50

Решение. Вычислим сначала

$$\overline{x}_в = \frac{20 \cdot 1 + 15 \cdot 2 + 10 \cdot 3 + 5 \cdot 4}{50} = 2; \text{ и}$$

$$\overline{x_в^2} = \frac{20 \cdot 1 + 15 \cdot 4 + 10 \cdot 9 + 5 \cdot 16}{50} = 5.$$

Тогда

$$D_в = \overline{x_в^2} - (\overline{x}_в)^2 = 1; \quad S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{k} n_i \cdot (x_i - \overline{x}_в)^2 = \frac{50}{49}.$$

В случае, когда первоначальные варианты x_i – большие числа, то целесообразно вычесть из всех вариант одно и то же число c, равное $\overline{x}_в$ или близкое к нему, т.е. перейти к условным вариантам $u_i = x_i - c$. При этом $Du_i = Dx_i$.

$$D_в(u_i) = \overline{u_i^2} - (\overline{u}_i)^2 = \frac{\sum_{i=1}^{k} n_i u_i^2}{n} - \left(\frac{\sum_{i=1}^{k} n_i u_i}{n} \right)^2 \qquad (2.17)$$

$$Mu_i = Mx_i - Mc = \overline{x}_в - c. \qquad (2.18)$$

Таким образом,

$$\overline{x}_в = c + \frac{\sum_{i=1}^{k} u_i \cdot n_i}{n}. \qquad (2.19)$$

Пример 2. По данному распределению выборки

x_i	1250	1270	1280
n_i	2	5	3

n=10

найти выборочную среднюю.

Решение. Поскольку первоначальные варианты – большие числа, то перейдем к условным вариантам. Выберем в качестве c=1270, тогда новые варианты u_i вычисляются по формулам $u_i = x_i - 1270$. Распределение выборки в условных вариантах принимает вид

u_i	-20	0	10
n_i	2	5	3

а выборочное среднее равно

$$\overline{x}_в = 1270 + \frac{-20 \cdot 2 + 10 \cdot 3}{10} = 1269.$$

Пример 3. Имеется распределение 80 предприятий по числу работающих на них (чел.):

x_i	150	250	350	450	550	650	750
n_i	1	3	7	30	19	15	5 .

Найти числовые характеристики распределения предприятий по числу работающих.

Решение.

Признак X – число работающих (чел.) на предприятии.	Число предприятий (n_i)	$x_i\, n_i$	$(x_i - \overline{x}_в)^2 n_i$	$x_i^2\, n_i$
150	1	150		22500
250	3	750	129600	187500
350	7	2450		857500
450	30	13500	202800	6045000
550	19	10450		5747500
650	15	9750	179200	6337500
750	5	3750		2812500
			108000	
			30400	
			294000	
			288000	
Итого	80	40800		22040000 .
			1232000	

$$\overline{x}_в = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i = \frac{40800}{80} = 510 \text{ (чел.)} – \text{среднее число}$$

работающих на предприятии.

Дисперсию рассчитываем двумя способами.

1) по формуле (2.10)

$$D_{\text{в}} = \frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x}_{\text{в}})^2 \cdot n_i = \frac{123200}{80} = 15400.$$

2) по формуле (2.16)

$$D_{\text{в}} = \overline{x^2} - \overline{x}_{\text{в}}^2, \text{ где } \overline{x^2} = \frac{1}{n}\sum_{i=1}^{k} x_i^2 \cdot n_i = \frac{22040000}{80} = 275500.$$

$$D_{\text{в}} = 275500 - (510)^2 = 15400.$$

$\sigma_{в} = \sqrt{D_{\text{в}}} = \sqrt{15400} \approx 124$ (численность работающих на каждом предприятии отклоняется от средней численности в среднем на 124 чел.)

$R = x_{\max} - x_{\min} = 750 - 150 = 600$ (чел.).

$$v = \frac{\sigma_{в}}{\overline{x}_{\text{в}}} \cdot 100\,\% = \frac{124}{510} \cdot 100\,\% \approx 24{,}3\,\%.$$

Так как $v \approx 24{,}3\,\% < 33\%$, то исследуемая совокупность однородная.

Пример Найти числовые характеристики распределения затрат времени на обработку одной детали (пример 2).

Решение. Признак X – затраты времени на обработку одной детали (мин) – непрерывный. Распределение задано интервальным рядом. Характеристики такого ряда находят по тем же формулам, что и для дискретного ряда, предварительно заменив интервальный ряд дискретным. Для этого для каждого интервала x_{i-1}–x_i вычисляют его середину x_i'. Расчеты представим в таблице:

Затраты времени на обработку 1 детали (X, мин): x_{i-1}–x_i	Число рабочих (n_i)	x_i'	$x_i' n_i$	$(x_i' - \overline{x}_в)^2$	$(x_i')^2 n_i$
22–24	2	23		50	1058
24–26	12	25	46	108	7500
26–28	34	27		34	24786
28–30	40	29	300	40	33640
30–32	10	31		90	9610
32–34	2	33	918	50	2178
			1160		
			310		
			66		
Итого	100	-	2800	372	78772

$\overline{x}_в = \frac{1}{n}\sum_{i=1}^{k} x_i' \cdot n_i = \frac{2800}{100} =$ 28 (мин) – среднее время на обработку одной детали.

Дисперсию рассчитываем двумя способами.

1) $D_в = \frac{1}{n}\sum_{i=1}^{k} (x_i' - \overline{x}_в)^2 \cdot m_i = \frac{372}{100} = 3,72;$

2)

$D_в = \overline{(x')^2} - \overline{x}_в^2$, где $\overline{(x')^2} = \frac{1}{n}\sum_{i=1}^{k} (x_i')^2 \cdot m_i = \frac{78772}{100} = 787,72;$

$D_в = 787,72 - (28)^2 = 3,72.$

$\sigma_в = \sqrt{D_в} = \sqrt{3,72} \approx 1,93$ (мин), то есть затраты времени на обработку одной детали каждым рабочим отклоняются от средних затрат времени в среднем на 1,93 мин.

$R = x_{max} - x_{min} = 34 - 22 = 12$ (мин).

$v = \frac{\sigma_в}{\overline{x}_в} \cdot 100\,\% = \frac{1{,}93}{28} \cdot 100\,\% \approx 6{,}9\,\%$ – совокупность однородная.

Интервальное оценивание параметров (доверительные интервалы)

Выше был рассмотрен вопрос об оценке неизвестного параметра θ одним числом θ^*, т.е. о *точечной* оценке. В ряде задач требуется не только найти для параметра θ подходящее численное значение, но и оценить его *точность* и *надежность*. Требуется знать, к каким ошибкам может привести замена θ его точечной оценкой θ^*, и с какой степенью уверенности можно ожидать, что эти оценки не выйдут за известные пределы.

Такого рода задачи особенно актуальны при малом числе наблюдений, когда точечная оценка θ^* в значительной мере случайна и приближенная замена θ на θ^* может привести к серьезным ошибкам. Для определения точности и надежности θ^* в МС вводят понятие *доверительного интервала* и *доверительной вероятности*. Часто из физических соображений делается вывод, что ξ распределена по нормальному закону с плотностью вероятности

$$f_\xi = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}, \quad a = M_\xi, \quad \sigma = \sqrt{D\xi}.$$

Возникает задача оценки параметров a и σ или одного из них, если известны наблюдаемые значения x_1 , x_2 ,..., x_n СВ ξ.

Пусть для параметра θ из опыта получена несмещенная оценка θ^*.

Оценим возможную при этом ошибку. Назначим некоторую достаточно большую вероятность γ ($\gamma = 0{,}95;\ 0{,}99;\ 0{,}9$)

такую, что событие с вероятностью γ можно считать практически достоверным. Найдём такое значение ε, $\varepsilon > 0$, для которого вероятность отклонения оценки на величину, не превышающую ε, равна γ:

$$P\left(\left|\theta^* - \theta\right| < \varepsilon\right) = \gamma. \qquad (2.20)$$

Тогда диапазон практически возможных значений ошибки, возникающий при замене θ на θ^*, будет равен $\pm\varepsilon$. Большие по абсолютной величине ошибки будут появляться с малой вероятностью $\alpha = 1 - \gamma$.

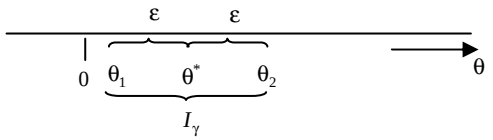
Перепишем уравнение (2.20) в виде:

$$P(\theta^* - \varepsilon < \theta < \theta^* + \varepsilon) = \gamma. \qquad (2.21)$$

Равенство (2.21) означает, что с вероятностью γ неизвестное значение параметра θ попадает в интервал I_γ, равный

$$I_\gamma = (\theta^* - \varepsilon; \theta^* + \varepsilon), \qquad (2.22)$$

который является случайным, т.к. случайным является центр θ^* интервала I_γ. Случайной является и его длина, равная 2ε, т.к. ε, как правило, вычисляется по опытным данным. Поэтому в (2.21) величину γ лучше толковать не как вероятность γ попадания точки θ в интервал I_γ, а как вероятность того, что случайный интервал I_γ накроет точку θ:



θ^* – центр доверительного интервала, $\theta_1 = \theta^* - \varepsilon$, $\theta_2 = \theta^* + \varepsilon$.

Вероятность γ принято называть *доверительной вероятностью* (надежностью), а интервал I_γ – *доверительным интервалом*.

Интервал (θ_1, θ_2) будем называть *доверительным* для оценки параметра θ при заданной доверительной вероятности γ или при заданном уровне значимости $\alpha = 1 - \gamma$, если он с вероятностью γ "накрывает" оцениваемый параметр θ, т.е.

$$P(\theta \in (\theta_1, \theta_2)) = P(\theta_1 < \theta < \theta_2) = \gamma. \qquad (2.23)$$

Границы интервала θ_1 и θ_2 называют *доверительными границами*. Доверительный интервал можно рассматривать как интервал значений параметра θ, совместимых с опытными данными и не противоречащих им. Метод доверительных интервалов был разработан Ю.Нейманом*, который использовал идеи Р.Фишера**.

Рассмотрим вопрос о нахождении доверительных границ θ_1 и θ_2. Пусть для параметра θ имеется несмещённая оценка θ^*. Если бы был известен закон распределения величины θ^*, задача нахождения доверительного интервала была бы весьма простой. Для этого достаточно было бы найти такое значение ε, для которого выполнено соотношение (2.20). Сложность состоит в том, что закон распределения оценки θ^* зависит от закона распределения СВ ξ, следовательно, от его неизвестных параметров, в частности, от параметра θ.

Доверительные интервалы для оценки неизвестного математического ожидания нормального распределения при известном σ

Пусть количественный признак ξ генеральной совокупности распределен нормально, причем известно σ – среднее квадратичное отклонение этого распределения. Оценим неизвестное математическое ожидание a по выборочной средней $\overline{x}_в$, т.е. найдем доверительные интервалы, покрывающие параметр a с надежностью γ. Будем рассматривать $\overline{x}_в$ как СВ $\overline{\xi}$ (т.е. $\overline{x}_в$ меняется от

* Ю.Нейманом (1894-1981) – американский математик-статистик
** Р.Фишера (1890-1962) – английский статистик и генетик

выборки к выборке), а выборочные значения признака $x_1, x_2, \dots x_n$ – как одинаково распределенные (т.е. имеющие одну и ту же функцию распределения $F(x)$) независимые в совокупности случайные величины $\xi_1, \xi_2, \dots \xi_n$ (эти числа также изменяются от выборки к выборке). Тогда математические ожидания каждой из величин $x_1, x_2, \dots x_n$ одинаковы и равны a, т.е. $M_{x_i} = a, \quad \sigma_{x_i} = \sigma, \ i = \overline{1,n}.$

Известно, что если СВ ξ распределена нормально, то выборочная средняя также распределена нормально и

$$M_{\overline{\xi}} = a, \ D_{\overline{\xi}} = \frac{\sigma^2}{n}, \ \sigma_{\overline{\xi}} = \frac{\sigma}{\sqrt{n}}.$$

Потребуем выполнение соотношения $P\left(\left|\xi - a\right| < \delta\right) = \gamma$, где γ – заданная надежность. Поскольку для нормально распределенной СВ ξ $P(\left|\xi - a\right| < \delta) = \Phi(\frac{\delta}{\sigma})$, то, сделав замену $\xi \to \overline{\xi}, \quad \sigma \to \sigma_{\overline{\xi}} = \frac{\sigma}{\sqrt{n}},$ получим

$$P(|\overline{\xi} - a| < \delta) = \Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right) = \Phi(t_\gamma),$$

где $t_\gamma = \frac{\delta\sqrt{n}}{\sigma}.$

Следовательно, $\boxed{\delta = \frac{t_\gamma \sigma}{\sqrt{n}}.}$

Таким образом, $P(|\overline{\xi} - a| < \frac{t_\gamma\sigma}{\sqrt{n}}) = \Phi(t_\gamma)$. Так как вероятность задана и равна γ, то, заменив $\overline{\xi}$ на $\overline{x}_в$, получим

$$P\left(\overline{x}_в - \frac{t_\gamma\sigma}{\sqrt{n}} < a < \overline{x}_в + \frac{t_\gamma\sigma}{\sqrt{n}}\right) = \Phi\left(t_\gamma\right) = \gamma. \qquad (2.24)$$

Таким образом, с доверительной вероятностью γ (надежностью γ) можно утверждать, что доверительный интервал $\left(\overline{x}_в - \frac{t_\gamma \sigma}{\sqrt{n}}; \overline{x}_в + \frac{t_\gamma \sigma}{\sqrt{n}}\right)$ покрывает неизвестное математическое ожидание a нормально распределенной СВ с известным среднеквадратичным отклонением σ с точностью $\delta = \frac{t_\gamma \sigma}{\sqrt{n}}$. Число t_γ определяется из соотношения $\Phi(t_\gamma) = \gamma$, где $\Phi(x)$ – функция Лапласа. По таблице значений функции Лапласа находим аргумент t_γ, которому соответствует значение функции Лапласа, равное γ.

Пример 3. СВ ξ распределена нормально с $\sigma = 3$. Найти доверительный интервал для оценки неизвестного математического ожидания a по выборочной средней $\overline{x}_в$, если объем выборки $n = 36$ и задана надежность оценки $\gamma = 0.95$.

Решение. Вычислим $\Phi\left(t_\gamma\right) = \gamma = 0.95$. По таблице значений функции Лапласа находим $t_\gamma = 1.96$; следовательно, точность оценки $\delta = \quad = \frac{t_\gamma \sigma}{\sqrt{n}} = \frac{1.96 \cdot 3}{6} = 0.98$. Тогда доверительный интервал $(\overline{x}_в - \delta, \overline{x}_в + \delta)$ имеет вид $(\overline{x}_в - 0.98, , \overline{x}_в + 0.98)$. Таким образом, с вероятностью 0.95 СВ ξ попадет в интервал $\left(\overline{x}_в - 0.98, \overline{x}_в + 0.98\right)$.

Смысл заданной надежности: надежность $\gamma = 0.95$ означает, что если проведено достаточно большое число выборок, то 95% из них определяют такие доверительные интервалы, в которых действительно заключен параметр a; в 5% случаев параметр a может выйти за границы доверительного интервала.

Пример 4. С целью определения среднего трудового стажа на предприятии методом случайной повторной выборки проведено обследование трудового стажа рабочих. Из всего коллектива рабочих завода случайным образом выбрано 400 рабочих, данные о трудовом стаже которых и составили выборку. Средний по выборке стаж оказался равным 9,4 года. Считая, что трудовой стаж рабочих имеет нормальный закон распределения, определить с вероятностью 0,97 границы, в которых окажется средний трудовой стаж для всего коллектива, если известно, что $\sigma = 1{,}7$ года.

Решение. Признак X – трудовой стаж рабочих. Этот признак имеет нормальный закон распределения с известным параметром $\sigma = 1{,}7$, параметр a неизвестен. Сделана выборка объемом $n = 400$, по данным выборки найдена точечная оценка параметра a: $\overline{x}_{\text{в}} = 9{,}4$. С надежностью $\gamma = 0{,}97$ найдем интервальную оценку параметра a по формуле:

$$\overline{x}_{\text{в}} - \frac{t \cdot \sigma}{\sqrt{n}} < a < \overline{x}_{\text{в}} + \frac{t \cdot \sigma}{\sqrt{n}}.$$

По таблице значений функции Лапласа $\Phi(t) \approx \frac{0{,}97}{2} = 0{,}485$ находим $t = 2{,}17$; тогда:

$$9{,}4 - \frac{2{,}17 \cdot 1{,}7}{\sqrt{400}} < a < 9{,}4 + \frac{2{,}17 \cdot 1{,}7}{\sqrt{400}},$$

$9{,}4 - 0{,}18 < a < 9{,}4 + 0{,}18$. Итак, $9{,}22 < a < 9{,}58$, то есть средний трудовой стаж рабочих всего коллектива лежит в пределах от 9,22 года до 9,58 года (с надежностью $\gamma = 0{,}97$).

С изменением надежности γ изменится и интервальная оценка.

Пусть $\gamma = 0{,}99$, тогда $\Phi(t) = 0{,}495$, отсюда $t = 2{,}58$. Тогда:

$$9{,}4 - \frac{2{,}58 \cdot 1{,}7}{20} < a < 9{,}4 + \frac{2{,}58 \cdot 1{,}7}{20}, \text{ или } 9{,}4 - 0{,}22 < a < 9{,}4 + 0{,}22.$$

Окончательно: $9{,}18 < a < 9{,}62$.

Доверительные интервалы для оценки математического ожидания нормального распределения при неизвестном σ

Для дальнейшего понадобятся новые специальные распределения СВ ξ: распределение χ^2 и t-распределение (*распределение Стьюдента*).

Распределение χ^2

Пусть ξ_i, $i=\overline{1,n}$, нормально распределенные независимые СВ, причем $M_{\xi_i}=0$, а $\sigma_{\xi_i}=1$, т.е. СВ ξ_i являются *нормированными*. Тогда говорят, что сумма квадратов этих величин

$$\chi^2=\sum_{i=1}^{n}\xi_i^{\,2} \qquad (2.25)$$

распределена по закону χ^2 с $k=n$ степенями свободы. Если же ξ_i, $i=\overline{1,n}$, связаны некоторым линейным соотношением, например, $\sum_{i=1}^{n}\xi_i^{\,2}=n\overline{x}_в$, то число степеней свободы $k=n$-1. Плотность распределения χ^2 равна

$$f_{\chi^2}(x)=\begin{cases}0, & x\le 0,\\ \dfrac{1}{2^{k/2}\,\Gamma\left(k/2\right)}e^{-x/2}\cdot x^{k/2-1}, & x>0,\end{cases}$$

где $\Gamma(x)=\int_0^{+\infty}t^{x-1}e^{t}dt$ – гамма-функция Эйлера. В частности, для $n\in\mathbb{N}$ $\Gamma(n+1)=n!$. Таким образом, распределение χ^2 определяется одним параметром – числом степеней свободы k. С увеличением числа степеней свободы k распределение χ^2 *медленно* приближается к нормальному распределению.

*Распределение Стьюдента**

Пусть ξ – СВ, распределенная по нормальному закону, причем $M_\xi = 0$, $\sigma_\xi = 1$, т.е. ξ является нормированной СВ, а v – независимая от ξ СВ, распределенная по закону χ^2 с k степенями свободы. Тогда СВ

$$\tau = \frac{\xi}{\sqrt{v/k}} \qquad (2.26)$$

имеет распределение, называемое *t-распределением* или *распределением Стьюдента* с k степенями свободы. Таким образом, *отношение нормированной нормальной величины* ξ *к квадратному корню от независимой СВ, распределенной по закону* χ^2 *с k степенями свободы, деленой на k, распределено по закону Стьюдента с k степенями свободы*. С возрастанием числа степеней свободы k распределение Стьюдента *быстро* приближается к нормальному.

Пусть количественный признак ξ генеральной совокупности распределен по нормальному закону, причем σ неизвестно. Требуется оценить неизвестное математическое ожидание a с помощью доверительных интервалов с заданной доверительной вероятностью (надежностью) γ.

Воспользоваться результатами предыдущего раздела нельзя, т.к. параметр σ в данном случае неизвестен. Для решения задачи по выборке $x_1, x_2, \ldots x_n$ вычислим $\overline{x}_в$ и исправленную дисперсию S^2. По данным выборки построим СВ T:

$$T = \frac{(\overline{x}_в - a)\sqrt{n}}{S}, \qquad (2.27)$$

где S – исправленное среднее квадратичное отклонение, которое не является несмещенной оценкой, n – объем выборки.

---

* Стьюдент-псевдоним английского статистика В. Госсета (1876-1937)

В математической статистике доказывается, что величина (статистика) (2.27) имеет распределение Стьюдента с $(n-1)$ степенью свободы.

Плотность распределения Стьюдента задается формулой $S_{n-1}(x)=b_n(1+\frac{x^2}{n})^{-\frac{n+1}{2}}$, где b_n находится из условия нормировки. Нетрудно видеть, что используя второй замечательный предел, легко получить $(1+\frac{x^2}{n})^{-\frac{n+1}{2}} \underset{n\to+\infty}{\to} e^{-\frac{x^2}{2}}$.

Возможные значения СВ τ будем обозначать через *t*. Можно доказать, что СВ τ имеет распределение Стьюдента с $k=n-1$ степенями свободы. Плотность *S(t,n)* распределения Стьюдента имеет вид:

$$S(t,n)=b_{n+1}\left(1+\frac{t^2}{n-1}\right)^{-\frac{n}{2}} \qquad (2.28)$$

где коэффициент b_{n+1} выражается через гамма-функцию Γ(x):

$$b_{n+1}=\frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi(n-1)}\Gamma\left(\frac{n-1}{2}\right)}.$$

Из (2.28) видно, что распределение Стьюдента определяется параметром *n* – объемом выборки (или, что то же самое, числом степеней свободы $k=n-1$) и не зависит от неизвестных параметров a и σ. Такая особенность распределения Стьюдента является его большим достоинством. Из (2.24) также следует, что $S(t,n)$ – четная функция переменной *t*. Следовательно, вероятность осуществления неравенства $\left|\frac{(\overline{x}_в-a)\sqrt{n}}{S}\right|<t_{\gamma,n}$ для четной

функции плотности распределения $f_\xi(x)$ в силу соотношения $P(|\xi|<\delta)=2\int_0^\delta f_\xi(x)dx$ определяется по формуле

$$P\left(\left|\frac{\overline{x}_в-a}{S/\sqrt{n}}\right|<t_{\gamma,n}\right)=2\int_0^{t_{\gamma,n}}S(t,n)\,dt=\gamma,$$

откуда следует

$$P\left(\overline{x}_в-\frac{t_{\gamma,n}S}{\sqrt{n}}<a<\overline{x}_в+\frac{t_{\gamma,n}S}{\sqrt{n}}\right)=\gamma. \qquad (2.29)$$

Таким образом, пользуясь распределением Стьюдента, найден доверительный интервал $\left(\overline{x}_в-\frac{t_{\gamma,n}S}{\sqrt{n}};\ \overline{x}_в+\frac{t_{\gamma,n}S}{\sqrt{n}}\right)$, покрывающий неизвестный параметр a с надежностью γ. По таблице распределения Стьюдента по заданным n и γ можно найти $t_{\gamma,n}$ $(t_{\gamma,n}=t(\gamma,n))$.

Пример 5. Количественный признак ξ генеральной совокупности распределен нормально. По выборке объема n=16 найдены $\overline{x}_в=20.2$ и исправленное среднее квадратичное отклонение $S=0.8$. Оценить неизвестное математическое ожидание a при помощи доверительного интервала с надежностью γ=0.95.

Решение. Для нахождения доверительного интервала следует по таблице t-распределения Стьюдента по n=16, γ=0.95 найти $t_{\gamma,n}$=2.13. Тогда границы доверительного интервала имеют вид

$$\overline{x}_в-t_{\gamma,n}\frac{S}{\sqrt{n}}=20.2-\frac{2.13\cdot0.8}{4}=19.774;\ \overline{x}_в+t_{\gamma,n}\frac{S}{\sqrt{n}}$$

$$=20.2+\frac{2.13\cdot0.8}{4}=20.626.$$

Таким образом, с надежностью γ=0.95 неизвестный параметр a заключается в доверительном интервале (19.774, 20.626).

Пример 6. С целью определения средней продолжительности рабочего дня на предприятии методом случайной повторной выборки проведено обследование продолжительности рабочего дня сотрудников. Из всего коллектива завода случайным образом выбрано 30 сотрудников. Данные табельного учета о продолжительности рабочего дня этих сотрудников и составили выборку. Средняя по выборке продолжительность рабочего дня оказалась равной 6,85 часа, а S = 0,7 часа. Считая, что продолжительность рабочего дня имеет нормальный закон распределения, с надежностью γ= 0,95 определить, в каких пределах находится действительная средняя продолжительность рабочего дня для всего коллектива данного предприятия.

Решение. Признак X – продолжительность рабочего дня. Признак имеет нормальное распределение с неизвестными параметрами. Сделана выборка объемом n = 30, по выборочным данным найдены точечные оценки параметров распределения: $\overline{x}_{\text{в}}$ = 6,85; S = 0,7. С надежностью γ= 0,95 найдем интервальную оценку параметра a по формуле:

$$\overline{x}_{\text{в}} - \frac{t_{\gamma,n} \cdot S}{\sqrt{n}} < a < \overline{x}_{\text{в}} + \frac{t_{\gamma,n} \cdot S}{\sqrt{n}},$$

$t_{\gamma,n}$ находим по таблице t-распределения Стьюдента $t_{\gamma,n}$ = t(0,95; 30)=2,045. Тогда:

$$6{,}85 - \frac{2{,}045 \cdot 0{,}7}{\sqrt{30}} < a < 6{,}85 + \frac{2{,}045 \cdot 0{,}7}{\sqrt{30}},$$

или $6{,}85 - 0{,}26 < a < 6{,}85 + 0{,}26$.

Итак, $6{,}59 < a < 7{,}11$, то есть с надежностью γ= 0,95 средняя продолжительность рабочего дня для всего коллектива лежит в пределах от 6,59 до 7,11 ч.

Определение объема выборки

Для определения необходимого объема выборки, при котором с заданной вероятностью γ можно утверждать, что выборочная средняя отличается от генерального по абсолютной величине меньше чем на δ, пользуются формулами:

а) в случае известной дисперсии из формулы (2.24):

$$n = \frac{t_\gamma{}^2 \sigma^2}{\delta^2}, \qquad (2.30)$$

где $\Phi\left(t_\gamma\right) = \gamma$.

б) в случае неизвестной дисперсии организуют специальную «пробную» выборку небольшого объема, находят оценку S^2 и, полагая $\sigma^2 \approx S^2$, находят объем «основной» выборки:

$$n = \frac{t_\gamma{}^2 S^2}{\delta^2}, \qquad (2.31)$$

Пример 7. Найти минимальный объем выборки, на основе которой можно было бы оценить математическое ожидание СВ с ошибкой, которая не превышает 0.2 и надежностью 0.98, если допускается что СВ имеет нормальное распределение с $\sigma = 4$.

Решение. Из равенства $\Phi(t_\gamma) = 0.98$ по таблице определяют $t_\gamma = 2.33$. По формуле (2.30) находим:

$$n = \frac{t_\gamma{}^2 \sigma^2}{\delta^2} = \frac{2.33^2 \cdot 16}{0.2^2} \approx 2171.$$

Проверка статистических гипотез

С теорией статистического оценивания параметров тесно связана проверка статистических гипотез. Она используется в том случае, когда необходим обоснованный вывод о преимуществах того или иного способа вложения инвестиций, об уровне доходности ценных бумаг, об эффективности лекарственных препаратов, о значимости построенной математической модели и т.д.

При изучении многих статистических данных необходимо знать *закон* распределения генеральной совокупности. Если закон распределения неизвестен и есть основания предположить, что он имеет определенный вид (например, A), то выдвигают гипотезу: генеральная совокупность распределена по закону A. В данной гипотезе речь идет о *виде* предполагаемого распределения.

Возможен случай, когда закон распределения известен, а его параметры неизвестны. Если есть основания предположить, что неизвестный параметр θ равен определенному значению θ_0, то выдвигают гипотезу: $\theta = \theta_0$. Здесь речь идет о *предполагаемой величине параметра* одного известного распределения. Возможны гипотезы о равенстве параметров двух или нескольких распределений, о независимости выборок и др.

Все выводы, которые делаются в МС, вообще говоря, являются гипотезами, т.е. предположениями о неизвестных параметрах известных распределений, об общем виде неизвестного теоретического распределения или функции распределения изучаемой СВ. Такие гипотезы называют *статистическими гипотезами*.

Различают *простые* и *сложные*, *параметрические* и *непараметрические* статистические гипотезы.

Статистическая гипотеза называется *простой*, если она однозначно определяет закон распределения СВ. *Сложной* называют гипотезу, состоящую из конечного или

бесконечного числа простых гипотез. Например, гипотезы "вероятность появления события A в схеме Бернулли равна $0,5$", "закон распределения СВ – нормальный с параметрами $a=0,\ \sigma^2=1$" являются *простыми* в отличие от *сложных* гипотез: "вероятность появления события A в схеме Бернулли заключена между $0,3$ и $0,5$", "закон распределения СВ не является нормальным".

Гипотеза называется *параметрической*, если в ней содержится некоторое условие о значении параметра известного распределения. Гипотезу, в которой сформулированы предположения относительно вида распределения, называют *непараметрической*.

Если исследовать всю генеральную совокупность, то, естественно, можно было бы наиболее точно установить справедливость выдвигаемой гипотезы. Однако такое исследование не всегда возможно, и суждение об истинности статистических гипотез проверяется на основании выборки.

Выдвигаемую (проверяемую) гипотезу называют *основной* или *нулевой* гипотезой H_0. Если, например, по полигону или гистограмме частот, построенным по некоторой выборке, можно предположить, что СВ распределена по нормальному закону, то может быть выдвинута гипотеза $H_0: a=a_0, \sigma=\sigma_0$. Одновременно с гипотезой H_0 выдвигается *альтернативная* (*конкурирующая*) гипотеза H_1. Если гипотеза H_0 будет отвергнута, то имеет место конкурирующая ей гипотеза.

Конкурирующей (*альтернативной*) называют гипотезу H_1, являющуюся логическим отрицанием H_0. *Нулевая* H_0 и *альтернативная* H_1 гипотезы представляют собой две возможности выбора, осуществляемого в задачах проверки статистических гипотез. Например, если $H_0: \theta=\theta_0$, то альтернативная гипотеза может иметь вид $H_1: \theta \neq \theta_0$, $H_1: \theta > \theta_0$, или $H_1: \theta < \theta_0$.

Выдвинутая гипотеза может быть правильной или неправильной, в связи с чем возникает необходимость ее проверки. Поскольку проверку осуществляют статистическими методами, ее называют *статистической.*

В результате статистической проверки гипотезы неправильное решение может быть принято в двух случаях: с одной стороны, на основании результатов опыта можно отвергнуть правильную гипотезу; с другой – можно принять неверную гипотезу. Очевидно, последствия этих ошибок могут оказаться различными. Отметим, что правильное решение может быть принято также в двух случаях:

1) гипотеза принимается, и она в действительности является правильной;
2) гипотеза отвергается, и она в действительности не верна.

По полученным значениям статистики основная гипотеза принимается или отклоняется. При этом в виду случайного характера выборки, могут быть допущены два вида ошибок:

–может быть отвергнута правильная гипотеза, в этом случае допускается *ошибка первого рода*,

–может быть принята неверная гипотеза, тогда допускается *ошибка второго рода* (см. схему).

H_0	H_0 – принимается	H_0 – отвергается
верна	правильное решение	ошибка I рода
ошибочна	ошибка II рода	правильное решение

Вероятность α совершить ошибку I рода, т.е. отвергнуть гипотезу H_0, когда она верна, называется *уровнем значимости* критерия.

Обычно принимают $\alpha = 0.1, 0.05, \cdots, 0.01$. Смысл α: при $\alpha = 0.05$ в 5 случаях из 100 имеется риск допустить ошибку I рода, т.е. отвергнуть правильную гипотезу.

Вероятность допустить ошибку II рода, т.е. принять гипотезу H_0, когда она неверна, обозначают β.

Вероятность $1-\beta$ не допустить ошибку II рода, т.е. отвергнуть гипотезу H_0, когда она ошибочна, называется *мощностью критерия*.

Используя терминологию статистического контроля качества продукции можно сказать, что вероятность α представляет "*риск поставщика*" (или *"риск производителя"*), связанный с вероятностью признать негодной по результатам выборочного контроля всю партию годных изделий, удовлетворяющих стандарту, а вероятность β – "*риск потребителя*", связанный с вероятностью принять по анализу выборки негодную партию, не удовлетворяющую стандарту. В некоторых прикладных исследованиях ошибка I рода α означает вероятность того, что сигнал, предназначенный наблюдателю, не будет принят, а ошибка II рода β – вероятность того, что наблюдатель примет ложный сигнал.

Для проверки справедливости нулевой гипотезы H_0 используют специально подобранную СВ K, точное или приближенное распределение которой известно. Эту СВ K, которая служит для проверки нулевой гипотезы H_0, называют *статистическим критерием* (или просто *критерием*).

Для проверки статистической гипотезы по данным выборок вычисляют частные значения входящих в критерий величин и получают частное (*наблюдаемое*) значение критерия $K_{набл}$.

После выбора определенного статистического критерия для решения вопроса о принятии или непринятии гипотезы множество его возможных значений разбивают на два непересекающихся подмножества, одно из которых называется *областью принятия гипотезы* (или *областью допустимых значений критерия*), а второе – *критической областью*.

Критической областью называется совокупность значений статистического критерия K, при которых нулевую гипотезу H_0 отвергают.

Областью принятия гипотезы (*областью допустимых значений критерия)* называется совокупность значений статистического критерия K, при которых нулевую гипотезу H_0 принимают.

Если наблюдаемое значение $K_{набл}$ статистического критерия K принадлежит критической области, то основная гипотеза отвергается в пользу альтернативной; если оно принадлежит области принятия гипотезы, то гипотезу принимают.

Поскольку статистический критерий K – одномерная СВ, то все ее возможные значения принадлежат некоторому интервалу. Следовательно, и критическая область, и область принятия гипотезы – также интервалы. Тогда должны существовать точки, их разделяющие.

Критическими точками (границами) $k_{кр}$ называют точки, отделяющие критическую область от области принятия гипотезы.

В отличие от рассмотренного интервального оценивания параметров, в котором имелась лишь одна возможность ошибки – получение доверительного интервала, не накрывающего оцениваемый параметр – при проверке статистических гипотез возможна двойная ошибка (как I рода α, так и II рода β). Вероятности оценок I и II рода (α и β) однозначно определяются выбором критической области. Естественным является желание сделать α и β сколь угодно малыми. Однако эти требования являются противоречивыми, ибо при фиксированном объеме выборки можно сделать сколь угодно малой лишь одну из величин – α или β, что сопряжено с неизбежным увеличением другой.

Одновременное уменьшение вероятностей α и β возможно лишь при увеличении объема выборки.

Поскольку одновременное уменьшение ошибок I и II рода невозможно, то при нахождении критических областей для данной статистики уровень значимости задают, стараясь подобрать такой критерий, чтобы вероятность ошибки II рода была наименьшей.

Различают *одностороннюю* (правостороннюю и левостороннюю) и *двустороннюю* критические области.

Правосторонней называют критическую область, определяемую неравенством $K > k_{кр}$, где $k_{кр} > 0$.

Левосторонней называют критическую область, определяемую неравенством $K < k_{кр}$, где $k_{кр} < 0$.

Двусторонней называют критическую область, определяемую неравенствами $K < k_1, K > k_2$, где $k_2 > k_1$.

Если критические точки симметричны относительно нуля, то двусторонняя критическая область определяется неравенствами $K < -k_{кр}$, $K > k_{кр}$, где $k_{кр} > 0$ или, что равносильно, $|K| > k_{кр}$.

Как найти критическую область?

Пусть $K = K(x_1, x_{2,} \ldots, x_n)$ – статистический критерий, выбранный для проверки нулевой гипотезы H_0, k_0 – некоторое число, $k_0 \in R$. Найдем правостороннюю критическую область, определяемую неравенством $K > k_{кр}$, где $k_{кр} > 0$. Для ее отыскания достаточно найти критическую точку $k_{кр}$. Рассмотрим вероятность $P(K > k_0)$ в предположении, что гипотеза H_0 верна. Очевидно, что с ростом k_0 вероятность $P(K > k_0)$ уменьшается. Тогда k_0 можно выбрать настолько большим, что вероятность $P(K > k_0)$ станет ничтожно малой. Другими словами, при

заданном уровне значимости α можно определить критическое значение $k_{кр}$ из неравенства $P(K > k_{кр}) = \alpha$.

Критическую точку $k_{кр}$ ищут из требования, чтобы при условии справедливости нулевой гипотезы H_0 вероятность того, что критерий K примет значение, большее $k_{кр}$, была равна принятому уровню значимости α:

$$P(K > k_{кр}) = \alpha. \qquad (3.1)$$

Для каждого из известных статистических критериев (нормального, Стьюдента, критерия Пирсона χ^2, Фишера-Снедекора, Кочрена и др.) имеются соответствующие таблицы, по которым находят $k_{кр}$, удовлетворяющее этим требованиям.

После нахождения $k_{кр}$ по данным выборок вычисляют реализовавшееся (наблюдаемое) значение $K_{набл}$ критерия K. Если окажется, что $K_{набл} > k_{кр}$, (т.е. реализовалось маловероятное событие), то нулевая гипотеза H_0 отвергается. Следовательно, принимается конкурирующая гипотеза H_1. Если же $K_{набл} < k_{кр}$, то в этом случае нет оснований отвергнуть выдвинутую гипотезу H_0. Следовательно, гипотеза H_0 принимается. Другими словами, *выдвинутая статистическая гипотеза согласуется с результатами эксперимента* (выборочными данными).

Левосторонняя критическая область определяется неравенством $K < k_{кр}$, где $k_{кр} < 0$. *Критическую точку $k_{кр}$ находят из требования, чтобы при условии справедливости нулевой гипотезы H_0 вероятность того, что критерий K примет значение, меньшее $k_{кр}$, была равна принятому уровню значимости α*:

$$P(K < k_{кр}) = \alpha. \qquad (3.2)$$

Двусторонняя критическая область определяется неравенствами $K < k_1, K > k_2$, где $k_2 > k_1$. *Критические точки* k_1, k_2 *находят из требования, чтобы при условии справедливости нулевой гипотезы* H_0 *сумма вероятностей того, что критерий* K *примет значение, меньшее* k_1 *или большее* k_2*, была равна принятому уровню значимости* α*:*

$$P(K < k_1) + P(K > k_2) = \alpha. \qquad (3.3)$$

Если распределение критерия симметрично относительно нуля, и для увеличения его мощности выбрать симметричные относительно нуля точки $-k_{кр}$ и $k_{кр}$, $k_{кр} > 0$, то $P(K < -k_{кр}) = P(K > k_{кр})$, и из $P(K < k_1) + P(K > k_2) = \alpha$ следует

$$P(K > k_{кр}) = \alpha/2. \qquad (3.4)$$

Это соотношение и служит для отыскания критических точек двусторонней критической области.

Отметим, что *принцип проверки статистической гипотезы не дает логического доказательства ее верности или неверности. Принятие гипотезы* H_0 *следует расценивать не как раз и навсегда установленный, абсолютно верный содержащийся в ней факт, а лишь как достаточно правдоподобное, не противоречащее опыту утверждение.*

Наиболее распространенным критерием проверки статистических гипотез о виде распределения генеральной совокупности (т.е. непараметрическим критерием) является *критерий Пирсона* χ^2.

Проверка гипотез о среднем значении нормально распределенной СВ при известной и неизвестной дисперсии

Пусть имеется генеральная совокупность X, распределенная по нормальному закону с *известной дисперсией* $D(X)=\sigma^2$.

Генеральная средняя a неизвестна, но есть основания предполагать, что она равна гипотетическому (предполагаемому) значению a_0.

Например, если X – совокупность размеров x_i партии деталей, изготавливаемых станком-автоматом, то можно предполагать, что генеральная средняя a этих размеров равна проектному размеру a_0.

Для проверки этого предположения (гипотезы) делают выборку, находят $\overline{x}_в$ и устанавливают, *значимо* или *незначимо* различаются $\overline{x}_в$ и a_0. Если различие окажется незначимым, то станок в среднем обеспечивает проектный размер; если же различие значимое, то станок требует наладки.

Из нормальной генеральной совокупности X извлечем выборку $x_1,\ldots,x_n$ объема n, по которой найдем $\overline{x}_в$. При этом дисперсия σ^2 известна. Поскольку предполагается, что $x_1,\ldots,x_n$ как СВ $X_1,\ldots,X_n$ взаимно независимы, то они имеют одинаковые нормальные распределения, а следовательно, и одинаковые характеристики (математическое ожидание, дисперсию, и т.д.).

Необходимо по известному $\overline{x}_в$ при заданном уровне значимости α проверить гипотезу $H_0: a=a_0$ о равенстве генеральной средней a гипотетическому значению a_0.

Поскольку $\overline{x}_в$ является несмещенной оценкой генеральной средней, т.е. $M(\overline{x}_в)=a$, то гипотезу $H_0: a=a_0$ можно записать в виде $H_0: M(\overline{x}_в)=a_0$. Таким образом, требуется проверить, что математическое ожидание выборочной средней $\overline{x}_в$ равно гипотетической генеральной средней a_0, т.е. *значимо* или *незначимо* различаются выборочная $\overline{x}_в$ и генеральная a_0 средние.

В качестве *критерия проверки* гипотезы H_0 примем СВ

$$U = \frac{\bar{X} - a_0}{\sigma(\bar{X})}.$$

В силу свойства $\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ для одинаково распределенных взаимно независимых СВ критерий проверки гипотезы H_0 принимает вид

$$U = \frac{\bar{X} - a_0}{\sigma} \cdot \sqrt{n}.$$

Случайная величина U распределена по стандартному нормальному закону с $a = 0, \sigma = 1$. Критическая область строится в зависимости от вида конкурирующей гипотезы H_1. Сформулируем правила проверки гипотезы H_0, обозначив через $U_{набл}$ значение критерия U, вычисленное по данным наблюдений.

Правило 1. Для того чтобы при заданном уровне значимости α проверить гипотезу $H_0 : a = a_0$ о равенстве неизвестной генеральной средней a нормальной совокупности с известной дисперсией σ^2 гипотетическому значению a_0 при конкурирующей гипотезе $H_1 : a \neq a_0$, необходимо вычислить

$$U_{набл} = \frac{\bar{x}_в - a_0}{\sigma} \cdot \sqrt{n} \qquad (3.5)$$

и по таблице значений функции Лапласа найти критическую точку *двусторонней критической области* из равенства

$$\Phi\left(u_{кр}\right) = 1 - \alpha. \qquad (3.6)$$

Если $\left|U_{набл}\right| < u_{кр}$ – нет оснований отвергнуть гипотезу H_0; если $\left|U_{набл}\right| > u_{кр}$ – гипотезу H_0 отвергают.

Правило 2. При конкурирующей гипотезе $H_1 : a > a_0$ критическую точку $u_{кр}$ *правосторонней критической области* находят из равенства

$$\Phi\left(u_{кр}\right) = 1 - 2\alpha. \qquad (3.7)$$

Если $U_{набл} < u_{кр}$ – нет оснований отвергнуть гипотезу H_0; если $U_{набл} > u_{кр}$ – гипотезу H_0 отвергают.

Правило 3. При конкурирующей гипотезе $H_1: a < a_0$ критическую точку $u_{кр}$ находят по правилу 2, а затем полагают границу *левосторонней критической области* $u'_{кр} = -u_{кр}$. Если $U_{набл} > -u_{кр}$ – нет оснований отвергнуть гипотезу H_0; если $U_{набл} < -u_{кр}$ – гипотезу H_0 отвергают.

Замечание. Из правила 1 следует, что если область принятия гипотезы H_0 есть интервал $-u_{кр} < U_{набл} < u_{кр}$, то область ее отклонения – $U \in (-\infty; u_{кр}) \cup (u_{кр}; +\infty)$

Пример 1. Из нормальной генеральной совокупности с известным $\sigma = 0.49$ извлечена выборка объема $n = 49$ и по ней найдено выборочное среднее $\overline{x}_в = 21.7$. При уровне значимости $\alpha = 0.05$ проверить гипотезу $H_0: a = a_0 = 21$ при конкурирующей гипотезе $H_1: a > 21$.

Решение. По данным задачи найдем

$$U_{набл} = \frac{\overline{x}_в - a_0}{\sigma} \cdot \sqrt{n} = \frac{21.7 - 21}{0.49} \cdot \sqrt{49} = 10.$$

Поскольку конкурирующая гипотеза H_1 имеет вид $H_1: a > 21$, то критическая область – правосторонняя. По правилу 2 критическую точку $u_{кр}$ находим из равенства $\Phi(u_{кр}) = 1 - 2\alpha = 1 - 2 \cdot 0.05 = 0.9$. По таблице значений функции Лапласа находим $u_{кр} = 1.65$. Так как $U_{набл} = 10 > 1.65$, то гипотезу H_0 отвергаем. Таким образом, различие между выборочной и гипотетической генеральной средней значимое.

Рассмотрим случай, когда *дисперсия* $D(X) = \sigma^2$ генеральной совокупности, распределенной по нормальному закону, *неизвестна* (т.е. σ неизвестно). Такая ситуация может

возникнуть, например, в случае малых выборок. В качестве проверки гипотезы H_0 принимают СВ

$$T = \frac{\overline{X} - a_0}{S} \cdot \sqrt{n-1}, \qquad (3.8)$$

где S – "исправленное" среднее квадратическое отклонение. Случайная величина T имеет распределение Стьюдента с $k = n-1$ степенями свободы. Критическая область, как и в рассмотренном выше случае с известной дисперсией $D(X) = \sigma^2$, строится в зависимости от вида конкурирующей гипотезы.

Правило 1. Для того, чтобы при заданном уровне значимости α проверить гипотезу $H_0 : a = a_0$ о равенстве неизвестной генеральной средней a нормальной совокупности с неизвестной дисперсией σ^2 гипотетическому значению a_0 при конкурирующей гипотезе $H_1 : a \neq a_0$, необходимо вычислить

$$T_{набл} = \frac{\overline{x}_в - a_0}{s} \cdot \sqrt{n-1} \qquad (3.9)$$

и по таблице критических точек распределения Стьюдента, по заданному уровню значимости α, помещенному в верхней строке таблицы, и числу степеней свободы $k = n-1$ найти критическую точку $t_{двуст.кр.}(\alpha; k)$ *двусторонней критической области*. Если $|T_{набл}| < t_{двуст.кр}$ – нет оснований отвергнуть гипотезу H_0; если $|T_{набл}| > t_{двуст.кр}$ – гипотезу H_0 отвергают.

Правило 2. При конкурирующей гипотезе $H_1 : a > a_0$ по заданному
уровню значимости α, помещенному в нижней строке таблицы критических точек распределения Стьюдента, и числу степеней свободы $k = n-1$ найти критическую точку $t_{правост.кр.}(\alpha; k)$ *правосторонней критической области*.

Если $T_{набл} < t_{правост.кр}$ – нет оснований отвергнуть гипотезу H_0; если $T_{набл} > t_{правост.кр}$ – гипотезу H_0 отвергают.

Правило 3. При конкурирующей гипотезе $H_1: a < a_0$ сначала по правилу 2 находят "вспомогательную" критическую точку $t_{правост.кр.}(\alpha; k)$, а затем полагают границу *левосторонней критической области*

$$t_{левост.кр.}(\alpha; k) = -t_{правост.кр.}(\alpha; k).$$

Если $T_{набл} > -t_{правост.кр}$ – нет оснований отвергнуть гипотезу H_0; если $T_{набл} < -t_{правост.кр}$ – гипотезу H_0 отвергают.

Для нахождения критической области необходимо знать критическое значение выборочной средней, которое можно найти из формулы для статистики (3.8):

$$\overline{x}_{кр} = a_0 + \frac{t_{кр}}{\sqrt{n-1}} \cdot s \qquad (3.10)$$

Пример 2. По выборке объема $n = 20$, извлеченной из нормальной генеральной совокупности, найдены выборочное среднее значение $\overline{x}_в = 18$ и "исправленное" среднее квадратическое отклонение $s = 4.5$. При уровне значимости 0.05 проверить гипотезу $H_0: a = a_0 = 17$ при конкурирующей гипотезе $H_1: a \neq 17$.

Решение. Вычислим наблюдаемое значение критерия:

$$T_{набл} = \frac{\overline{x}_в - a_0}{s} \cdot \sqrt{n-1} = \frac{18-17}{4.5} \cdot \sqrt{19} = 0.99.$$

Поскольку конкурирующая гипотеза $H_1: a \neq 17$ – двусторонняя, то по таблице критических точек распределения Стьюдента по уровню значимости $\alpha = 0.05$, помещенному в верхней строке таблицы, и по числу степеней свободы $k = 20 - 1 = 19$, согласно правилу 1, находим критическую точку $t_{двуст.кр.}(\alpha; k) = t_{двуст.кр.}(0.05; 19) = 2.09$. Так как $|T_{набл}| = 0.99 < \quad < t_{двуст.кр} = 2.09$, то нет оснований

отвергнуть гипотезу $H_0: a = a_0 = 17$. Следовательно, выборочное среднее *незначимо* отличается от гипотетической генеральной средней.

Пример 3. На основании сделанного прогноза ***средняя дебиторская задолженность*** однотипных предприятий региона должна составить $a_0 = 120$ ден.ед. Выборочная проверка 10 предприятий дала среднюю задолженность $\overline{x}_в = 135$ ден.ед. и среднее квадратическое отклонение $s = 20$ ден.ед. При уровне значимости 0.05 выяснить, можно ли принять данный прогноз. Найти критическую область для $\overline{x}$, если в действительности средняя дебиторская задолженность всех предприятий региона равна 130 ден.ед.

Решение. Проверяемая гипотеза $H_0: a = \overline{x}_в = 120$ при конкурирующей гипотезе $H_1: a > 120$. Так как генеральная дисперсия σ^2 неизвестна, то используем $t-$критерий Стьюдента. Вычислим наблюдаемое значение критерия:

$$T_{набл} = \frac{\overline{x}_в - a_0}{s} \cdot \sqrt{n-1} = \frac{135-120}{20} \cdot \sqrt{10-1} = 2.25.$$

Поскольку конкурирующая гипотеза $H_1: a > 120$ – правосторонняя, то по таблице критических точек распределения Стьюдента по уровню значимости $\alpha = 0.05$, помещенному в нижней строке таблицы, и по числу степеней свободы $k = 10-1 = 9$, согласно правилу 2, находим критическую точку

$$t_{правостор.кр.}(\alpha; k) = t_{правостор.кр.}(0.05; 9) = 1.83.$$

Так как $T_{набл} = 2.25 > 1.83$, то гипотеза H_0 отвергается, т.е. на 5%-ом уровне значимости сделанный прогноз должен быть отвергнут.

Так как выдвинутая альтернативная гипотеза $H_1: a > 120$, то критическая область – правосторонняя и критическое

значение выборочной средней можно найти из формулы (3.10): $\overline{x}_{кр} = a_0 + \frac{t_{кр}}{\sqrt{n-1}} \cdot s = 120 + 1.83\frac{20}{\sqrt{10-1}} = 132.2$ ден.ед.

Таким образом, критическая область значений для $\overline{x}$ есть интервал $(132.2; +\infty)$.

Пример 4. На основании исследований одного залегания ученым-археологам стало известно, что диаметр раковин ископаемого моллюска равен 18.2 мм. В распоряжении ученых оказалась выборка из 50 раковин моллюсков из другого залегания, для которой было вычислено $\overline{x}_в = 18.9$ мм, $S = 2.18$ мм. Можно ли сделать предположение при $\alpha = 0.05$, что конкретное местообитание раковин не оказало влияние на диаметр их раковин?

Решение. Нулевая гипотеза $H_0 : a = a_0 = 18.2$ при конкурирующей гипотезе $H_1 : a \neq 18.2$. По правилу 1 имеем двустороннюю критическую область. Найдем и сравним $T_{набл}$ и $t_{двуст.кр.}$. Имеем:

$$T_{набл} = \frac{\overline{x}_в - a_0}{s} \cdot \sqrt{n-1} = \frac{18.9 - 18.2}{2.18} \cdot \sqrt{50-1} \approx 0.5$$

$$t_{двуст.кр.}(\alpha; k) = t_{двуст.кр.}(0.05; 49) = 2.02$$

Так как $T_{набл} = 0.5 < t_{двуст.кр.} = 2.02$, то нет оснований принимать гипотезу H_0. Т.е. с 95% уверенностью можно утверждать, что конкретное местообитание раковин оказало влияние на диаметр их раковин.

Исключение грубых ошибок наблюдений

Грубые ошибки могут возникнуть из-за ошибок показаний измерительных приборов, ошибок регистрации, случайного сдвига запятой в десятичной записи числа и т.д.

Пусть, например, $x^*, x_1, x_2, ..., x_n$ – совокупность имеющихся наблюдений, причем x^* резко выделяется.

Необходимо решить вопрос о принадлежности резко выделяющегося значения к остальным наблюдениям.

Для ряда наблюдений $x_1, x_2, ..., x_n$ рассчитывают $\overline{x}_в$ и исправленное среднее квадратическое отклонение S. При справедливости гипотезы $H_0 : \overline{x} = x^*$ о принадлежности x^* к остальным наблюдениям статистика

$$T = \frac{\overline{x} - x^*}{S} \qquad (3.11)$$

имеет $t-$распределение Стьюдента с $\nu = n-1$ степенями свободы. Конкурирующая гипотеза H_1 имеет вид: $\overline{x} > x^*$ или $\overline{x} < x^*-$ в зависимости от того, является ли резко выделяющееся значение больше или меньше остальных наблюдений. Гипотеза H_0 отвергается, если $|T_{набл.}| > t_{кр.}$, и принимается, если $|T_{набл.}| < t_{кр.}$

Пример 5. Имеются следующие данные об урожайности пшеницы на 8 опытных участках одинакового размера (ц/га):

26.5 26.2 35.9 30.1 32.3 29.3 26.1 25.0

Есть основание предполагать, что значение урожайности третьего участка $x^* = 35.9$ зарегистрировано неверно. Является ли это значение аномальным (резко выделяющимся) на 5%-ом уровне значимости?

Решение. Исключив значение $x^* = 35.9$, найдем для оставшихся наблюдений

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{26.5 + 26.2 + 30.1 + 32.3 + 29.3 + 26.1 + 25.0}{7} = 27.93$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{6}\Big((26.5-27.93)^2 + (26.2-27.93)^2 + (30.1-27.93$$

$$+(32.3-27.93)^2 + (29.3-27.93)^2 + (26.1-27.93)^2 + (25.0-27.93)^2\Big)$$

$$= 7.109$$

Значит, $S = 2.67$ (ц/га).

Фактически наблюдаемое значение

$$T_{набл.} = \frac{\overline{x} - x^*}{S} = \frac{27.93 - 35.9}{2.67} \approx -2.98.$$

Табличное значение $t_{кр.} = t_{1-2\alpha;\,n-1} = t_{0.9;\,6} = 1.94$.

Сравнивая $|T_{набл.}|$ и $t_{кр.}$ ($|T_{набл.}| > t_{кр.}$), гипотезу $H_0 : \overline{x} = x^*$ (о принадлежности x^* к остальным наблюдениям) отвергаем. Следовательно, значение $x^* = 35.9$ является аномальным, и его следует отбросить.

Критерий для проверки гипотезы о вероятности события

Пусть проведено n независимых испытаний (n – достаточно большое число), в каждом из которых некоторое событие A появляется с одной и той же, но неизвестной вероятностью p, и найдена относительная частота $\frac{m}{n}$ появлений A в этой серии испытаний. Проверим при заданном уровне значимости α нулевую гипотезу H_0, состоящую в том, что вероятность p равна некоторому значению p_0.

Примем в качестве статистического критерия случайную величину

$$U = \frac{\left(\frac{m}{n} - p_0\right)\sqrt{n}}{\sqrt{p_0 q_0}}, \qquad (3.12)$$

имеющую нормальное распределение с параметрами $M(U)=0$, $\sigma(U)=1$ (то есть нормированную). Здесь $q_0 = 1 - p_0$.

Вывод о нормальном распределении критерия следует из теоремы Лапласа (при достаточно большом n относительную частоту можно приближенно считать нормально распределенной с математическим ожиданием p и средним квадратическим отклонением $\sqrt{\frac{pq}{n}}$).

Критическая область строится в зависимости от вида конкурирующей гипотезы.

1) Если H_0: $p = p_0$, а H_1: $p \neq p_0$, то критическую область нужно построить так, чтобы вероятность попадания критерия в эту область равнялась заданному уровню значимости α. При этом наибольшая мощность критерия достигается тогда, когда критическая область состоит из двух интервалов, вероятность попадания в каждый из которых равна $\frac{\alpha}{2}$. Поэтому $u_{кр}$ определяется по таблице значений функции Лапласа из условия $\Phi(u_{кр}) = 1-\alpha$, а критическая область имеет вид $(-\infty;-u_{кр}) \cup (u_{кр};+\infty)$.

Далее нужно вычислить наблюдаемое значение критерия:

$$U_{набл} = \frac{\left(\frac{m}{n} - p_0\right)\sqrt{n}}{\sqrt{p_0 q_0}}. \qquad (3.13)$$

Если $|U_{набл}|<u_{кр}$, то нулевая гипотеза принимается.
Если $|U_{набл}|>u_{кр}$, то нулевая гипотеза отвергается.

2) Если конкурирующая гипотеза H_1: $p>p_0$, то критическая область определяется неравенством $U>u_{кр}$, то есть является правосторонней, причем $p(U>u_{кр})=\alpha$. Тогда $p(0<U<u_{кр}) = 1-2\alpha$. Следовательно, $u_{кр}$ можно найти по таблице значений функции Лапласа из условия, что $\Phi(u_{кр}) = 1-2\alpha$. Вычислим наблюдаемое значение критерия по формуле (3.13).Если $U_{набл}<u_{кр}$, то нулевая гипотеза принимается. Если $U_{набл}>u_{кр}$, то нулевая гипотеза отвергается.

3) Для конкурирующей гипотезы H_1: $p < p_0$ критическая область является левосторонней и задается неравенством $U <- u_{кр}$, где $u_{кр}$ вычисляется как в предыдущем случае.
Если $U_{набл} > - u_{кр}$, то нулевая гипотеза принимается.
Если $U_{набл} < - u_{кр}$, то нулевая гипотеза отвергается.

Пример 6. Пусть проведено 50 независимых испытаний, и относительная частота появления события A оказалась равной 0,12. Проверить при уровне значимости α=0,01 нулевую гипотезу H_0: p=0,1 при конкурирующей гипотезе H_1: p>0,1.

Решение. Найдем $U_{набл} = \dfrac{(0,12-0,1)\sqrt{50}}{\sqrt{0,1\cdot 0,9}} = 0,471.$

Критическая область является правосторонней, а $u_{кр}$ находим из равенства $\Phi(u_{кр})=1-2\cdot 0,01=0,98$. Из таблицы значений функции Лапласа определяем $u_{кр}$=2,33. Итак, $U_{набл}<u_{кр}$, и гипотеза о том, что p=0,1, принимается.

Непараметрические критерии

Если закон распределения генеральной совокупности неизвестен, но имеются основания предположить, что предполагаемый закон имеет определенный вид (например, A), то проверяют нулевую гипотезу: H_0: {**генеральная совокупность распределена по закону А**}.
Проверка гипотезы о предполагаемом законе распределения так же, как и проверка гипотезы о неизвестных параметрах известного закона распределения, производится при помощи специально подобранной случайной величины – *критерия согласия*.

Как бы хорошо ни был подобран теоретический закон распределения, между эмпирическим и теоретическим распределениями неизбежны расхождения.

Поэтому возникает вопрос: объясняются ли эти расхождения случайными обстоятельствами, связанными с ограниченным числом наблюдений, или они являются существенными и связаны с тем, что теоретический закон распределения

подобран неудачно. Для ответа на этот вопрос и служат критерии согласия.

Критерием согласия называют критерий проверки гипотезы о предполагаемом законе неизвестного распределения.

Одним из основных критериев согласия является *критерий* χ^2 *(критерий Пирсона*).

Критерий Пирсона

Критерий Пирсона позволяет, в частности, проверить гипотезу о нормальном распределении генеральной совокупности. Для проверки этой гипотезы будем сравнивать *эмпирические* n_i (т.е. наблюдаемые) и *теоретические* n_i' (т.е. вычисленные в предположении нормального закона распределения) *частоты*, которые, как правило, различаются.

Случайно (*незначимо*) или неслучайно (*значимо*) это расхождение?

Ответ на этот вопрос и дает критерий согласия Пирсона.

Предположим, что генеральная совокупность X распределена нормально.

Приведем ***алгоритм нахождения теоретических частот***.

1. Весь интервал наблюдаемых значений СВ X (выборки объема n) делят на k частичных интервалов (x_i, x_{i+1}) одинаковой длины, находят $x_i^* = {(x_i + x_{i+1})}/{2}$ – середины частичных интервалов. В качестве частоты n_i варианты x_i^* принимают число вариант, попавших в i-ый интервал. Получают последовательность равноотстоящих вариант и соответствующих им частот:

x_i	x_1^*	x_2^*	…	x_k^*	
n_i	n_1	n_2	…	n_k	$\sum_{i=1}^{k} n_i = n$

2. Вычисляют $\overline{x}_в^*$ и выборочное среднее квадратическое отклонение σ^*.

3. Нормируют СВ X, т.е. переходят к величине $Z=\frac{(X-\overline{x}_в^*)}{\sigma^*}$ и вычисляют концы интервалов (z_i, z_{i+1}): $z_i=(x_i-\overline{x}_в^*)/\sigma^*$, $z_{i+1}=(x_{i+1}-\overline{x}_в^*)/\sigma^*$, причем полагают наименьшее значение $z_1=-\infty$, а наибольшее $z_k=+\infty$.

4. Вычисляют теоретические вероятности p_i попадания СВ X в интервалы (x_i, x_{i+1}) по формуле $p_i=\frac{1}{2}\left(\Phi(z_{i+1})-\Phi(z_i)\right)$ ($\Phi(z)$–функция Лапласа). Находят *теоретические частоты* $n_i'=np_i$.

Пусть по выборке объема n нормально распределенной генеральной совокупности X получено эмпирическое распределение

x_i	x_1	x_2	…	x_k	
n_i	n_1	n_2	…	n_k	$\sum_{i=1}^{k} n_i = n$

и вычислены теоретические частоты n_i'.

Задача: при уровне значимости α проверить справедливость нулевой гипотезы H_0: {генеральная совокупность распределена нормально}.

В качестве критерия проверки гипотезы H_0 примем СВ

$$\chi^2=\sum_{i=1}^{k}\frac{\left(n_i-n_i'\right)^2}{n_i'} \qquad (3.14)$$

Это – СВ, т.к. в различных опытах она принимает различные, не известные заранее значения. Ясно, что $\chi^2\to 0$

при $n_i \to n_i'$, т.е. чем меньше различаются эмпирические n_i и теоретические n_i' частоты, тем меньше значение критерия χ^2. Таким образом, критерий (3.14) характеризует близость эмпирического и теоретического распределения.

Известно, что при $n \to \infty$ закон распределения СВ (3.14) стремится к закону распределения χ^2 с ν степенями свободы.

Поэтому СВ в (3.14) обозначается через χ^2, а сам критерий называют *критерием согласия* χ^2. Число степеней свободы ν находят по равенству $\nu = k - r - 1$, где k – число групп (частичных интервалов), r – число параметров предполагаемого распределения, которые оценены по данным выборки (для нормального закона распределения $r = 2$, поэтому $k = l - 3$).

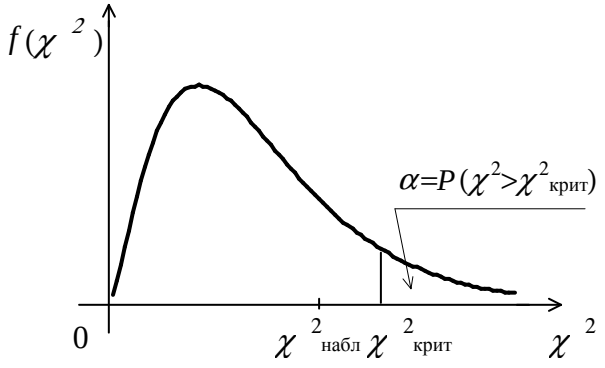
Построим *правостороннюю критическую область* (т.к. односторонний критерий более "жестко" отвергает гипотезу H_0), исходя из требования, чтобы, в предположении справедливости гипотезы H_0, вероятность попадания критерия в эту область была равна принятому уровню значимости α: $P\,[\chi^2 > \chi^2_{кр}(\alpha; k)] = \alpha$. Следовательно, правосторонняя критическая область определяется неравенством $\chi^2 > \chi^2_{кр}(\alpha; k)$, а область принятия гипотезы H_0 – неравенством $\chi^2 < \chi^2_{кр}(\alpha; k)$.

Значение критерия (3.14), вычисленное по данным наблюдений, обозначим $\chi^2_{набл}$. Сформулируем

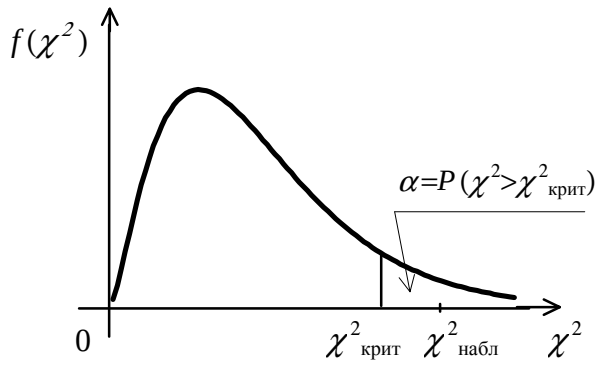
Правило проверки нулевой гипотезы H_0. *Для того чтобы при заданном уровне значимости* α *проверить нулевую гипотезу* H_0: *{генеральная совокупность распределена нормально}, необходимо вычислить теоретические частоты* n_i' *и наблюдаемое значение критерия согласия* χ^2 *Пирсона*

$\chi^2_{набл} = \sum_{i=1}^{k} \frac{(n_i - n'_i)^2}{n'_i}$. *По таблице критических точек распределения* χ^2 *по заданному уровню значимости* α *и числу степеней свободы* $k = l - 3$ *найти критическую точку* $\chi^2_{кр}(\alpha; k)$.

- *Если наблюдаемое значение критерия* $\chi^2_{€‡‘}$ *попало в область принятия гипотезы* $\chi^2_{набл} < \chi^2_{кр}(\alpha; k)$, *то нет оснований отвергнуть нулевую гипотезу* H_0 *(рис. 5а)).*
- *Если наблюдаемое значение критерия* $\chi^2_{€‡‘}$ *попало в критическую область* $\chi^2_{набл} > \chi^2_{кр}(\alpha; k)$, *то нулевую гипотезу* H_0 *отвергают (рис. 5б)).*



а).



б)

Рис. 5

Покажем, что для контроля вычислений наблюдаемого критерия $\chi^2_{набл}$ можно использовать равенство

$$\sum_{i=1}^{k} \frac{n_i^2}{n_i'} - n = \chi^2_{набл}. \qquad (3.15)$$

Действительно, из (3.14) вытекает:

$$\chi^2 = \sum_{i=1}^{l} \frac{\left(n_i - n_i'\right)^2}{n_i'} = \sum_{i=1}^{l} \left(\frac{n_i^2}{n_i'} - \frac{2n_i n_i'}{n_i'} + \frac{n_i'^2}{n_i'} \right) = \sum_{i=1}^{l} \left(\frac{n_i^2}{n_i'} - 2n_i + n_i' \right) =$$

$$= \sum_{i=1}^{l} \frac{n_i^2}{n_i'} - 2\sum_{i=1}^{l} n_i + \sum_{i=1}^{l} np_i = \sum_{i=1}^{l} \frac{n_i^2}{n_i'} - 2n + n\sum_{i=1}^{l} p_i = \sum_{i=1}^{l} \frac{n_i^2}{n_i'} - 2n + n =$$

$$= \sum_{i=1}^{l} \frac{n_i^2}{n_i'} - n$$

Пример 7. Используя критерий Пирсона при уровне значимости 0,05, установить, случайно или значимо расхождение между эмпирическими и теоретическими частотами, которые вычислены, исходя из предположения о нормальном распределении признака X генеральной совокупности:

n_i	14	18	32	70	20	36	10
n_i'	10	24	34	80	18	22	12.

Решение. Выдвигаем нулевую гипотезу H_0 и ей конкурирующую H_1.

H_0: признак X имеет нормальный закон распределения.

H_1: признак X имеет закон распределения, отличный от нормального.

В данном случае рассматривается правосторонняя критическая область. Проверим гипотезу с помощью случайной величины $\chi^2_{набл} = \sum_{i=1}^{k} \frac{(n_i - n_i')^2}{n_i'}$, которая имеет распределение χ^2 с $\nu = k - 3 = 7 - 3 = 4$ степенями свободы. Вычислим наблюдаемое значение критерия χ^2 по выборочным данным. Расчеты представим в таблице:

	n_i	n_i'	$\frac{(n_i - n_i')^2}{n_i'}$
	14	10	1,6
	18	24	1,5
	32	34	0,118
	70	80	1,25
	20	18	0,222
	36	22	8,909
	10	12	0,333
Итого	200	200	13,932

.

Итак имеем, $\chi^2_{\text{набл}} \approx 13{,}93$; $\chi^2_{\text{крит}}(0{,}05;4) = 9{,}5$ (по таблице).

Сравниваем $\chi^2_{\text{набл}}$ и $\chi^2_{\text{крит}}(0{,}05;\ 4)$.

Так как $\chi^2_{\text{набл}} > \chi^2_{\text{крит}}(0{,}05;\ 4)$, то есть наблюдаемое значение критерия попало в критическую область (см. рис. 5б)), нулевая гипотеза отвергается, справедлива конкурирующая гипотеза, то есть признак X имеет закон распределения, отличный от нормального, расхождение между эмпирическими и теоретическими частотами значимо.

Критерий Колмогорова

На практике кроме критерия χ^2 часто используют *критерий Колмогорова* (*критерий согласия* λ), в котором в качестве меры расхождения между теоретическим и эмпирическим распределениями рассматривают максимальное значение абсолютной величины разности между эмпирической функцией распределения $F^*(x)$ и соответствующей теоретической функцией распределения $F(x)$: $D^* = \sup_x |F^*(x) - F(x)|$. Случайная величина D^* называется *статистикой Колмогорова*. Для *непрерывных СВ* с помощью этого критерия можно проверить гипотезу о виде функции распределения, а также:

1. согласие эмпирического распределения с гипотетическим (т.е. с теоретическим);
2. гипотезу о том, что две выборки взяты из одной и той же генеральной совокупности, т.е. определяются одним и тем же теоретическим распределением.

Теорема Колмогорова. *Пусть* $F(x)$ – *гипотетическая функция распределения случайной величины* X; $x_1, x_2, \ldots, x_n$. – *выборка объема* n *и* $F^*(x)$ – *эмпирическая функция распределения. Тогда*

$$P(\sqrt{n}\,D^* < x) \xrightarrow[n\to\infty]{} K(x) = \sum_{k=+\infty}^{+\infty} (-1)^k e^{-2k^2x^2}, \; x > 0,$$

где $D^* = \sup_x |F^*(x) - F(x)|$.

При использовании этого критерия следует учитывать, что функция $F(x)$ не должна зависеть от параметров выборки. *Статистика Колмогорова* используется для решения следующих задач:

1.проверка гипотезы $H_0 : F^*(x) = F(x)$, где $F^*(x)$ – эмпирическая функция распределения СВ X, вычисленная по

выборке $x_1, x_2, \ldots, x_n$; $F(x)$ – гипотетическая функция распределения СВ X;

2.построение доверительных границ для $F(x)$.

При использовании статистики Колмогорова следует иметь в виду таблицу приложений [9] критических значений λ_α статистики Колмогорова, в которой для типичных уровней значимости α приведены значения *критической точки* λ_α, удовлетворяющей уравнению $P(D^* = \sup_x |F^*(x) - F(x)| > \lambda_\alpha) = \alpha.$

Так как распределение статистики Колмогорова D^* не зависит от
оцениваемой функции $F(x)$ и в качестве меры расстояния между $F^*(x)$ и $F(x)$ используют максимальное отклонение, то величину D^* можно применять для построения доверительных границ непрерывной функции
распределения $F(x)$.

Для любой неизвестной непрерывной функции $F(x)$ при произвольном x имеем

$$P(F^*(x) - \lambda_\alpha \le F(x) \le F^*(x) + \lambda_\alpha) = 1 - \alpha.$$

Значит, доверительная область есть полоса шириной $2\lambda_\alpha$, в центре которой находится выборочная функция распределения $F^*(x)$, причем с вероятностью $1-\alpha$ истинная функция распределения $F(x)$ целиком лежит внутри этой полосы. Сформулированный результат дает основание для оценивания максимального объема выборки, необходимого для аппроксимации неизвестной функции распределения $F(x)$ с заданной точностью. Например, при $\alpha = 0.01, n = 100$ эмпирическая функция распределения $F^*(x)$ повсюду отстоит от истинной $F(x)$ не более, чем на 0.161.

Схема применения критерия Колмогорова:

1. Строятся эмпирическая функция распределения $F^*(x)$ и предполагаемая теоретическая функция распределения $F(x)$.

2. Определяется мера расхождения между теоретическим и эмпирическим распределением D^*: $D^* = \sup_x |F^*(x) - F(x)|$, вычисляется величина $\lambda = D^* \sqrt{n}$.

3. Если $\lambda > \lambda_\alpha$ для заданного уровня значимости α, то нулевая гипотеза H_0 о том, что СВ X имеет заданный закон распределения, отвергается. Если $\lambda \le \lambda_\alpha$, то считают, что гипотеза H_0 не противоречит опытным данным.

Элементы теории корреляции. Линейная регрессия

Одной из основных задач МС является нахождение зависимости между двумя или несколькими СВ. В естественных науках часто речь идет о ***функциональной зависимости*** между величинами X и Y, когда *каждому значению одной переменной соответствует вполне определенное значение другой* (например, скорость свободного падения тела в вакууме зависит от времени падения). Однако строгая функциональная зависимость реализуется редко, т.к. обе СВ или одна из них подвержены действию случайных факторов. В этом случае возникает ***статистическая зависимость***.

Статистической (*вероятностной, стохастической*) называют зависимость, при которой *изменение одной из величин влечет изменение распределения другой*.

Примером статистической зависимости может служить зависимость всхожести семян некоторых культур от количества микроэлементов при их обработке, зависимость производительности труда на предприятии от его энерговооруженности и.т.д.

Таким образом, статистическая зависимость между двумя СВ Y и X неоднозначна. Статистическая зависимость, в частности, проявляется в том, что при изменении одной из величин изменяется *среднее значение* другой.

Для исследователя представляет интерес *усредненная* по x схема зависимости, т.е. закономерность в изменении *условного математического ожидания* $M_x(Y)$ (или в других обозначениях – $M(Y|X=x)$, т.е. математического ожидания СВ Y, вычисленной в предположении, что СВ X приняла значение x) в зависимости от x. Такую статистическую зависимость называют ***корреляционной***.

Корреляционной (или ***регрессионной***) ***зависимостью*** между двумя переменными величинами называется функциональная зависимость между значениями одной из них и условным математическим ожиданием другой.

Корреляционная зависимость, например, имеется:

- между ростом и весом человека – с увеличением роста *средний вес* также возрастает;
- между надежностью автомобиля и его возрастом – чем больше возраст, тем *в среднем* меньше его надежность.

Рассмотрим пример СВ Y, которая связана с другой СВ X не функционально, а корреляционно.

Пример 1. Пусть Y – успеваемость студентов, X – посещаемость ими учебных занятий. У одинаковых студенческих групп (по количеству студентов и количеству часов лекционных и практических занятий) по результатам экзаменационной сессии успеваемость разная, т.е. Y не является функцией от X – посещаемости учебных занятий: $Y \neq f(X)$. Однако, как показывает опыт, результаты экзаменационной сессии лучше у тех студентов, которые систематически посещали учебные занятия. Это означает, что Y связано с X корреляционной зависимостью (связью).

Для уточнения определения корреляционной зависимости введем понятие *условной средней*.

Условной средней $\vec{Y}_x = M(Y \mid X = x)$ называется среднее значение СВ Y при $X = x$.

В качестве *оценок* условных математических ожиданий принимают *условные средние*, которые находят по данным наблюдений, т.е. по выборке. Так например, если при $x_1 = 2$ СВ Y приняла значения $y_1 = 3,\ y_2 = 7, y_3 = 5$, то *условное среднее* $\overline{y}_{x_1} = \dfrac{y_1 + y_2 + y_3}{3} = \dfrac{3+7+5}{3} = 5.$

Условное математическое ожидание $M_x(Y)$ СВ Y есть функция от x: $M_x(Y) = f(x)$, которую называют *функцией регрессии Y на X*.

Поскольку каждому значению x соответствует одно значение условного среднего, т.е. $\vec{Y}_x = f(x)$ является функцией от x, то можно сказать, что СВ Y зависит от СВ X *корреляционно***.**

Корреляционной зависимостью Y от X называется функциональная зависимость условной средней $\vec{Y}_x$ от x.

Уравнение $y = f(x)$ называется *уравнением регрессии Y на X*. Функция $f(x)$ называется *регрессией Y на X*, а ее график – *линией регрессии СВ Y на* СВ X.

Аналогично для СВ X определяются *условное математическое ожидание* $M_y(X)$ (или в других обозначениях – $M(X|Y=y)$, *условное среднее* $\bar{X}_y = M(X|Y=y)$, *корреляционная зависимость* СВ X от СВ Y, *функция регрессии* СВ X на СВ Y: $\varphi(y) = \bar{X}_y$.

Основными задачами теории корреляции являются:

1. Установление формы корреляционной связи, т.е. вида функции регрессии (линейная, квадратичная, показательная и т.д.).

2. Оценка тесноты корреляционной связи Y от X, которая оценивается величиной рассеяния значений Y около $\vec{Y}_x$. Большое рассеяние означает слабую зависимость Y от X либо вообще отсутствие таковой. Малое рассеяние указывает на существование достаточно сильной зависимости Y от X.

Важной с точки зрения приложений является ситуация, когда обе функции регрессии $f(x), \varphi(y)$ являются линейными. Тогда говорят, что СВ X и Y связаны линейной корреляционной зависимостью (*линейной корреляцией*).

Такая ситуация возникает, например, если система СВ (X,Y) имеет совместное нормальное распределение. Тогда модельные уравнения регрессии являются линейными, а их графики – прямыми.

Рассмотрим методы нахождения линейной регрессии, представляющей наибольший интерес.

Пусть даны результаты n измерений двух СВ X и Y: (x_1, y_1), (x_2, y_2), ..., (x_n, y_n). Предварительное представление о характере зависимости между X и Y можно получить, если элементы выборки $(x_i, y_i), i = \overline{1,n}$, изобразить графически точками координатной плоскости в выбранной системе координат. В результате получим *точечную диаграмму* статистической зависимости, которая называется ***корреляционным полем***. По его виду можно составить предварительное мнение о степени и типе зависимости двух СВ.

Как известно, для описания системы двух СВ (X,Y) вводят математические ожидания $M(X)$, $M(Y)$ и дисперсии $D(X) = \sigma_X^2$, $D(Y) = \sigma_Y^2$ для каждой из составляющих, а также *корреляционный момент* (*ковариацию*)

$$K(X,Y) = = M\{[X - M(X)][Y - M(Y)]\}$$

и *коэффициент корреляции*

$$r = \frac{K(X,Y)}{\sigma(X)\sigma(Y)}$$

Корреляционный момент $K(X,Y)$ служит для характеристики связи между СВ X и Y: если X и Y независимы, то $K(X,Y) = 0$, а следовательно, и $r = 0$.

Две случайные величины X и Y называются *коррелированными*, если их корреляционный момент (или, что то же самое, коэффициент корреляции) отличен от нуля. СВ X и Y называются *некоррелированными*, если их корреляционный момент равен нулю.

Рассмотрим вопрос о силе связи между признаками X и Y. Для этой цели введем *выборочный коэффициент корреляции* $r_в$. На основе определения *теоретического коэффициента корреляции* $r = \dfrac{K(X,Y)}{\sigma(X)\sigma(Y)}$ и оценок параметров теоретического распределения через выборочные, *выборочный коэффициент корреляции* $r_в$ может быть представлен в виде

$$r_в = \frac{K_в(X,Y)}{\sigma_в(X)\sigma_в(Y)}, \qquad (4.1)$$

где $K_в(X,Y)$ – выборочный корреляционный момент,

$$K_в(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_в)(y_i - \overline{y}_в) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}_в \cdot \overline{y}_в = \overline{xy}_в - \overline{x}_в \cdot \overline{y}_в$$

,

$\sigma_в(X) = \sqrt{D_в(X)} = \sqrt{\overline{x_в^2} - (\overline{x}_в)^2}$, $\sigma_в(Y) = \sqrt{D_в(Y)} = \sqrt{\overline{y_в^2} - (\overline{y}_в)^2}$ – выборочные среднеквадратические отклонения признаков X и Y. Следовательно, из (4.1) имеем

$$r_в = \frac{\overline{xy}_в - \overline{x}_в \cdot \overline{y}_в}{\sqrt{\overline{x_в^2} - (\overline{x}_в)^2}\sqrt{\overline{y_в^2} - (\overline{y}_в)^2}} \qquad (4.2)$$

Формула (4.2) симметрична относительно двух переменных, т.е. x и y можно менять местами. Выборочный коэффициент корреляции $r_в$ обладает теми же свойствами, что и теоретический коэффициент корреляции $r_Г$, и является мерой линейной зависимости между двумя наблюдаемыми величинами.

Основные свойства коэффициента корреляции:

1. Коэффициент корреляции $r_в$ принимает значения на отрезке $[-1;1]$, т.е. $-1 \le r_в \le 1$.

2. Если все значения переменных увеличить (уменьшить) на одно и то же число или в одно и то же число раз, то

величина выборочного коэффициента корреляции не изменится.

3. При $r_в = \pm 1$ корреляционная связь представляет *линейную функциональную зависимость*. При этом линии регрессии Y на X и X на Y совпадают, все наблюдаемые значения располагаются на общей прямой.

4. Если с ростом значений одной СВ значения второй возрастают, то $r_в > 0$, если убывают, то $r_в < 0$.

5. При $r_в = 0$ *линейная* корреляционная связь отсутствует, групповые средние переменных совпадают с их общими средними, а линии регрессии Y на X и X на Y параллельны осям координат.

Замечание. Равенство $r_в = 0$ говорит лишь об отсутствии *линейной* корреляционной зависимости (т.е. о некоррелированности СВ X и Y), но не вообще об отсутствии корреляционной, а тем более статистической зависимости.

Выборочный коэффициент корреляции $r_в$ является *оценкой генерального коэффициента корреляции*

$$r_Г = \frac{K(X,Y)}{\sigma(X)\sigma(Y)},$$

характеризующего тесноту связи между СВ X и Y генеральной совокупности. На практике о тесноте корреляционной зависимости между рассматриваемыми СВ судят не по величине $r_Г$, которая, как правило, неизвестна, а по величине ее выборочного аналога $r_в$. Так как $r_в$ вычисляется по значениям переменных, случайно попавших в выборку из генеральной совокупности, то в отличие от параметра $r_Г$ параметр $r_в$ – *величина случайная*.

Проверка значимости выборочного коэффициента корреляции в случае выборки из нормального двумерного распределения

Пусть двумерная генеральная совокупность (X,Y) распределена нормально.

Из этой совокупности извлечена выборка объема n и по ней найден выборочный коэффициент корреляции $r_в$, причем оказалось, что $r_в \neq 0$. Поскольку выборка произведена случайно, нельзя утверждать, что $r_Г \neq 0$. Требуется проверить, значимо или незначимо отличие выборочного коэффициента корреляции $r_в$ от нуля либо это вызвано только случайными изменениями выборочных значений.

При заданном уровне значимости α проверить справедливость гипотезы $H_0 : r_Г = 0$ о равенстве нулю генерального коэффициента корреляции при конкурирующей гипотезе $H_1 : r_Г \neq 0$.

Для проверки гипотезы H_0 задается уровень значимости α – допустимая вероятность ошибки (α=0.05; 0.01; 0.1). Вычисляют статистику $T_{набл} = \dfrac{|r_в|}{\sqrt{1-r_в^2}} \cdot \sqrt{n-2}$, где n – объем выборки. По таблице распределения Стьюдента по заданному уровню значимости α и числу степеней свободы $k = n-2$ находят $t_{кр} = t_{\alpha;n-2}$. Если $T_{набл} < t_{кр}$ – нет оснований отвергнуть гипотезу H_0. Если $T_{набл} > t_{кр}$, то гипотезу H_0 о равенстве коэффициента корреляции нулю отвергают. Другими словами, $r_в$ значимо отличается от нуля, т.е. СВ X и Y коррелированны. В этом случае считают, что зависимость между наблюдаемыми величинами можно приблизить линейной зависимостью.

Пример 2. По выборке объема $n = 102$, извлеченной из нормальной двумерной совокупности, найден выборочный

коэффициент корреляции $r_в = 0.3$. При уровне значимости 0.05 проверить нулевую гипотезу о равенстве нулю генерального коэффициента $r_Г$ при конкурирующей гипотезе $H_1 : r_Г \neq 0$.

Решение. Найдем наблюдаемое значение критерия $T_{набл} = \frac{|r_в|}{\sqrt{1-r_в^2}} \cdot \sqrt{n-2} = \frac{|0.3|}{\sqrt{1-0.3^2}} \cdot \sqrt{102-2} = \frac{30}{0.954} = 28.62$.

По условию, конкурирующая гипотеза имеет вид $r_Г \neq 0$, следовательно, критическая область – двусторонняя. По уровню значимости $\alpha = 0.05$ и числу степеней свободы $k = n-2 = 100$ по таблице критических точек распределения Стьюдента для двусторонней критической области находим критическую точку $t_{кр} = t_{\alpha;n-2} = t_{кр}(0.05;100) = 1.98$. Так как $T_{набл} = 28.62 > t_{кр} = 1.98$, то нулевую гипотезу H_0 отвергаем, т.е. выборочный коэффициент корреляции значимо отличается от нуля. Таким образом, СВ X и Y коррелированны.

Определение коэффициентов уравнения линейной регрессии

Рассмотрим систему двух СВ (X,Y). Если обе функции регрессии $f(x), \varphi(y)$ Y на X и X на Y линейны, то говорят, что СВ X и Y связаны *линейной корреляционной зависимостью*. Тогда графиками линейных функций регрессии являются прямые линии. Пусть, например, приближенное представление СВ Y представлено в виде линейной функции СВ X: $Y \approx f(x) = a + bX$, где a, b – параметры, подлежащие определению. Их можно определить различными способами. Наиболее употребительным является метод наименьших квадратов (МНК). Функцию $f(x)$ называют "*наилучшим приближением*" Y в смысле МНК, если математическое ожидание $M[Y - f(X)]^2$ принимает

наименьшее возможное значение. Поэтому $f(x)$ называют *среднеквадратической регрессией* Y на X.

Теорема 1. *Линейная средняя квадратическая регрессия* Y *на* X *имеет вид*

$$f(x) = m_y + r\frac{\sigma_y}{\sigma_x}(x - m_x), \qquad (4.3)$$

где

$$m_x = M(X), m_y = M(Y), \sigma_x = \sqrt{D(X)}, \sigma_y = \sqrt{D(Y)}, r = \frac{K(X,Y)}{\sigma_x \sigma_y}$$

– коэффициент корреляции СВ X *и* Y.

Из теоремы 1 следует, что в силу (4.3) параметры a, b имеют вид: $a = m_y - r\frac{\sigma_y}{\sigma_x}m_x$, $b = r\frac{\sigma_y}{\sigma_x}$.

Коэффициент $b = r\frac{\sigma_y}{\sigma_x}$ называется *коэффициентом регрессии* Y *на* X,

прямая $y - m_y = r\frac{\sigma_y}{\sigma_x}(x - m_x)$ *– прямой среднеквадратической регрессии* Y *на* X.

Аналогично можно получить прямую среднеквадратической регрессии X на Y:

$$x - m_x = r\frac{\sigma_x}{\sigma_y}(y - m_y),$$

где $r\frac{\sigma_x}{\sigma_y}$ *– коэффициент регрессии* X *на* Y.

Из сопоставления уравнений линейной регрессии Y на X и X на Y видно, что при $r = \pm 1$ они совпадают. Кроме того очевидно, что обе прямые

$$y - m_y = r\frac{\sigma_y}{\sigma_x}(x - m_x),\ x - m_x = r\frac{\sigma_x}{\sigma_y}(y - m_y)$$

проходят через точку $(m_x; m_y)$, называемую *центром совместного распределения СВ X и Y*.

Выше были введены уравнения регрессии Y на X: $M(Y|X=x)=f(x)$ и X на Y: $M(X|Y=y)=\varphi(y)$.

Поскольку условное математическое ожидание $M(Y|X=x)$ является функцией от x, то и его оценка $M(M(Y|X=x))$, т.е. условное среднее $\overline{y}_x$, также является функцией от x: $\overline{y}_x = f^*(x)$; где $f^*(x)$ в силу (4.3) имеет вид

$$f^*(x) = \overline{y}_в + r_в \frac{s_y}{s_x}(x - \overline{x}_в).$$

Полученное уравнение называют *выборочным уравнением регрессии Y на X*; функцию $f^*(x)$ – *выборочной регрессией Y на X*, а ее график – *выборочной линией регрессии Y на X*. Аналогично, уравнение $\overline{x}_y = \varphi^*(y)$, где $\varphi^*(y) = \overline{x}_в + r_в \frac{s_x}{s_y}(y - \overline{y}_в)$, называют *выборочным уравнением регрессии X на Y*; функцию $\varphi^*(y)$ – *выборочной регрессией X на Y*, а ее график – *выборочной линией регрессии X на Y*.

Как по данным наблюдений $(x_1; y_1), (x_2; y_2), \ldots, (x_n; y_n)$, полученным в результате n независимых опытов, найти параметры функций $f^*(x)$, $\varphi^*(y)$, если их вид известен? Как оценить тесноту связи между СВ X и Y и установить, коррелированны ли эти величины?

Пусть принята гипотеза о линейной зависимости между величинами X и Y. По данным наблюдений найдем, например, выборочное уравнение прямой линии среднеквадратической регрессии Y на X: $f^*(x) = Y = a + bx$. Будем полагать, что все результаты измерений $(x_i; y_i)$, $i = \overline{1,n}$ различны. Подберем параметры a, b так, чтобы точки $(x_1; y_1), (x_2; y_2), \ldots, (x_n; y_n)$, построенные на плоскости xOy по

данным наблюдений, лежали как можно ближе к прямой $Y = a + bx$ в смысле МНК. Сформулированное требование означает, что параметры a, b будем выбирать из условия, чтобы сумма квадратов отклонений $Y_i - y_i, i = \overline{1,n}$, была минимальной. Здесь Y_i – ордината, вычисленная по эмпирическому (выборочному) уравнению $Y_i = a + bx_i$, соответствующая наблюдаемому значению x_i, y_i – наблюдаемая ордината, соответствующая x_i. Следовательно, рассмотрим функцию

$$F(a,b) = \sum_{i=1}^{n} (Y_i - y_i)^2 = \sum_{i=1}^{n} ((a + bx_i) - y_i)^2 \to \min_{a,b} \qquad (4.4)$$

Необходимое условие экстремума сводится к условиям

$$\begin{cases} \dfrac{\partial F(a,b)}{\partial a} = 0, \\ \dfrac{\partial F(a,b)}{\partial b} = 0. \end{cases}$$

Находя соответствующие частные производные и приравнивая их нулю, получаем:

$$\begin{cases} n \cdot a + b \sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} y_i, \\ a \sum\limits_{i=1}^{n} x_i + b \sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} x_i y_i. \end{cases}, \qquad (4.5)$$

или, после деления обеих частей уравнений системы на n:

$$\begin{cases} a + b\overline{x}_в = \overline{y}_в, \\ a\overline{x}_в + b\overline{x_в^2} = \overline{xy}_в \end{cases},$$

где $\overline{x}_в = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i$, $\overline{y}_в = \dfrac{1}{n}\sum\limits_{i=1}^{n} y_i$, $\overline{x_в^2} = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i^2$, $\overline{xy}_в = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i y_i$.

Вычисляя определитель Δ данной системы

$$\Delta = \begin{vmatrix} 1 & \overline{x}_в \\ \overline{x}_в & \overline{x_в^2} \end{vmatrix} = \overline{x_в^2} - \left(\overline{x}_в\right)^2 = D_в,$$

находим неизвестные параметры a, b:

$$a = \frac{\Delta a}{\Delta} = \frac{\begin{vmatrix} \overline{y}_в & \overline{x}_в \\ \overline{xy}_в & \overline{x_в^2} \end{vmatrix}}{\overline{x_в^2} - \left(\overline{x}_в\right)^2} = \frac{\overline{y}_в \cdot \overline{x_в^2} - \overline{x}_в \cdot \overline{xy}_в}{\overline{x_в^2} - \left(\overline{x}_в\right)^2} =$$

$$= \frac{(\overline{y}_в \cdot \overline{x_в^2} - \overline{y}_в \cdot \left(\overline{x}_в\right)^2) + \overline{y}_в \cdot \left(\overline{x}_в\right)^2 - \overline{x}_в \cdot \overline{xy}_в}{\overline{x_в^2} - \left(\overline{x}_в\right)^2} =$$

$$= \overline{y}_в - \frac{\overline{x}_в \cdot \overline{xy}_в - \overline{y}_в \cdot \left(\overline{x}_в\right)^2}{\overline{x_в^2} - \left(\overline{x}_в\right)^2} = \overline{y}_в - \frac{\overline{x}_в \cdot (\overline{xy}_в - \overline{y}_в \cdot \overline{x}_в)}{\overline{x_в^2} - \left(\overline{x}_в\right)^2} \overset{(4.1)}{=}$$

$$= \overline{y}_в - \frac{\overline{x}_в \cdot K_в(X,Y)}{\sigma_в^2(X)} = \overline{y}_в - \frac{\overline{x}_в \cdot r_в(X,Y) \cdot \sigma_в(X) \cdot \sigma_в(Y)}{\sigma_в^2(X)} =$$

$$= \overline{y}_в - r_в(X,Y)\frac{\overline{x}_в \cdot \sigma_в(Y)}{\sigma_в(X)},$$

$$b = \frac{\Delta b}{\Delta} = \frac{\begin{vmatrix} 1 & \overline{y}_в \\ \overline{x}_в & \overline{xy}_в \end{vmatrix}}{\overline{x_в^2} - \left(\overline{x}_в\right)^2} = \frac{\overline{xy}_в - \overline{y}_в \cdot \overline{x}_в}{\overline{x_в^2} - \left(\overline{x}_в\right)^2} = \frac{K_в(X,Y)}{\sigma_в^2(X)} \overset{(4.1)}{=}$$

$$= \frac{r_в(X,Y) \cdot \sigma_в(X) \cdot \sigma_в(Y)}{\sigma_в^2(X)} = r_в(X,Y)\frac{\sigma_в(Y)}{\sigma_в(X)}.$$

Таким образом, выборочное уравнение линейной регрессии $y = a + bx$ Y на X имеет вид:

$$y = \overline{y}_в - r_в(X,Y)\frac{\overline{x}_в \cdot \sigma_в(Y)}{\sigma_в(X)} + r_в(X,Y)\frac{\sigma_в(Y)}{\sigma_в(X)}x,$$

или в иной записи – $y - \overline{y}_в = r_в(X,Y)\frac{\sigma_в(Y)}{\sigma_в(X)}(x - \overline{x}_в)$.

Аналогичным образом можно получить уравнение линейной регрессии X на Y: $x-\overline{x}_в = r_в(X,Y)\dfrac{\sigma_в(X)}{\sigma_в(Y)}\left(y-\overline{y}_в\right)$.

На практике совместное распределение СВ (X,Y) зачастую неизвестно, а известны только результаты наблюдений, т.е. выборка пар (x_i,y_i), $i=\overline{1,n}$ значений СВ (X,Y). Тогда все рассмотренные величины $m_x, m_y, \sigma_x, \sigma_y, r$ заменяем их выборочными аналогами: в полученных уравнениях $\sigma_в(X)$, $\sigma_в(Y)$ – их несмещенными оценками

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x}_в)^2 = \frac{1}{n-1}\sum_{i=1}^{n}x_i^2 - \frac{n}{n-1}\overline{x}_в^2,$$

$$S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y}_в)^2 = \frac{1}{n-1}\sum_{i=1}^{n}y_i^2 - \frac{n}{n-1}\overline{y}_в^2,$$

а x, y в левых частях – на соответствующие условные средние $\overline{y}_x$, $\overline{x}_y$, получим эмпирические функции линейной регрессии в виде

$$\overline{y}_x - \overline{y}_в = r_в\frac{S_y}{S_x}(x-\overline{x}_в), \qquad (4.6)$$

$$\overline{x}_y - \overline{x}_в = r_в\frac{S_x}{S_y}\left(y-\overline{y}_в\right). \qquad (4.7)$$

Заметим, что если нанести обе линии регрессии на корреляционное поле, то прямые должны пересечься в точке $\left(\overline{x}_в, \overline{y}_в\right)$.

Уравнения линейной регрессии получены в предположении, что все измерения встречаются по одному разу. При большом числе наблюдений одно и то же значение СВ X может повторяться n_x раз, а СВ Y – n_y раз. Одинаковая пара чисел (x,y) может наблюдаться n_{xy} раз. Поэтому результаты наблюдений группируют, подсчитывая

частоты n_x, n_y, n_{xy}. Все данные записывают в *корреляционную таблицу*. Построение корреляционной таблицы следующее: в клетки верхней строки записывают наблюдаемые значения $x_i, i=\overline{1,k}$, а в первый столбец – наблюдаемые значения y_j, $j=\overline{1,m}$. На пересечении строк и столбцов записывают *кратности* $n_{x_i y_j}, i=\overline{1,k}$, $j=\overline{1,m}$ наблюдаемых пар значений признаков. В правом нижнем углу расположена сумма всех частот n_{x_i}, $i=\overline{1,k}$ и n_{y_j}, $j=\overline{1,m}$, равная общему числу всех наблюдений n (объему выборки).

x_i y_i	x_1	x_2	…	x_k	n_{y_j}
y_1	n_{11}	n_{12}	…	n_{1k}	$\sum_{i=1}^{k} n_{1i}$
y_2	n_{21}	n_{22}	…	n_{2k}	$\sum_{i=1}^{k} n_{2i}$
…	…	…	…	…	…
y_m	n_{m1}	n_{m2}	…	n_{mk}	$\sum_{i=1}^{k} n_{mi}$
n_{x_i}	$\sum_{j=1}^{m} n_{j1}$	$\sum_{j=1}^{m} n_{j2}$	…	$\sum_{j=1}^{m} n_{j1}$	$n=\sum_{i=1}^{k} n_{x_i}=\sum_{j=1}^{m} r$

Для непрерывных СВ корреляционная таблица имеет вид

x_i / y_i	$[x_1, x_2)$	$[x_2, x_3)$	…	$[x_{k-1}, x_k]$	n_{y_j}
$[y_1, y_2)$	n_{11}	n_{12}	…	n_{1k}	$\sum_{i=1}^{k} n_{1i}$
$[y_2, y_3)$	n_{21}	n_{22}	…	n_{2k}	$\sum_{i=1}^{k} n_{2i}$
…	…	…	…	…	…
$[y_{m-1}, y_m]$	n_{m1}	n_{m2}	…	n_{mk}	$\sum_{i=1}^{k} n_{mi}$
n_{x_i}	$\sum_{j=1}^{m} n_{j1}$	$\sum_{j=1}^{m} n_{j2}$	…	$\sum_{j=1}^{m} n_{j1}$	$n = \sum_{i=1}^{k} n_{x_i} = \sum_{j=1}^{m} n$

где n_{ij} – частоты (кратности) наблюдаемых пар значений признаков, попавших в соответствующие интервалы $[x_i, x_{i+1})$, $[y_j, y_{j+1})$, $i = \overline{1, k-1}$, $j = \overline{1, m-1}$. В этом случае таблица сводится к предыдущей путем перехода к серединам интервалов группировки статистических данных.

Если на основании наблюдаемых значений (x_i, y_i), $i = \overline{1, n}$ СВ (X, Y) можно предположить, что зависимость y_i от x_i квадратичная ($y = ax^2 + \ + bx + c$, то применение МНК

$$S(a,b,c) = \sum_{i=1}^{n} \left(ax_i^{\,2} + bx_i + c - y_i \right)^2 \to \min$$

дает возможность найти неизвестные параметры a,b,c.

Отметим, что и в этом случае схема для нахождения параметров a,b,c является линейной.

Если же рассматривается нелинейная зависимость наблюдаемых значений y_i от x_i, то обычно используют методы линеаризации, т.е. переходят к условным переменным, где зависимость от параметров становится линейной, а затем применяют МНК. Пусть, например, на основании наблюдаемых значений (x_i, y_i), $i=\overline{1,n}$ СВ (X,Y) выдвинута гипотеза H_0: зависимость y_i от x_i имеет вид a,b $y=ae^{bx}$. Прологарифмировав данное нелинейное уравнение, получим $\ln y=\ln a+bx\ln e$. Введя обозначения $Y=\ln a$, $A=\ln a, B=b,$ получим линейную зависимость $Y=A+Bx,$ для которой можно применить описанный выше МНК нахождения неизвестных параметров A,B. Из введенной замены переменных находим $a=e^A, b=B,$ а следовательно, и предполагаемую зависимость $y=e^{A+Bx}$.

Помимо зависимости (корреляционной) между двумя СВ можно рассматривать корреляционную зависимость одной СВ от двух и более СВ. В таких случаях говорят о *множественной регрессии*. Например, множественная регрессия от двух переменных: $z=ax+by+c$.

В этом случае параметры уравнения находятся по МНК и рассматривают корреляционную связь между каждыми признаками и отдельно тесноту связи между признаком z и общими признаками x и y. Для этого вычисляется совместный выборочный коэффициент корреляции, который выражается через выборочные коэффициенты корреляции компонент. При нахождении выборочного уравнения регрессии необходимо проверять статистическую гипотезу о значимости коэффициента корреляции, т.е. о том, как связано выборочное уравнение регрессии с регрессией, изучаемой генеральной совокупностью.

Уравнение регрессии можно использовать для *прогнозирования* (*предсказания*).

Пример 3. Изучается зависимость себестоимости одного изделия (Y, р.) от величины выпуска продукции (X, тыс. шт.) по группе предприятий за отчетный период. Получены следующие данные:

X	2	3	4	5	6
Y	1,9	1,7	1,8	1,6	1,4

.

Провести корреляционно-регрессионный анализ зависимости себестоимости одного изделия от выпуска продукции.

Решение. Признак X – объем выпускаемой продукции, тыс. шт. (факторный признак). Признак Y – себестоимость одного изделия, р. (результативный признак). Предполагаем, что признаки имеют нормальный закон распределения. Признаки находятся в статистической зависимости, так как себестоимость одного изделия зависит не только от объема выпускаемой продукции, но и от многих других факторов, которые в данном случае не учитываются. Определим форму связи. Построим точки с координатами (x_i , y_i) и по их расположению определим форму связи (рис. 6).

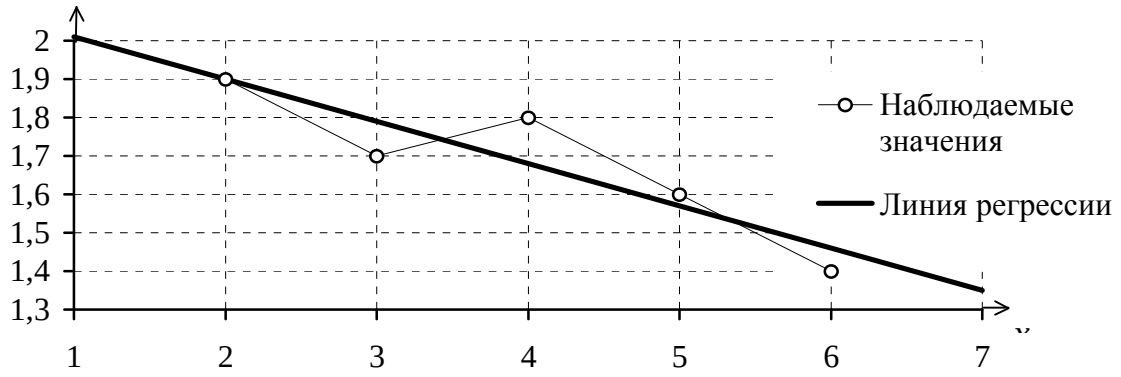


Рис. 6

Итак, форма связи линейная.

Проведем корреляционный анализ. Рассчитаем выборочный линейный коэффициент корреляции:

$$r_в = \frac{\overline{xy}_в - \overline{x}_в \cdot \overline{y}_в}{\sqrt{\overline{x_в^2} - (\overline{x}_в)^2}\sqrt{\overline{y_в^2} - (\overline{y}_в)^2}}$$

Расчеты представим в таблице:

x_i	y_i	$x_i \cdot y_i$	x_i^2	y_i^2
2	1,9	3,8	4	
3	1,7	5,1	9	3,61
4	1,8	7,2	16	
5	1,6	8,0	25	2,89
6	1,4	8,4	36	
				3,24
				2,56
				1,96
Итого 20	8,4	32,5	90	14,26.

$$\overline{x}_в = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{20}{5} = 4; \quad \overline{y}_в = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{8,4}{5} = 1,68;$$

$$\overline{xy}_в = \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \frac{32,5}{5} = 6,5;$$

$$\overline{x_в^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{90}{5} = 18; \quad \overline{y_в^2} = \frac{1}{n}\sum_{i=1}^{n} y_i^2 = \frac{14,26}{5} = 2,852;$$

$$\sigma_x^2 = \overline{x_в^2} - \overline{x}_в^2 = 18 - 16 = 2; \quad \sigma_y^2 = \overline{y_в^2} - \overline{y}_в^2 = 2,852 - (1,68)^2 = 0,029$$

$$r_{‰} = \frac{6,5 - 4 \cdot 1,68}{\sqrt{2 \cdot 0,0296}} \approx -\ 0,90.$$

Проверим значимость выборочного коэффициента корреляции. Для этого выдвигаем гипотезы:

H_0: $r_\Gamma = 0$,

H_1: $r_\Gamma \neq 0$. Примем уровень значимости $\alpha = 0,05$.

Для проверки нулевой гипотезы используем случайную величину $T_{набл} = \dfrac{|r_в|}{\sqrt{1-r_в^2}} \cdot \sqrt{n-2}$, имеющую распределение Стьюдента с $k = n - 2 = 3$ степенями свободы. По выборочным данным находим наблюдаемое значение критерия $T_{набл} = \dfrac{|-0,90| \cdot \sqrt{3}}{\sqrt{1-0,81}} \approx 3,58$. По таблице критических точек распределения Стьюдента находим $t_{кр.дв}(0,05; 3) = 3,18$. Сравниваем $T_{набл}$ и $t_{кр}(0,05; 3)$. Так как $T_{набл} > t_{кр}$, то есть $T_{набл}$ попало в критическую область, нулевая гипотеза отвергается, справедлива конкурирующая гипотеза: $r_Г \neq 0$, значит, $r_в$ значим. Признаки X и Y коррелированны. Так как $|r_в|$ близок к единице, следовательно, себестоимость одного изделия и объем выпускаемой продукции находятся в тесной корреляционной зависимости.

Найдем коэффициент детерминации. $D = r_в^2 \cdot 100\ \% = 0,81\ \%$, то есть вариация себестоимости единицы продукции в среднем на 81 % объясняется вариацией объема выпускаемой продукции.

Выразим эту связь аналитически приблизительно в виде линейного уравнения регрессии:

$$\bar{y}_x - \bar{y}_в \approx b(x - \bar{x}_в),$$

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x^2} = \frac{-0,22}{2} = -0,11.$$

$$\bar{y}_x - 1,68 = -0,11\,(x - 4) \text{ или } \bar{y}_x \approx -0,11x + 2,12.$$

Из уравнения следует, что с увеличением выпуска продукции на 1 тыс. шт. себестоимость одного изделия снизится в среднем на 0,11 р.

Найдем по уравнению регрессии себестоимость одного изделия, если выпуск продукции составит 5,2 тыс. шт. :

$$\overline{y}_x \approx -0{,}11 \cdot 5{,}2 + 2{,}12 = 1{,}55 \text{ (р.)}$$

Пример 4. Для нормирования труда проведено статистическое исследование связи между количеством изготавливаемых изделий (X, шт.) и затратами времени на обработку одного изделия (Y, мин). Сделана выборка объемом $n = 51$ и получены следующие данные: $r_в = 0{,}8$, $\overline{x} = 8$, $\sigma_x = 3{,}2$, $\overline{y} = 40$, $\sigma_y = 8$. Проверить значимость коэффициента корреляции при $\alpha = 0{,}02$. Построить уравнение регрессии.

Решение. Признак X – количество изготавливаемых изделий, шт. Признак Y – затраты времени на обработку одного изделия, мин.

Предполагаем, что признаки имеют нормальный закон распределения. Они находятся в статистической зависимости, так как затраты времени зависят не только от количества изготавливаемых изделий, но и от многих других факторов, которые в данном случае не учитываются. В данном случае связь линейная, теснота связи характеризуется линейным коэффициентом корреляции $r_в = 0{,}8$. Но прежде чем делать вывод о тесноте взаимосвязи, необходимо проверить значимость коэффициента корреляции. Выдвигаем нулевую гипотезу и ей конкурирующую:

H_0: $r_Г = 0$,

H_1: $r_Г \neq 0$.

Проверяем нулевую гипотезу с помощью случайной величины, имеющей распределение Стьюдента с $k = n - 2 = 49$ степенями свободы: $T = \dfrac{|r_в|}{\sqrt{1 - r_в^2}} \cdot \sqrt{n-2}$.

По выборочным данным найдем наблюдаемое значение критерия $T_{набл} = \dfrac{0{,}8 \cdot \sqrt{49}}{\sqrt{1-0{,}64}} \approx 9{,}33$. По таблице критических точек распределения Стьюдента находим $t_{кр.дв}(\alpha, k) = t_{кр.дв}(0{,}02;\ 49) = 2{,}40$. Сравниваем $T_{набл}$ и $t_{кр.дв}(0{,}02;\ 49)$. Так как $|T_{набл}| > t_{кр.дв}(0{,}02;\ 49)$, то есть наблюдаемое значение критерия попало в критическую область, нулевая гипотеза отвергается, справедлива конкурирующая гипотеза: $r_г \neq 0$, признаки X и Y коррелированны, $r_в$ значим.

$D = r_в^2 \cdot 100\ \% = 64\ \%$, то есть вариация затрат времени на обработку одного изделия в среднем на 64 % объясняется за счет вариации количества изготавливаемых изделий.

Выразим эту взаимосвязь аналитически в виде уравнения регрессии вида:

$$\bar{y}_x - \bar{y} \approx b(x - \bar{x}).$$

Коэффициент b выразим через парный линейный коэффициент корреляции:

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x^2}; \quad r_{‰} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y}.$$

Сравнивая эти две формулы, можем записать:

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y} \cdot \frac{\sigma_y}{\sigma_x} = r_в \cdot \frac{\sigma_y}{\sigma_x}.$$

Тогда по выборочным данным будем иметь:

$b = 0{,}8 \cdot 8/32 = 2$; $\bar{y}_x - 40 \approx 2(x - 8)$ или $\bar{y}_x \approx 24 + 2x$.

Из уравнения следует, что с увеличением количества выпускаемых изделий на 1 шт., затраченное время в среднем увеличится на 2 мин.

КОНТРОЛЬНЫЕ ВОПРОСЫ

по курсу «Математическая статистика»

1. Предметы и методы математической статистики. Задачи математической статистики. Генеральная и выборочная совокупности.
2. Выборочные аналоги интегральной функций распределения.
3. Выборочные аналоги дифференциальной функций распределения.
4. Статистические характеристики вариационных рядов.
5. Понятие о точечной оценке числовой характеристики случайной величины. Свойства точечных оценок.
6. Точечные оценки математического ожидания и их свойства.
7. Точечные оценки дисперсии и их свойства.
8. Частость как точечная оценка вероятности.
9. Понятие об интервальной оценке параметров распределения.
10. Доверительный интервал для математического ожидания при известном σ.
11. Доверительный интервал для математического ожидания при неизвестном σ.
12. Интервальная оценка вероятности.
13. Определение объема выборки.
14. Понятие статистических гипотез их виды. Понятие ошибки первого и второго рода.
15. Основной принцип проверки статистических гипотез
16. Понятие односторонней и двусторонней критической области. Правило нахождения критических точек.
17. Проверка гипотез о среднем значении нормально распределенной СВ при известной дисперсии
18. Проверка гипотез о среднем значении нормально распределенной СВ при неизвестной дисперсии
19. Исключение грубых ошибок наблюдений

20. Построение теоретического закона распределения по опытным данным.
21. Критерий согласия Пирсона.
22. Понятие функциональной, стохастической и корреляционной зависимости. Функция регрессии.
23. Генеральное и выборочное корреляционные отношения.
24. Линейное уравнение регрессии.
25. Генеральный и выборочный коэффициенты корреляции.
26. Нелинейные функции регрессии.
27. Понятие о множественной регрессии.

ЛИТЕРАТУРА

1. *Гмурман В.Е.* Теория вероятностей и математическая статистика. –М.: Высшая школа, 1977.
2. *Гмурман В.Е.* Руководство к решению задач по теории вероятностей и математической статистике. –М.: Высшая школа, 1997.
3. *Калинина В.Н., Панкин В.Ф.* Математическая статистика. – М.: Высшая школа, 1994.
4. *Мацкевич И.П., Свирид Г.П., Булдык Г.М.* Сборник задач и упражнений по высшей математике (Теория вероятностей и математическая статистика).– Минск: Вышейша школа, 1996.
5. *Тимофеева Л.К., Суханова Е.И.* Математика для экономистов. Сборник задач по теории вероятностей и математической статистике. –М.: УМиИЦ «Учебная литература», 1998. –182 с.
6. *Кремер Н.Ш.* Теория вероятностей и математическая статистика. –М.: ЮНИТИ, 2001.–542с.
7. *Гусак А.А., Бричикова Е.А.* Справочное пособие к решению задач. Теория вероятностей.– Минск: ТетраСистемс, 2000.–
8. *Гурский Е.И.* Сборник задач по теории вероятностей и математической статистике.–Минск: Вышэйшая школа, 1975.–250с.
9. *Вентцель Е.С.* Теория вероятностей.–М: из-во физико-математической литературы, 1962–564с.

Перечень вопросов по курсу высшей математики
(раздел «Теория вероятностей и математическая статистика»)

1. Случайные события, их классификация и действия над ними.
2. Классическое, статистическое и геометрическое определение вероятности.
3. Элементы комбинаторики: размещения, перестановки и сочетания. Свойства сочетаний.
4. Теорема сложения и умножения вероятностей.
5. Зависимые и независимые события. Условная вероятность. Теоремы умножения вероятностей.
6. Теорема сложения вероятностей совместных событий. Вероятность наступления только одного, хотя бы одного события.
7. Формула полной вероятности и формулы Байеса.
8. Повторные независимые испытания. Формула Бернулли.
9. Найвероятнейшее число появления события.
10. Понятие дискретной случайной величины и ее закона распределения. Многоугольник распределения. Примеры.
11. Функция распределения случайной величины и ее свойства. График функции распределения дискретной случайной величины.
12. Математическое ожидание дискретной случайной величины и его свойства.
13. Дисперсия дискретной случайной величины и ее свойства. Среднее квадратическое отклонение.
14. Биномиальный закон распределения и его числовые характеристики.
15. Закон Пуассона и его числовые характеристики.
16. Равномерное дискретное распределение и его характеристики.
17. Плотность распределения вероятностей непрерывной случайной величины и ее свойства.
18. Математическое ожидание и дисперсия непрерывной случайной величины.
19. Равномерный закон распределения и его числовые характеристики.
20. Показательный закон распределения и его числовые характеристики.
21. Нормальный закон распределения, его параметры и их вероятностный смысл. Влияние параметров a и σ на форму нормальной кривой.
22. Вероятность попадания нормально распределенной случайной величины в заданный интервал; вероятность заданного отклонения.
23. Правило трех сигм и его значение для практики.
24. Функция Лапласа и ее связь с функцией распределения нормальной случайной величины.
25. Моменты случайных величин. Асимметрия и эксцесс.
26. Неравенство Маркова.
27. Неравенство Чебышева. Следствия.
28. Теорема Чебышева и ее следствия.
29. Теорема Бернулли. Значение закона больших чисел.
30. Понятие о центральной предельной теореме и ее следствиях.
31. Предмет и задачи математической статистики. Генеральная и выборочная совокупности. Способ отбора.
32. Построение дискретного вариационного ряда. Эмпирическая функция распределения и ее свойства.
33. Построение интервального вариационного ряда. Гистограмма частот и относительных частот.
34. Средняя арифметическая и ее свойства.
35. Дисперсия вариационного ряда и ее свойства. Исправленная выборочная дисперсия.
36. Понятие доверительного интервала.
37. Основные понятия статистической проверки гипотез. Гипотезы и , критерий проверки, ошибки первого и второго рода, критическая область, мощность критерия.
38. Критерий согласия Пирсона о предполагаемом законе распределения случайной величины.
39. Модели и основные понятия регрессионного и корреляционного анализа.
40. Нахождение параметров линейного уравнения регрессии методом наименьших квадратов.
41. Понятие коэффициента линейной корреляции и его свойства.

ЛИТЕРАТУРА

Электронные издания

\\C\dymkov\2007-2008\МЭО\3 семестр
(Лекции, Литература, Вопросы и пр.)***.pdf

Основная литература

1. Мацкевич И.П., Свирид Г.П. Высшая математика: Теория вероятностей и математическая статистика. - Минск: Выш. шк., 1993. - 269 с.
2. Барковская Л.С, Станишевская Л.В., Черторицкий Ю.Н. Теория вероятностей: Практикум. 2-ое изд.- Минск: БГЭУ, 2005.- 142 с
3. Станишевская Л.В., Черторицкий Ю.Н. Теория вероятностей: Практикум. 2-ое изд.- Минск: БГЭУ, 2006.- 174 с.
4. Гмурман В.Е. Теория вероятностей и математическая статистика. – М..: Наука, 1993. – 326 с.
5. Кремер Н.Ш. Теория вероятностей и математическая статистика- 2-е изд., перераб. и доп. М.: Юнити, 2004. – 573 с.
6. Мацкевич И.П., Свирид Г.П., Булдык Г.М. Сборник задач и упражнений по высшей математике: Теория вероятностей и математическая статистика. - Минск: Выш. шк„ 1996.-318 с.
7. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. – М.: Высш. шк., 1979. – 400 с.

Дополнительная литература

1. Вентцель Е.С. Теория вероятностей: Учеб. для вузов.– 7-е изд.стер. – М.: Высш. шк., 2001. – 575 с.
2. Феллер В. Введение в теорию вероятностей и ее приложения. – М.:, 1964.
3. Белько И.В., Свирид Г.П. Теория вероятностей и математическая статистика. Примеры и задачи. – Минск: Новое знание, 2002. – 250 с.
4. Гурский Е.И. Сборник задач по теории вероятностей и математической статистике. – Минск: Выш. шк., 1984. – 223 с.