Н.Д.Выск

Теория вероятностей и математическая статистика

учебное пособие

МАТИ-РГТУ им. К.Э. Циолковского
Кафедра «Высшая математика»
2011

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Методические указания

Начинайте каждое занятие с изучения лекции. При этом:

- ✓ вначале внимательно прочтите определения и осознайте смысл используемых терминов
- ✓ затем прочтите формулировки теорем, которые задают свойства изучаемых объектов
- ✓ разберите доказательства теорем и выводы формул
- ✓ в завершение работы прочтите всю лекцию еще раз, чтобы убедиться, что теоретический материал освоен.

Следующий этап работы – выполнение заданий практикума.

- ✓ каждую задачу попробуйте решить самостоятельно
- ✓ в случае неудачи посмотрите указание и вновь повторите попытку
- ✓ в случае повторной неудачи внимательно разберите приведенное решение
- ✓ если вы решили задачу самостоятельно (во всяком случае, ваш ответ оказался верным), все равно обязательно прочтите решение, данное в учебном курсе – это поможет вам проверить правильность примененного метода решения
- ✓ закончив решение всех задач практикума, обязательно вернитесь к тем из них, которые не получились в первый раз, и попробуйте вновь самостоятельно решить их.

При выполнении домашнего задания используйте материал лекции и практикума.

1. ТЕОРИЯ ВЕРОЯТНОСТЕЙ

1.1. СЛУЧАЙНЫЕ СОБЫТИЯ

1.1.1. *Случайные события. Классическое и статистическое определение вероятности*

В различных разделах науки и техники нередко возникают ситуации, когда результат каждого из многих проводимых опытов заранее предугадать невозможно, однако можно исследовать закономерности, возникающие при проведении серии опытов. Нельзя, например, точно сказать, какая сторона монеты окажется сверху при данном броске: герб или цифра – но при большом количестве бросков число выпадений герба приближается к половине количества бросков; нельзя заранее предсказать результат одного выстрела из данного орудия по данной цели, но при большом числе выстрелов частота попадания приближается к некоторому постоянному числу. Исследование вероятностных закономерностей массовых однородных явлений составляет предмет **теории вероятностей**.

Основным интуитивным понятием классической теории вероятностей является **случайное событие**. События, которые могут произойти в результате опыта, можно подразделить на три вида:
а) **достоверное событие** – событие, которое всегда происходит при проведении опыта;
б) **невозможное событие** – событие, которое в результате опыта произойти не может;
в) **случайное событие** – событие, которое может либо произойти, либо не произойти. Например, при броске игральной кости достоверным событием является выпадение числа очков, не превышающего 6, невозможным – выпадение 10 очков, а случайным – выпадение 3 очков.

Алгебра событий

Определение. **Суммой A+B** двух событий A и B называют событие, состоящее в том, что произошло хотя бы одно из событий A и B. **Суммой нескольких событий**, соответственно, называется событие, заключающееся в том, что произошло хотя бы одно из этих событий.

Пример 1.
Два стрелка делают по одному выстрелу по мишени. Если событие A – попадание первого стрелка, а событие B – второго, то сумма A+B – это хотя бы одно попадание при двух выстрелах.

Пример 2.

Если при броске игральной кости событием A_i назвать выпадение i очков, то выпадение нечетного числа очков является суммой событий $A_1+A_2+A_3$.

Назовем все возможные результаты данного опыта его *исходами* и предположим, что множество этих исходов, при которых происходит событие A (исходов, *благоприятных* событию A), можно представить в виде некоторой области на плоскости. Тогда множество исходов, при которых произойдет событие A+B, является объединением множеств исходов, благоприятных событиям A или B (рис. 1).

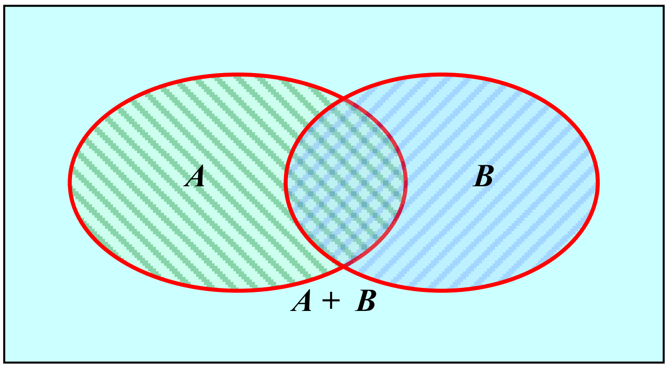


Рис.1.

Определение. **Произведением AB** событий A и B называется событие, состоящее в том, что произошло и событие A, и событие B. Аналогично **произведением нескольких событий** называется событие, заключающееся в том, что произошли все эти события.

Пример 3.

В примере 1 (два выстрела по мишени) событием AB будет попадание обоих стрелков.

Пример 4.

Если событие A состоит в том, что из колоды карт извлечена карта пиковой масти, а событие B – в том, что из колоды вынута дама, то событием AB будет извлечение из колоды дамы пик.

Геометрической иллюстрацией множества исходов опыта, благоприятных появлению произведения событий A и B, является пересечение областей, соответствующих исходам, благоприятным A и B.

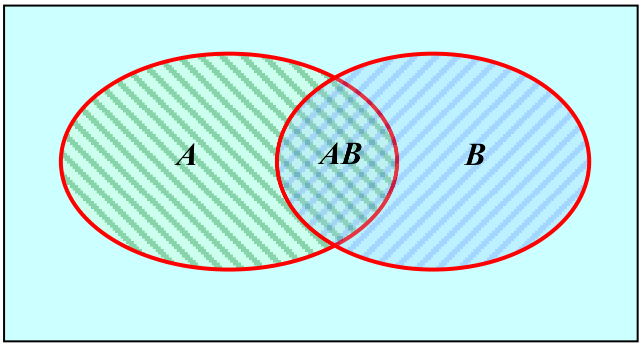


Рис.2.

Определение. **Разностью A\B** событий A и B называется событие, состоящее в том, что A произошло, а B – нет.

Пример 5.

Вернемся к примеру 1, где A\ B – попадание первого стрелка при промахе второго.

Пример 6.

В примере 4 A\B – извлечение из колоды любой карты пиковой масти, кроме дамы. Наоборот, B \A – извлечение дамы любой масти, кроме пик.

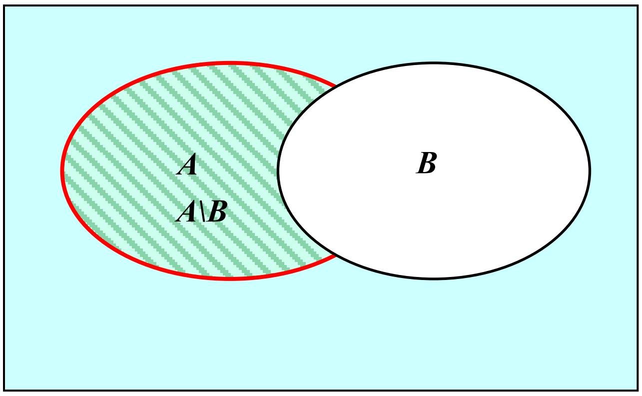


Рис.3.

Если все исходы опыта, при которых происходит событие B, содержатся в множестве исходов, при которых происходит событие A, то говорят, что $B \subseteq A$ (рис. 4).

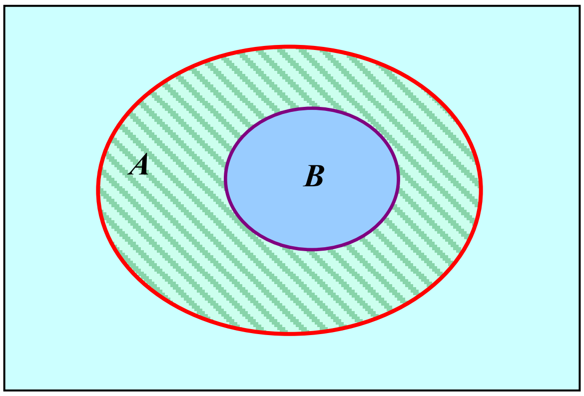


Рис. 4

Введем еще несколько категорий событий.

Определение. События А и В называются **совместными**, если они могут произойти оба в результате одного опыта. В противном случае (то есть если они не могут произойти одновременно) события называются **несовместными.**

Примеры: совместными событиями являются попадания двух стрелков в примере 1 и появление карты пиковой масти и дамы в примере 4; несовместными – события A_1 – A_6 в примере 2.

Замечание 1. Если изобразить графически области исходов опыта, благоприятных несовместным событиям, то они не будут иметь общих точек.

Замечание 2. Из определения несовместных событий следует, что их произведение является невозможным событием.

Определение. Говорят, что события A_1, A_2,…,A_n образуют **полную группу**, если в результате опыта обязательно произойдет хотя бы одно из событий этой группы.

Замечание. В частности, если события, образующие полную группу, попарно несовместны, то в результате опыта произойдет *одно и только одно* из них. Такие события называют **элементарными событиями**.

Пример 7.

В примере 2 события A_1 – A_6 (выпадение одного, двух,…, шести очков при одном броске игральной кости) образуют полную группу несовместных событий.

Определение. События называются **равновозможными**, если нет оснований считать, что одно из них является более возможным, чем другое.

Примеры: выпадение любого числа очков при броске игральной кости, появление любой карты при случайном извлечении из колоды, выпадение герба или цифры при броске монеты и т.п.

Классическое определение вероятности

При изучении случайных событий возникает необходимость количественно сравнивать возможность их появления в результате опыта. Например, при последовательном извлечении из колоды пяти карт более возможна ситуация, когда появились карты разных мастей, чем появление пяти карт одной масти; при десяти бросках монеты более возможно чередование гербов и цифр, нежели выпадение подряд десяти гербов, и т.д. Поэтому с каждым таким событием связывают по определенному правилу некоторое число, которое тем больше, чем более возможно событие. Это число называется **вероятностью события** и является вторым основным понятием теории вероятностей.

Отметим, что само понятие вероятности, как и понятие случайного события, является аксиоматическим и поэтому не поддается строгому определению. То, что в дальнейшем будет называться различными определениями вероятности, представляет собой способы вычисления этой величины.

Если все события, которые могут произойти в результате данного опыта,

а) попарно несовместны;

б) равновозможны;

в) образуют полную группу,

то говорят, что имеет место **схема случаев.**

Можно считать, что случаи представляют собой все множество исходов опыта. Пусть их число равно n (число возможных исходов), а при m из них происходит некоторое событие А (число благоприятных исходов).

Определение. **Вероятностью события А** называется отношение числа исходов опыта, благоприятных этому событию, к числу возможных исходов:

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

- *классическое определение вероятности*.

Свойства вероятности

Из определения вытекают следующие свойства вероятности:

Свойство 1. Вероятность достоверного события равна единице.

Доказательство. Так как достоверное событие всегда происходит в результате опыта, то все исходы этого опыта являются для него благоприятными, то есть $m = n$, следовательно,

$P(A) = 1$.

Свойство 2. Вероятность невозможного события равна нулю.

Доказательство. Для невозможного события ни один исход опыта не является благоприятным, поэтому $m = 0$ и $p(A) = 0$.

Свойство 3. Вероятность случайного события есть число, заключенное между нулем и единицей.

Доказательство. Случайное событие происходит при некоторых исходах опыта, но не при всех, следовательно, $0 < m < n$, и из (1) следует, что $0 < p(A) < 1$.

Пример 8.

Из урны, содержащей 6 белых и 4 черных шара, наудачу вынут шар. Найти вероятность того, что он белый.

Решение.
Будем считать элементарными событиями, или исходами опыта, извлечение из урны каждого из имеющихся в ней шаров. Очевидно, что эти события удовлетворяют всем условиям, позволяющим считать их схемой случаев. Следовательно, число возможных исходов равно 10, а число исходов, благоприятных событию А (появлению белого шара) – 6 (таково количество белых шаров в урне). Значит,

$$p(A)=\frac{m}{n}=\frac{6}{10}=0,6.$$

Относительная частота. Статистическое определение вероятности

Классическое определение вероятности применимо только для очень узкого класса задач, где все возможные исходы опыта можно свести к схеме случаев. В большинстве реальных задач эта схема неприменима. В таких ситуациях требуется определять вероятность события иным образом. Для этого введем вначале понятие **относительной частоты W(A)** события А как отношения числа опытов, в которых наблюдалось событие А, к общему количеству проведенных испытаний:

$$W(A)=\frac{M}{N}, \qquad (2)$$

где N – общее число опытов, M – число появлений события A.
Большое количество экспериментов показало, что если опыты проводятся в одинаковых условиях, то для большого количества испытаний относительная частота изменяется мало, колеблясь около некоторого постоянного числа. Это число можно считать вероятностью рассматриваемого события.

Определение. **Статистической вероятностью события** считают число, вокруг которого группируются относительные частоты.

Замечание 1. Из формулы (2) следует, что свойства вероятности, доказанные для ее классического определения, справедливы и для статистического определения вероятности.
Замечание 2. Для существования статистической вероятности события A требуется:
1) возможность производить неограниченное число испытаний;
2) устойчивость относительных частот появления A в различных сериях достаточно большого числа опытов.

Замечание 3. Недостатком статистического определения является некоторая неоднозначность значения статистической вероятности.

Пример 9.
Если в задаче задается вероятность попадания в мишень для данного стрелка (скажем, p = 0,7), то эта величина получена в результате изучения статистики большого количества серий выстрелов, в которых этот стрелок попадал в мишень около семидесяти раз из каждой сотни выстрелов.

Основные формулы комбинаторики

При вычислении вероятностей часто приходится использовать некоторые формулы *комбинаторики* – науки, изучающей комбинации, которые можно составить по определенным правилам из элементов некоторого конечного множества. Определим основные такие комбинации.

Перестановки – это комбинации, составленные из всех n элементов данного множества и отличающиеся только порядком их расположения. Число всех возможных перестановок

$$P_n = n! \qquad (3)$$

Пример 10.
Сколько различных списков (отличающихся порядком фамилий) можно составить из 7 различных фамилий?
Решение.

$$P_7 = 7! = 2\cdot 3\cdot 4\cdot 5\cdot 6\cdot 7 = 5040.$$

Размещения – комбинации из m элементов множества, содержащего n различных элементов, отличающиеся либо составом элементов, либо их порядком. Число всех возможных размещений

$$A_n^m = n(n-1)(n-2)...(n-m+1). \qquad (4)$$

Пример 11.
Сколько возможно различных вариантов пьедестала почета (первое, второе, третье места), если в соревнованиях принимают участие 10 человек?
Решение.

$$A_{10}^3 = 10\cdot 9\cdot 8 = 720.$$

Сочетания – неупорядоченные наборы из m элементов множества, содержащего n различных элементов (то есть наборы, отличающиеся только составом элементов). Число сочетаний

$$C_n^m = \frac{n!}{m!(n-m)!}. \qquad (5)$$

Пример 12.
В отборочных соревнованиях принимают участие 10 человек, из которых в финал выходят трое. Сколько может быть различных троек финалистов?
Решение.
В отличие от предыдущего примера, здесь не важен порядок финалистов, следовательно, ищем число сочетаний из 10 по 3:

$$C_{10}^3 = \frac{10!}{3!\cdot 7!} = \frac{8\cdot 9\cdot 10}{6} = 120.$$

ПРИМЕРЫ РЕШЕНИЯ ЗАДАЧ

Задача 1.
Среди студентов, собравшихся на лекцию по теории вероятностей, выбирают наудачу одного. Пусть событие B заключается в том, что он не курит, а событие C в том, что он живет в общежитии. Когда будет справедливо соотношение

$$\overline{C} \subseteq B?$$

УКАЗАНИЕ

Событие, противоположное событию C, происходит только в случае, если произошло событие B.

РЕШЕНИЕ

Событие $\overline{C}$ заключается в том, что выбранный студент НЕ живет в общежитии. Оно должно происходить всегда, если произошло событие B, следовательно, все студенты, не живущие в общежитии, не курят. Однако событие B может произойти и в случае, если $\overline{C}$ не произошло, то есть нельзя утверждать, что не курят только студенты, не живущие в общежитии. Следовательно, верный ответ: все студенты, не живущие в общежитии, не курят.
ОТВЕТ: все студенты, не живущие в общежитии, не курят.

Задача 2.
Пусть A, B, C – три произвольно выбранных события. Найти выражение для события, состоящего в том, что из A, B, C произошло не более двух событий.

УКАЗАНИЕ

Фраза «не более двух событий» означает не то, что произошли именно ДВА события, а то, что не произошли все три – любой другой результат удовлетворяет условиям задачи.

РЕШЕНИЕ

По условию задачи требуется, чтобы не произошли все три события вместе. Это требование можно сформулировать как событие, противоположное произведению ABC, то есть $\overline{ABC}$, или как условие, что произошло хотя бы одно из событий, противоположных данным: $\overline{A}+\overline{B}+\overline{C}$. Поэтому правильными являются ответы 1 и 4.

ОТВЕТ: $\overline{ABC}$ или $\overline{A}+\overline{B}+\overline{C}$.

Задача 3.

Какова вероятность того, что наудачу взятую кость домино можно приставить к данной: (2; 5)?

УКАЗАНИЕ

Воспользуйтесь классическим определением вероятности:

$$p(A)=\frac{m}{n},$$

где n – число возможных исходов опыта, а m – число исходов, благоприятных событию A. Напоминаем, что в наборе домино 28 костей, соответствующих неупорядоченным парам чисел (0,0), (0,1),…, (6,6).

РЕШЕНИЕ

Поскольку одна кость извлечена из набора, число n возможных исходов опыта равно 27 – числу оставшихся костей. Каждое из чисел 2 и 5 присутствует на семи костях, одна из которых: (2,5) – общая и в следующем выборе не участвует. Следовательно, в наборе осталось 6 + 6 = 12 костей, каждую из которых можно приставить к данной, то есть число m благоприятных исходов равно 12. Поэтому

$$p=\frac{m}{n}=\frac{12}{27}=\frac{4}{9}.$$

ОТВЕТ: $\frac{4}{9}$.

Задача 4.

Найти вероятность того, что при трех бросках монеты герб выпадет 1 раз.

УКАЗАНИЕ

Воспользуйтесь классическим определением вероятности:

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

Число исходов опыта в решаемой задаче невелико, и их можно пересчитать вручную.

РЕШЕНИЕ

Каждый бросок монеты может иметь 2 различных исхода, поэтому общее число n исходов при трех бросках равно $2 \cdot 2 \cdot 2 = 8$. Очевидно, что при этом выпадение ровно одного герба возможно трижды: при первом, втором или третьем броске. Следовательно, $m = 3$, а вероятность

$$p=\frac{m}{n}=\frac{3}{8}.$$

ОТВЕТ: $\frac{3}{8}$.

Задача 5.

Сколькими способами можно расставить в одну шеренгу 6 человек?

УКАЗАНИЕ

Требуется найти число перестановок из 6 элементов.

РЕШЕНИЕ

Первого человека в шеренге можно выбрать шестью способами, второго – пятью и так далее, то есть число способов

$$N=P_6=6!=1\cdot 2\cdot 3\cdot 4\cdot 5\cdot 6=720.$$

ОТВЕТ: 720.

Задача 6.

Сколько различных трехзначных чисел можно составить, используя по одному разу цифры 1,2,3,4,5,6,7,8?

УКАЗАНИЕ

Требуется найти число размещений из 8 по 3.

РЕШЕНИЕ

Первую цифру можно выбрать 8 способами, вторую – 7, третью – 6, то есть

$$N=A_8^3=\frac{8!}{(8-3)!}=\frac{8!}{5!}=\frac{1\cdot 2\cdot 3\cdot 4\cdot 5\cdot 6\cdot 7\cdot 8}{1\cdot 2\cdot 3\cdot 4\cdot 5}=$$

$$=6\cdot 7\cdot 8=336.$$

ОТВЕТ: 336.

Задача 7.

Сколько различных букетов, состоящих из трех цветов, можно составить, если имеется 10 цветов разных сортов?

УКАЗАНИЕ
Требуется найти число сочетаний из 10 по 3.

РЕШЕНИЕ
Букеты различаются только составом входящих в них цветов, следовательно, это неупорядоченные наборы, поэтому

$$N = C_{10}^{3} = \frac{10!}{3!(10-3)!} = \frac{10!}{3!\cdot 7!} = \frac{8\cdot 9\cdot 10}{1\cdot 2\cdot 3} = 120.$$

ОТВЕТ: 120.

Задача 8.
Из коробки, в которой лежат 6 красных и 5 синих карандашей, случайным образом извлечены 3 карандаша. Найти вероятность того, что все они красные.

УКАЗАНИЕ
Число возможных исходов опыта – число сочетаний из 11 (общее число карандашей в коробке) по 3, а число благоприятных исходов - число сочетаний из 6 (количество карандашей нужного цвета) по 3.

РЕШЕНИЕ

$$n = C_{11}^{3} = \frac{11!}{3!\cdot 8!} = \frac{9\cdot 10\cdot 11}{1\cdot 2\cdot 3} = 165,$$

$$m = C_{6}^{3} = \frac{6!}{3!\cdot 3!} = \frac{4\cdot 5\cdot 6}{1\cdot 2\cdot 3} = 20, \quad p = \frac{m}{n} = \frac{20}{165} = \frac{4}{33}.$$

Задача 9.
Из 10 служащих туристического агентства 7 знают французский язык. Какова вероятность того, что среди трех человек, находящихся в данный момент в офисе, французским владеют двое?

УКАЗАНИЕ
Число благоприятных исходов – это произведение количества способов, которыми можно выбрать двух человек, знающих французский язык, из семи, и способов, которыми можно выбрать одного человека, не знающего этого языка, из оставшихся трех.

РЕШЕНИЕ

$$n = C_{10}^{3} = \frac{10!}{3!\cdot 7!} = 120,$$

$$m = C_{7}^{2}\cdot C_{3}^{1} = \frac{7!}{2!\cdot 5!}\cdot 3 = 63, \quad p = \frac{63}{120} = \frac{21}{40}.$$

ОТВЕТ: $\frac{21}{40}$.

1.1.2. *Геометрическая вероятность. Теоремы сложения и умножения*

Одним из недостатков классического определения вероятности является то, что оно неприменимо к испытаниям с бесконечным количеством исходов. В таких случаях можно воспользоваться понятием **геометрической вероятности**.

Пусть на отрезок L наудачу брошена точка. Это означает, что точка обязательно попадет на отрезок L и с равной возможностью может совпасть с любой точкой этого отрезка. При этом вероятность попадания точки на любую часть отрезка L не зависит от расположения этой части на отрезке и пропорциональна его длине. Тогда вероятность того, что брошенная точка попадет на отрезок l, являющийся частью отрезка L, вычисляется по формуле:

$$p = \frac{|l|}{|L|},$$

где $|l|$ – длина отрезка l, а $|L|$ – длина отрезка L.

Можно дать аналогичную постановку задачи для точки, брошенной на плоскую область S и вероятности того, что она попадет на часть этой области s:

$$p = \frac{|s|}{|S|},$$

где $|s|$ – площадь части области, а $|S|$ – площадь всей области.

В трехмерном случае вероятность того, что точка, случайным образом расположенная в теле V, попадет в его часть v, задается формулой:

$$p = \frac{|v|}{|V|},$$

где $|v|$ – объем части тела, а $|V|$ – объем всего тела.

Пример 1.

Найти вероятность того, что точка, наудачу брошенная в круг, не попадет в правильный шестиугольник, вписанный в него.

Решение.
Пусть радиус круга равен R , тогда сторона шестиугольника тоже равна R. При этом площадь круга $S = \pi R^2$, а площадь шестиугольника

$$s = \frac{3\sqrt{3}}{2} R^2.$$

Следовательно,

$$p = \frac{S - s}{S} = \frac{\pi R^2 - \frac{3\sqrt{3}}{2} R^2}{\pi R^2} = \frac{\pi - 3\sqrt{3}}{2\pi} \approx 0{,}174.$$

Пример 2.
На отрезок AB случайным образом брошены три точки: C, D и M. Найти вероятность того, что из отрезков AC, AD и AM можно построить треугольник.
Решение.
Обозначим длины отрезков AC, AD и AM через x, y и z и рассмотрим в качестве возможных исходов множество точек трехмерного пространства с координатами (x, y, z). Если принять длину отрезка равной 1, то эти множество возможных исходов представляет собой куб с ребром, равным 1. Тогда множество благоприятных исходов состоит из точек, для координат которых выполнены неравенства треугольника: $x + y > z$, $x + z > y$, $y + z > x$. Это часть куба, отрезанная от него плоскостями $x + y = z$, $x + z = y$, $y + z = x$

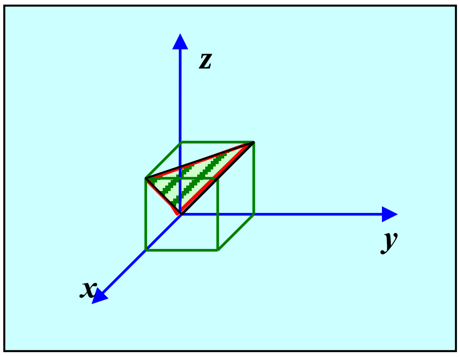


Рис. 4

(одна из них, плоскость $x + y = z$, проведена на рис.4). Каждая такая плоскость отделяет от куба пирамиду, объем которой равен

$$\frac{1}{3} \cdot \frac{1}{2} \cdot 1 = \frac{1}{6}.$$

Следовательно, объем оставшейся части

$$v = 1 - 3 \cdot \frac{1}{6} = \frac{1}{2}.$$

Тогда

$$p = \frac{v}{V} = \frac{1}{2} : 1 = \frac{1}{2}.$$

Теорема сложения вероятностей

Теорема сложения. Вероятность $p(A + B)$ суммы событий A и B равна

$$p(A+B)=p(A)+p(B)-p(AB). \quad (1)$$

Доказательство.
Докажем теорему сложения для схемы случаев. Пусть n – число возможных исходов опыта, m_A – число исходов, благоприятных событию A, m_B – число исходов, благоприятных событию B, а m_{AB} – число исходов опыта, при которых происходят оба события (то есть исходов, благоприятных произведению AB). Тогда число исходов, при которых имеет место событие $A + B$, равно $m_A + m_B - m_{AB}$ (так как в сумме $(m_A + m_B)$ m_{AB} учтено дважды: как исходы, благоприятные A, и исходы, благоприятные B). Следовательно, вероятность суммы можно определить по формуле (1):

$$p(A+B)=\frac{m_A+m_B-m_{AB}}{n}=\frac{m_A}{n}+\frac{m_B}{n}-\frac{m_{AB}}{n}=p(A)+p(B)-p(AB),$$

что и требовалось доказать.

Следствие 1. Теорему сложения можно распространить на случай суммы любого числа событий. Например, для суммы трех событий A, B и C

$$P(A + B + C) = p(A) + p(B) + p(C) - p(AB) - p(AC) - p(BC) + p(ABC)$$

и т.д.

Следствие 2. Если события A и B несовместны, то $m_{AB} = 0$, и, следовательно, вероятность суммы несовместных событий равна сумме их вероятностей:

$$p(A+B)=p(A)+p(B). \quad (2)$$

Определение. **Противоположными событиями** называют два несовместных события, образующих полную группу. Если одно из них назвать A, то второе принято обозначать $\bar{A}$.

Замечание. Таким образом, $\bar{A}$ заключается в том, что событие A не произошло.

Теорема. Сумма вероятностей противоположных событий равна 1:

$$p(A)+p(\bar{A})=1. \qquad (3)$$

Доказательство.

Так как A и $\bar{A}$ образуют полную группу, то одно из них обязательно произойдет в результате опыта, то есть событие $A+\bar{A}$ является достоверным. Следовательно, $P(A+\bar{A}) = 1$. Но, так как A и $\bar{A}$ несовместны, из (7) следует, что $P(A+\bar{A}) = p(A) + p(\bar{A})$.

Значит, $p(A) + p(\bar{A}) = 1$, что и требовалось доказать.

Замечание. В ряде задач проще искать не вероятность заданного события, а вероятность события, противоположного ему, а затем найти требуемую вероятность по формуле (3).

Пример 3.

Из урны, содержащей 2 белых и 6 черных шаров, случайным образом извлекаются 5 шаров. Найти вероятность того, что вынуты шары разных цветов.

Решение.

Событие $\bar{A}$, противоположное заданному, заключается в том, что из урны вынуто 5 шаров одного цвета, а так как белых шаров в ней всего два, то этот цвет может быть только черным. Множество возможных исходов опыта найдем по формуле для числа сочетаний:

$$n=C_8^5=\frac{8!}{5!\cdot 3!}=\frac{6\cdot 7\cdot 8}{6}=56,$$

а множество исходов, благоприятных событию $\bar{A}$ - это число возможных наборов по 5 шаров только из шести черных:

$$m_{\bar{A}}=C_6^5=6.$$

Тогда $p(\bar{A})=\frac{6}{56}=\frac{3}{28}$, а $p(A)=1-\frac{3}{28}=\frac{25}{28}$.

Теорема умножения вероятностей

Определение. Назовем **условной вероятностью $p_A(B)$ события B** вероятность события B при условии, что событие A произошло.

Замечание.

Понятие условной вероятности используется в основном в случаях, когда осуществление события A изменяет вероятность события B.

Пример 4.

Пусть событие A – извлечение из колоды в 32 карты туза, а событие B – то, что и вторая вынутая из колоды карта окажется тузом. Тогда, если после первого раза карта была возвращена в колоду, то вероятность вынуть вторично туз не меняется: $p(B)=p(A)=\dfrac{4}{32}=\dfrac{1}{8}=0{,}125.$ Если же первая карта в колоду не возвращается, то осуществление события A приводит к тому, что в колоде осталась 31 карта, из которых только 3 туза. Поэтому

$$p_A(B)=\frac{3}{31}\approx 0{,}097.$$

Пример 5.

Если событие A – попадание в самолет противника при первом выстреле из орудия, а B – при втором, то первое попадание уменьшает маневренность самолета, поэтому $p_A(B)$ увеличится по сравнению с $p(A)$.

Теорема умножения. Вероятность произведения двух событий равна произведению вероятности одного из них на условную вероятность другого при условии, что первое событие произошло:

$$p(AB)=p(A)\cdot p_A(B). \qquad (4)$$

Доказательство.

Воспользуемся обозначениями теоремы сложения. Тогда для вычисления $p_A(B)$ множеством возможных исходов нужно считать m_A (так как A произошло), а множеством благоприятных исходов – те, при которых произошли и A, и B (m_{AB}). Следовательно,

$$p_A(B)=\frac{m_{AB}}{m_A}=\frac{m_{AB}}{n}\cdot\frac{n}{m_A}=p(AB):p(A),$$

откуда следует утверждение теоремы.

Пример 6.

Для поражения цели необходимо попасть в нее дважды. Вероятность первого попадания равна 0,2, затем она не меняется при промахах, но после первого попадания увеличивается вдвое. Найти вероятность того, что цель будет поражена первыми двумя выстрелами.

Решение.

Пусть событие A – попадание при первом выстреле, а событие B – попадание при втором. Тогда

$$p\,(A) = 0{,}2,\ p_A(B) = 0{,}4,\ p\,(AB) = 0{,}2\cdot 0{,}4 = 0{,}08.$$

Следствие. Если подобным образом вычислить вероятность события BA, совпадающего с событием AB, то получим, что

$$p(BA)=p(B)\cdot p_B(A).$$

Следовательно,

$$p(A)\cdot p_A(B)=p(B)\cdot p_B(A).\quad (5)$$

Определение. Событие B называется **независимым** от события A, если появление события A не изменяет вероятности B, то есть $p_A(B) = p\ (B)$.

Замечание. Если событие B не зависит от A, то и A не зависит от B. Действительно, из (10) следует при этом, что

$$p\ (A)\cdot p\ (B) = p\ (B)\cdot p_B(A),$$

откуда $p_B(A) = p\ (A)$. Значит, **свойство независимости событий взаимно**.

Теорема умножения для независимых событий имеет вид:

$$p(AB)=p(A)\cdot p(B),\qquad (6)$$

то есть вероятность произведения независимых событий равна произведению их вероятностей.

При решении задач теоремы сложения и умножения обычно применяются вместе.

Пример 7.

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 H_1 – попадание первого стрелка, H_2 – попадание второго. Тогда

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

События H_1 и H_2 совместны и независимы, поэтому теорема сложения применяется в общем виде, а теорема умножения – в виде (6). Следовательно, $p(C)$ = 0,6·0,7 = 0,42, $p(A)$ = 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), $p(D)$ = 0,4·0,3 = 0,12. Заметим, что события A и D являются противоположными, поэтому $p(A) = 1 – p(D)$.

Вероятность появления хотя бы одного события

Теорема. Вероятность появления хотя бы одного из попарно независимых событий $A_1, A_2,\ldots, A_n$ равна

$$p(A)=1-q_1q_2\ldots q_n,\qquad (7)$$

где q_i – вероятность события $\overline{A}_i$, противоположного событию A_i.
Доказательство.
Если событие A заключается в появлении хотя бы одного события из A_1, A_2,…, A_n, то события A и $\overline{A}_1\overline{A}_2...\overline{A}_n$ противоположны, поэтому сумма их вероятностей равна 1. Кроме того, поскольку A_1, A_2,…, A_n независимы, то независимы и $\overline{A}_1, \overline{A}_2, ..., \overline{A}_n$, следовательно,

$$p\left(\overline{A}_1\overline{A}_2...\overline{A}_n\right) = p(\overline{A}_1)p(\overline{A}_2)...p(\overline{A}_n) = q_1q_2...q_n.$$

Отсюда следует справедливость формулы (7).

Пример 8.
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 вероятности противоположного события (выпадения цифры) и равна 0,5. Тогда вероятность выпадения хотя бы одного герба при n выстрелах равна 1-(0,5)^n . Тогда из решения неравенства 1- (0,5)^n > 0,9 следует, что n > $\log_2 10 \geq 4$.

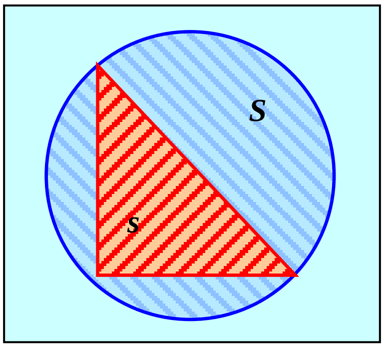
ПРИМЕРЫ РЕШЕНИЯ ЗАДАЧ

Задача 1.
Найти вероятность того, что точка, брошенная наудачу в круг, попадет в равнобедренный прямоугольный треугольник, вписанный в этот круг.

УКАЗАНИЕ

Искомая вероятность равна отношению площадей треугольника и круга.

РЕШЕНИЕ



Используем формулу геометрической вероятности:

$$p = \frac{s}{S},$$

где S – мера (в данном случае площадь) области возможных исходов, а s – площадь области благоприятных исходов.

Если радиус круга равен R, то гипотенуза прямоугольного треугольника является диаметром окружности и и равна $2R$, а высота, опущенная на гипотенузу, равна R. Следовательно,

$$s_{\Delta} = \frac{1}{2} \cdot 2R \cdot R = R^2, \quad S_{круга} = \pi R^2,$$

$$p = \frac{s_{\Delta}}{S_{круга}} = \frac{R^2}{\pi R^2} = \frac{1}{\pi}.$$

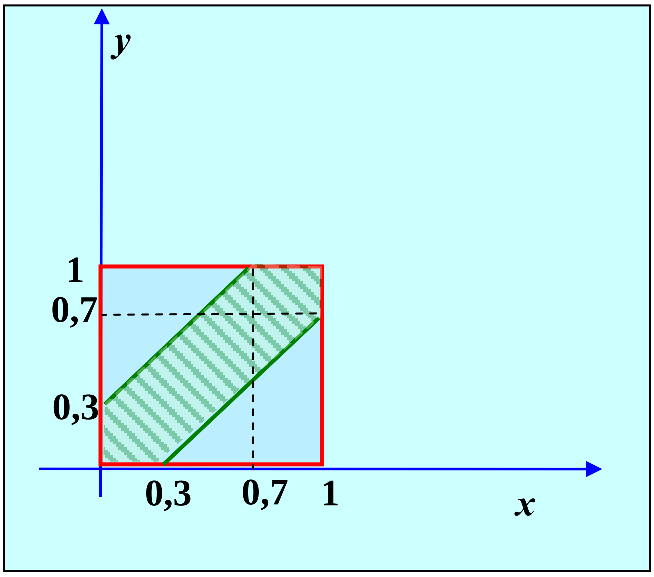
Ответ: $\frac{1}{\pi}$.

Задача 2.
Числа x и y могут равновозможно принимать любые значения из промежутка [0;1]. Найти вероятность того, что $|x - y| < 0,3$.

Указание

Будем считать, что данные числа являются координатами точки на плоскости. Тогда множеством возможных исходов является квадрат со стороной, равной единице, а множество благоприятных исходов – область, заданная неравенством $|x - y| < 0,3$.

Решение



Зададим множество возможных исходов опыта в виде квадрата $0 \leq x \leq 1$, $0 \leq y \leq 1$. Тогда множество благоприятных исходов представляет собой часть этого квадрата, расположенную между прямыми $x - y = 0,3$ и $x - y = -0,3$. Найдем площадь этой области как разность площади квадрата, равной единице, и суммы площадей прямоугольных треугольников, катеты которых равны 0,7:

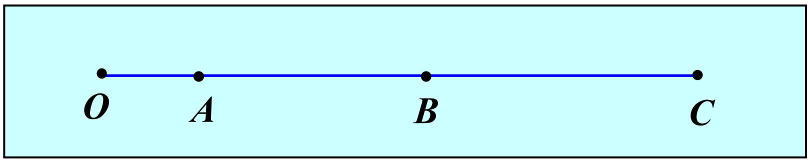
$$s = 1 - 2 \cdot \frac{1}{2} \cdot 0{,}7^2 = 1 - 0{,}49 = 0{,}51.$$

$$p = \frac{s}{S} = \frac{0{,}51}{1} = 0{,}51.$$

Ответ: 0,51.

Задача 3.
На отрезке длиной l наудачу ставятся две точки. Найти вероятность того, что из полученных трех отрезков можно построить треугольник.

Указание



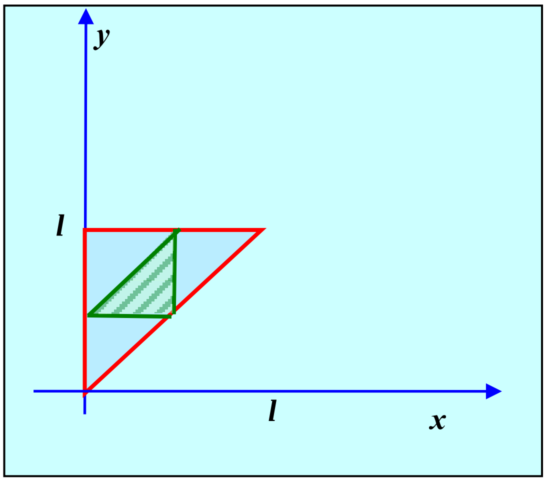
Пусть OC – данный отрезок, A и B – поставленные на нем точки. Рассмотрите в качестве множества возможных исходов прямоугольный треугольник с катетами, равными l, в котором координаты каждой точки равны длинам отрезков OA и OB. Множество возможных исходов задайте с помощью неравенства треугольника (сумма любых двух сторон треугольника больше третьей стороны).

Решение

Пусть OC – данный отрезок, A и B – поставленные на нем точки. Обозначим длины отрезков OA и OB соответственно за x и y ($y > x$). Тогда множество возможных исходов опыта представляет собой треугольник $0 \le x \le l$, $0 \le y \le l$, $y > x$. Определим множество благоприятных исходов с помощью неравенства треугольника:

$$\begin{cases} OA + AB > BC \\ OA + BC > AB \\ AB + BC > OA \end{cases} \Rightarrow \begin{cases} y > l - y \\ x + l - y > y - x \\ l - x > x \end{cases} \Rightarrow$$

$$\Rightarrow \begin{cases} y > \dfrac{l}{2} \\ y - x < \dfrac{l}{2}. \\ x < \dfrac{l}{2} \end{cases}$$



Из рисунка видно, что множество благоприятных исходов является треугольником, площадь которого в 4 раза меньше площади множества возможных исходов. Следовательно, искомая вероятность равна 1/4.

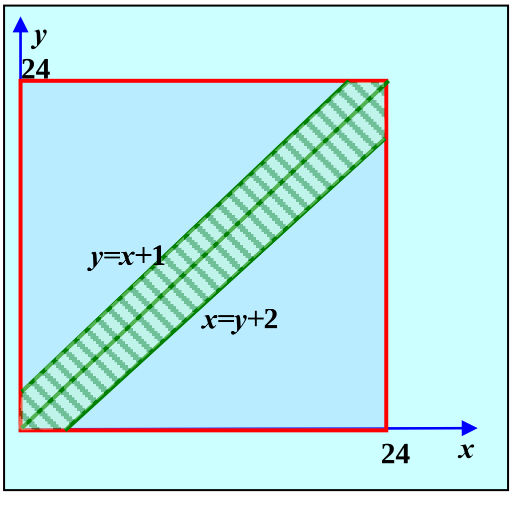
Ответ: $\frac{1}{4}$.

Задача 4.
Два парохода должны подойти к одному и тому же причалу. Время прихода обоих пароходов независимо и равновозможно в течение данных суток. Найти вероятность того, что одному из пароходов придется ожидать освобождения причала, если время стоянки первого парохода – один час, а второго – два часа (ответ округлить до третьего знака после запятой).

Указание

Примите за x время прибытия первого парохода, за y – время прибытия второго, и рассмотрите в качестве множества возможных исходов опыта квадрат $0 \le x \le 24$, $0 \le y \le 24$.

Решение



Если принять за x время прибытия первого парохода, за y – время прибытия второго, то множество возможных исходов опыта является квадратом $0 \le x \le 24$, $0 \le y \le 24$. Множество благоприятных исходов задается условиями $x \le y \le x + 1$, $y \le x \le y + 2$.
Найдем площадь полученной области:

$$s = 24^2 - \frac{1}{2} \cdot 23^2 - \frac{1}{2} \cdot 22^2 = 69,5.$$

$$S = 24^2 = 576, \quad p = \frac{s}{S} = \frac{69,5}{576} \approx 0,121.$$

Ответ: 0,121.

Задача 5.
Из полного набора костей домино наудачу извлекаются три кости. Найти вероятность того, что все они – дубли.

Указание

Примените теорему умножения для зависимых событий.

Решение

Будем считать, что кости извлекаются по одной. Событием A назовем извлечение дубля в первый раз, событием B – извлечение второго дубля, событием C – третьего. Эти события являются зависимыми, так как после извлечения каждой очередной кости изменяется и число оставшихся костей (28, затем 27, затем 26), и количество дублей среди них (7, затем 6, затем 5).
Нам требуется найти вероятность произведения ABC.

$$p(A) = \frac{7}{28} = \frac{1}{4}; \quad p_A(B) = \frac{6}{27} = \frac{2}{9}; \quad p_{AB}(C) = \frac{5}{26};$$

$$p(ABC) = p(A) \cdot p_A(B) \cdot p_{AB}(C) = \frac{1}{4} \cdot \frac{2}{9} \cdot \frac{5}{26} = \frac{5}{468}.$$

Ответ: $\frac{5}{468}$.

Задача 6.
В первой урне находится 5 белых шаров, 11 черных и 8 красных, во второй соответственно 10, 8 и 6. Из каждой урны наудачу извлекается по одному шару. Найти вероятность того, что они одного цвета (ответ округлить до третьего знака после запятой).

Указание

Примените теорему сложения для трех несовместных событий: A – извлечены белые шары, B – черные, C – красные.

Решение

Событие, вероятность которого требуется определить, представляет собой сумму для трех несовместных событий: A – извлечены белые шары, B – черные, C – красные. Каждое их них, в свою очередь, является произведением двух независимых событий: извлечение шара нужного цвета из первой и второй урны. Применим теорему умножения для независимых событий и теорему сложения для несовместных событий:

$$p(A)=\frac{5}{24}\cdot\frac{10}{24}=\frac{25}{288};\quad p(B)=\frac{11}{24}\cdot\frac{8}{24}=\frac{11}{72};$$

$$p(C)=\frac{8}{24}\cdot\frac{6}{24}=\frac{1}{12};$$

$$p(A+B+C)=\frac{25}{288}+\frac{11}{72}+\frac{1}{12}=\frac{31}{96}\approx 0,323.$$

Ответ: 0,323.

Задача 7.

Найти вероятность выпадения хотя бы одной шестерки при четырех бросках игральной кости (ответ округлить до третьего знака после запятой).

Указание

Найдите вначале вероятность события, противоположного данному.

Решение

Событие, противоположное данному, заключается в том, что при четырех бросках шестерка не выпала ни разу. Найдем его вероятность, используя теорему умножения для независимых событий:

$$p(\bar{A})=\left(\frac{5}{6}\right)^4=\frac{625}{1296};\quad p(A)=1-p(\bar{A})=\frac{671}{1296}\approx 0,518.$$

Ответ: 0,518.

1.1.3. *Формула полной вероятности и формула Байеса. Формула Бернулли*

Пусть событие A может произойти только совместно с одним из событий $H_1, H_2,\dots, H_n$, образующих полную группу несовместных событий. Тогда события $H_1, H_2,\dots, H_n$ называются **гипотезами.**

Теорема. Вероятность события A, наступающего совместно с гипотезами $H_1, H_2,\dots, H_n$, равна:

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

где $p(H_i)$ – вероятность i- й гипотезы, а $p(A/H_i)$ – вероятность события A при условии реализации этой гипотезы. Формула (1) носит название **формулы полной вероятности**.

Доказательство.

Можно считать событие A суммой попарно несовместных событий AH_1, $AH_2, \ldots, AH_n$. Тогда из теорем сложения и умножения следует, что

$$p(A) = p(AH_1 + AH_2 + \ldots + AH_n) =$$

$$= p(AH_1) + p(AH_2) + \ldots + p(AH_n) = \sum_{i=1}^{n} p(H_i)p(A/H_i),$$

что и требовалось доказать.

Пример 1.

Имеются три одинаковые урны с шарами. В первой из них 3 белых и 4 черных шара, во второй – 2 белых и 5 черных, в третьей – 10 черных шаров. Из случайно выбранной урны наудачу вынут шар. Найти вероятность того, что он белый.

Решение.

Будем считать гипотезами H_1, H_2 и H_3 выбор урны с соответствующим номером. Так как по условию задачи все гипотезы равновозможны, то

$$p(H_1) = p(H_2) = p(H_3) = \frac{1}{3}.$$

Найдем условную вероятность A при реализации каждой гипотезы:

$$p(A/H_1) = \frac{3}{7}, p(A/H_2) = \frac{2}{7}, \ p(A/H_3) = 0.$$

Тогда

$$p(A) = \frac{1}{3} \cdot \frac{3}{7} + \frac{1}{3} \cdot \frac{2}{7} + \frac{1}{3} \cdot 0 = \frac{5}{21} \approx 0,238.$$

Формула Байеса (теорема гипотез)

Пусть известен результат опыта, а именно то, что произошло событие A. Этот факт может изменить априорные (то есть известные до опыта) вероятности гипотез. Например, в предыдущем примере извлечение из урны белого шара говорит о том, что этой урной не могла быть третья, в которой нет белых шаров, то есть $p\ (H_3/A) = 0$. Для переоценки вероятностей гипотез при известном результате опыта используется **формула Байеса**:

$$p(H_i / A) = \frac{p(H_i)p(A/H_i)}{p(A)}. \qquad (2)$$

Действительно, из теоремы умножения получим, что

$$p(A)p(H_i / A) = p(H_i)p(A/H_i),$$

откуда следует справедливость формулы (2).

Пример 2.

После двух выстрелов двух стрелков, вероятности попаданий которых равны 0,6 и 0,7, в мишени оказалась одна пробоина. Найти вероятность того, что попал первый стрелок.

Решение.

Пусть событие A – одно попадание при двух выстрелах, а гипотезы: H_1 – первый попал, а второй промахнулся, H_2 – первый промахнулся, а второй попал, H_3 – оба попали, H_4 – оба промахнулись. Вероятности гипотез: $p(H_1) = 0,6 \cdot 0,3 = 0,18$, $p(H_2) = 0,4 \cdot 0,7 = 0,28$, $p(H_3) = 0,6 \cdot 0,7 = 0,42$, $p(H_4) = 0,4 \cdot 0,3 = 0,12$. Тогда

$$p(A/H_1) = p(A/H_2) = 1,\; p(A/H_3) = p(A/H_4) = 0.$$

Следовательно, полная вероятность

$$p(A) = 0,18 \cdot 1 + 0,28 \cdot 1 + 0,42 \cdot 0 + 0,12 \cdot 0 = 0,46.$$

Применяя формулу Байеса, получим:

$$p(H_1 / A) = \frac{0,18 \cdot 1}{0,46} = \frac{9}{23} \approx 0,391.$$

Схема повторения испытаний. Формула Бернулли

Рассмотрим серию из n испытаний, в каждом из которых событие A появляется с одной и той же вероятностью p, причем результат каждого испытания не зависит от результатов остальных. Подобная постановка задачи называется **схемой повторения испытаний**. Найдем вероятность того, что в такой серии событие A произойдет ровно $к$ раз (неважно, в какой последовательности). Интересующее нас событие представляет собой сумму равновероятных несовместных событий, заключающихся в том, что A произошло в некоторых $к$ испытаниях и не произошло в остальных $n - к$ испытаниях. Число таких событий равно числу сочетаний из n по $к$, то есть C_n^k, а вероятность каждого из них: $p^k q^{n-k}$, где $q = 1 - p$ – вероятность того, что в данном опыте A не произошло. Применяя теорему сложения для несовместных событий, получим **формулу Бернулли**:

$$\boxed{p_n(k) = C_n^k \cdot p^k \cdot q^{n-k}. \qquad (3)}$$

Пример 3.

Для получения приза нужно собрать 5 изделий с особым знаком на этикетке. Найти вероятность того, что придется купить 10 изделий, если этикетки с этим знаком имеют 5% изделий.

Решение.

Из постановки задачи следует, что последнее купленное изделие имеет особый знак. Следовательно, из предыдущих девяти эти знаки имели 4 изделия. Найдем вероятность этого по формуле Бернулли:

$$p_9(4)=C_9^4\cdot(0,05)^4\cdot(0,95)^5=0,0006092.$$

Тогда p = 0,0006092·0,05 = 0,0000304.

Приближение Пуассона для схемы Бернулли

Формула Бернулли требует громоздких расчетов при большом количестве испытаний. Можно получить более удобную для расчетов приближенную формулу, если при большом числе испытаний вероятность появления A в одном опыте мала, а произведение $np = \lambda$ сохраняет постоянное значение для разных серий опытов (то есть среднее число появлений события A в разных сериях испытаний остается неизменным). Применим формулу Бернулли:

$$p_n(k)=\frac{n(n-1)(n-2)...(n-k+1)}{k!}p^k(1-p)^{n-k}=$$

$$=\frac{n(n-1)...(n-k+1)}{k!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}.$$

Найдем предел полученного выражения при $n\to\infty$:

$$p_n(k)\approx\frac{\lambda^k}{k!}\lim_{n\to\infty}\left(1\cdot\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)...\left(1-\frac{k-1}{n}\right)\left(1-\frac{\lambda}{n}\right)^{n-k}\right)=$$

$$=\frac{\lambda^k}{k!}\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k}=\frac{\lambda^k}{k!}\cdot e^{-\lambda}\cdot 1.$$

Таким образом, **формула Пуассона**

$$p_n(k)=\frac{\lambda^k e^{-\lambda}}{k!}\qquad(4)$$

позволяет найти вероятность k появлений события A для массовых (n велико) и редких (p мало) событий.

ПРИМЕРЫ РЕШЕНИЯ ЗАДАЧ

Задача 1.

В первой урне 4 белых шара и 6 черных, во второй 1 белый и 9 черных, в третьей 10 черных шаров. Из наудачу выбранной урны вынут шар. Найти вероятность того, что он белый.

Указание

Используйте формулу полной вероятности, считая гипотезами выбор урны (все три гипотезы равновероятны).

Решение

Пусть событие A – извлечение белого шара, гипотезы H_1, H_2, H_3 – выбор соответствующей урны. Тогда

$$p(H_1)=p(H_2)=p(H_3)=\frac{1}{3},$$

$$p_{H_1}(A)=\frac{4}{10},\quad p_{H_2}(A)=\frac{1}{10},\quad p_{H_3}(A)=\frac{0}{10}=0,$$

$$p(A)=\frac{1}{3}\cdot\frac{4}{10}+\frac{1}{3}\cdot\frac{1}{10}+\frac{1}{3}\cdot 0=\frac{1}{6}.$$

Ответ: $\frac{1}{6}$.

Задача 2.

В ящике 15 теннисных мячей, из которых 9 новых. Для первой игры наугад берутся три мяча, которые после игры возвращаются в ящик. Для второй игры также наугад берутся три мяча. Найти вероятность того, что все мячи, взятые для второй игры, новые (ответ округлить до третьего знака после запятой).

Указание

Используйте формулу полной вероятности, задавая четыре гипотезы по количеству новых мячей среди трех, выбранных для первой игры.

Решение

Пусть событие A – выбор для второй игры трех новых мячей.

Зададим пространство гипотез следующим образом:

H_1 – все мячи, взятые для первой игры, старые (тогда после первой игры в ящике по-прежнему 9 новых мячей);

H_2 – среди мячей, взятых для первой игры, один новый (тогда после первой игры новых мячей осталось 8);

H_3 – среди мячей, взятых для первой игры, два новых (для второй игры остается 7 новых);

H_4 – все мячи, взятые для первой игры, новые (для второй игры остается 6 новых).

Найдем вероятности гипотез, используя классическое определение вероятности и формулу для числа сочетаний:

$$p(H_1)=\frac{C_6^3}{C_{15}^3}=\frac{20}{455}=\frac{4}{91};\quad p(H_2)=\frac{C_6^2\cdot C_9^1}{C_{15}^3}=\frac{15\cdot 9}{455}=\frac{27}{91};$$

$$p(H_3)=\frac{C_6^1\cdot C_9^2}{C_{15}^3}=\frac{6\cdot 36}{455}=\frac{216}{455};\quad p(H_4)=\frac{C_9^3}{C_{15}^3}=\frac{84}{455}=\frac{12}{65}.$$

Аналогичным образом, учитывая изменение количества новых мячей перед второй игрой в зависимости от выбора гипотезы, найдем условную вероятность события A при реализации каждой гипотезы:

$$p_{H_1}(A)=\frac{C_9^3}{C_{15}^3}=\frac{12}{65},\quad p_{H_2}(A)=\frac{C_8^3}{C_{15}^3}=\frac{56}{455}=\frac{8}{65},$$

$$p_{H_3}(A)=\frac{C_7^3}{C_{15}^3}=\frac{35}{455}=\frac{1}{13},\quad p_{H_1}(A)=\frac{C_6^3}{C_{15}^3}=\frac{4}{91}.$$

Теперь можно найти полную вероятность события A:

$$p(A)=\frac{4}{91}\cdot\frac{12}{65}+\frac{27}{91}\cdot\frac{8}{65}+\frac{216}{455}\cdot\frac{1}{13}+\frac{12}{65}\cdot\frac{4}{91}=$$

$$=\frac{528}{5915}\approx 0,089.$$

Ответ: 0,089.

Задача 3.

На двух станках изготовляются одинаковые детали. Вероятность брака для детали, изготовленной на первом станке, равна 0,02, на втором – 0,05. Из ящика, в котором находятся 20 деталей, изготовленных на первом станке, и 30, изготовленных на втором, случайно выбрана деталь, которая после проверки оказалась доброкачественной. Найти вероятность того, что она изготовлена на втором станке.

Указание

Примените формулу Байеса, выбрав в качестве гипотез изготовление детали на первом или втором станке. Тогда требуется, зная результат опыта, переоценить вероятность гипотезы H_2.

Решение

Пусть гипотеза H_1 – деталь изготовлена на первом станке, H_2 – на втором, событие A – выбранная деталь доброкачественная. Найдем полную вероятность события A:

$$p(H_1)=\frac{20}{50}=0,4,\quad p(H_2)=\frac{30}{50}=0,6,$$

$$p_{H_1}(A)=1-0,02=0,98,\quad p_{H_2}(A)=1-0,05=0,95,$$

$$p(A)=0,4\cdot 0,98+0,6\cdot 0,95=0,962.$$

Применяя формулу Байеса, найдем вероятность гипотезы H_2 при условии, что событие A произошло:

$$p_A(H_2)=\frac{0,6\cdot 0,95}{0,962}=\frac{285}{481}\approx 0,593.$$

Ответ: 0,593.

Задача 4.

В некоторой местности 20% дней в году – дождливые и 80% - солнечные. Прогноз погоды дают два предсказателя. Первый ошибается в 30% случаев, второй – в 10%. Первый утверждает, что завтра будет солнце, второй – что дождь. Чей прогноз сбудется с большей вероятностью?

Указание

Примените формулу Байеса, считая гипотезами солнечную или дождливую погоду на завтрашний день, а событием A – выданные прогнозы.

Решение

Пусть гипотеза H_1 заключается в том, что в выбранный день будет солнечная погода, а гипотеза H_2 – дождливая. Тогда при реализации первой гипотезы прогноз первого предсказателя верен, а второго – ошибочен, а при реализации второй гипотезы – наоборот. Найдем полную вероятность события A (выдачи соответствующих прогнозов):

$$p(H_1)=0{,}8, \quad p(H_2)=0{,}2,$$

$$p_{H_1}(A)=0{,}7\cdot 0{,}1=0{,}07, \quad p_{H_2}(A)=0{,}3\cdot 0{,}9=0{,}27,$$

$$p(A)=0{,}8\cdot 0{,}07+0{,}2\cdot 0{,}27=0{,}11.$$

Итак, вероятность получения именно таких взаимоисключающих прогнозов равна 0,11. Переоценим вероятность первой гипотезы, зная, что событие A произошло. Если она окажется больше 0,5, то с большей вероятностью сбудется прогноз первого предсказателя, если меньше 0,5 – прогноз второго.

$$p_A(H_1)=\frac{0{,}8\cdot 0{,}07}{0{,}11}=\frac{56}{110}>\frac{1}{2} \quad -$$

более вероятно, что сбудется прогноз первого предсказателя.

Ответ: прогноз первого.

Задача 5.

Найти вероятность выпадения 2-х шестерок при 5 бросках игральной кости.

Указание

Воспользуйтесь формулой Бернулли.

Решение

$$n=5, \quad k=2, \quad p=\frac{1}{6}, \quad q=1-\frac{1}{6}=\frac{5}{6},$$

$$p_5(2)=C_5^2\cdot\left(\frac{1}{6}\right)^2\cdot\left(\frac{5}{6}\right)^3=10\cdot\frac{1}{36}\cdot\frac{125}{216}=\frac{625}{3888}\approx 0{,}161.$$

Ответ: 0,161.

Задача 6.

Стрелку выдано 5 патронов для поражения трех мишеней. Вероятность попадания при каждом выстреле равна 0,4. Найти вероятность того, что он израсходует 5 патронов и поразит все цели.

Указание

Событие, вероятность которого требуется определить, заключается в том, что последним, пятым выстрелом поражена последняя, третья мишень. Следовательно, при первых четырех выстрелах было ровно два попадания.

Решение

Представим событие, вероятность которого требуется определить, в виде произведения двух независимых событий: A – два попадания при четырех выстрелах, B – попадание при пятом выстреле. Тогда

$$p(A) = C_4^2 \cdot 0,4^2 \cdot 0,6^2 = 6 \cdot 0,16 \cdot 0,36 = 0,3456;$$

$$p(B) = 0,4; \quad p(AB) = 0,3456 \cdot 0,4 = 0,13824 \approx 0,138.$$

Ответ: 0,138.

Задача 7.

Для поражения цели произвели 7 выстрелов. Вероятность попадания при каждом выстреле равна 0,6. При одном попадании цель выходит из строя с вероятностью 0,1, при двух – с вероятностью 0,5, при трех – всегда выходит из строя. Найти вероятность выхода цели из строя после 7 выстрелов.

Указание

Рассмотрите гипотезы H_1 – одно попадание из семи выстрелов, H_2 – два попадания, H_3 – три или более попаданий. Вероятности гипотез можно найти по формуле Бернулли. Затем для определения вероятности события A – выхода цели из строя – примените формулу полной вероятности.

Решение

Пусть событие A – выход цели из строя. Вероятность этого события зависит от количества попаданий, поэтому рассмотрим гипотезы H_1 – одно попадание из семи выстрелов, H_2 – два попадания, H_3 – три или более попаданий. Вероятности гипотез можно найти по формуле Бернулли:

$$p(H_1) = C_7^1 \cdot 0,6 \cdot 0,4^6 \approx 0,0172;$$

$$p(H_2) = C_7^2 \cdot 0,6^2 \cdot 0,4^5 \approx 0,0774;$$

$$p(H_3) = 1 - p(H_1) - p(H_2) - p(0) \approx$$

$$\approx 1 - 0,0172 - 0,0774 - 0,4^7 \approx 0,9038;$$

$$p(A) = 0,0172 \cdot 0,1 + 0,0774 \cdot 0,5 + 0,9038 \cdot 1 \approx 0,944.$$

Ответ: 0,944.

1.2. СЛУЧАЙНЫЕ ВЕЛИЧИНЫ

1.2.1. *Случайные величины. Функция распределения и плотность вероятности*

Закон распределения и функция распределения дискретной случайной величины

Наряду с понятием случайного события в теории вероятности используется и более удобное понятие *случайной величины*.

> **Случайной величиной** называется величина, принимающая в результате опыта одно из своих возможных значений, причем заранее неизвестно, какое именно.

Будем обозначать случайные величины заглавными буквами латинского алфавита ($X, Y, Z, \ldots$), а их возможные значения – соответствующими малыми буквами (x_i, y_i, …).

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 при попадании.

Можно заметить, что множество возможных значений для перечисленных случайных величин имеет разный вид: для первых двух величин оно конечно (соответственно 6 и 11 значений), для третьей величины множество значений бесконечно и представляет собой множество натуральных чисел, а для четвертой – все точки отрезка, длина которого равна радиусу мишени. Таким образом, 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**, если она принимает отдельные, изолированные возможные значения с определенными вероятностями.

Случайная величина называется **непрерывной**, если множество ее возможных значений целиком заполняет некоторый конечный или бесконечный промежуток.

Дискретные случайные величины

Для задания дискретной случайной величины нужно знать ее возможные значения и вероятности, с которыми принимаются эти значения. Соответствие между ними называется **законом распределения** случайной величины. Он может иметь вид таблицы, формулы или графика.

Таблица, в которой перечислены возможные значения дискретной случайной величины и соответствующие им вероятности, называется **рядом распределения**:

x_i	x_1	x_2	…	x_n	…
p_i	p_1	p_2	…	p_n	…

Заметим, что событие, заключающееся в том, что случайная величина примет одно из своих возможных значений, является достоверным, поэтому $\sum\limits_{i=1}^{n(\infty)} p_i = 1$.

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. Их вероятности найдены в примере, рассмотренном в лекции 3. Следовательно, ряд распределения имеет вид:

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

Графически закон распределения дискретной случайной величины можно представить в виде **многоугольника распределения** – ломаной, соединяющей точки плоскости с координатами (x_i, p_i).

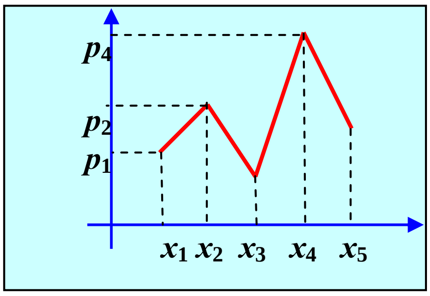


Рис. 1

Функция распределения

Функцией распределения *F*(*x*) случайной величины X называется вероятность того, что случайная величина примет значение, меньшее x:

$$F(X) = p(X < x). \qquad (1)$$

Свойства функции распределения:

1) $0 \le F(x) \le 1$.
Действительно, так как функция распределения представляет собой вероятность, она может принимать только те значения, которые принимает вероятность.
2) Функция распределения является неубывающей функцией, то есть $F(x_2) \ge F(x_1)$ при $x_2 > x_1$. Это следует из того, что $F(x_2) = p(X < x_2) = p(X < x_1) + p(x_1 \le X < x_2) \ge F(x_1)$.
3) $\lim\limits_{x \to -\infty} F(x) = 0$, $\lim\limits_{x \to +\infty} F(x) = 1$. В частности, если все возможные значения X лежат на интервале $[a, b]$, то $F(x) = 0$ при $x \le a$ и $F(x) = 1$ при $x \ge b$. Действительно, $X < a$ – событие невозможное, а $X < b$ – достоверное.
4) Вероятность того, что случайная величина примет значение из интервала $[a, b]$, равна разности значений функции распределения на концах интервала:

$$p\,(\,a \le X < b\,) = F(b) - F(a).$$

Справедливость этого утверждения следует из определения функции распределения (см. свойство 2).
Для дискретной случайной величины значение $F(x)$ в каждой точке представляет собой сумму вероятностей тех ее возможных значений, которые меньше аргумента функции.

Пример 2.
Найдем $F(x)$ для примера 1:

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

Соответственно график функции распределения имеет ступенчатый вид:

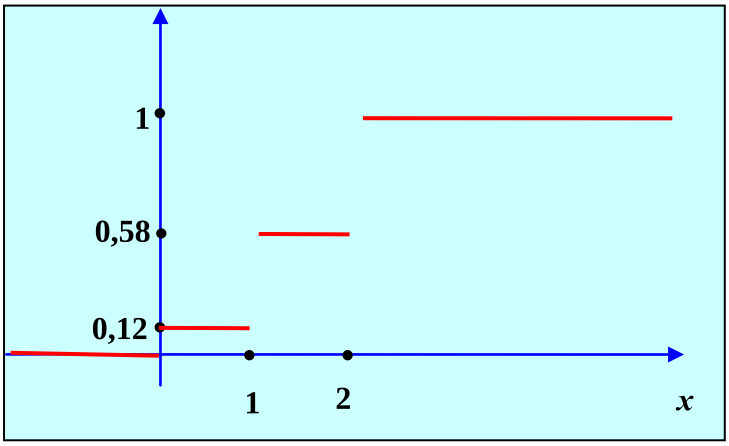


Рис. 2

Функция распределения и плотность распределения непрерывной случайной величины

Определение и свойства функции распределения сохраняются и для непрерывной случайной величины, для которой функцию распределения можно считать одним из видов задания закона распределения. Но для непрерывной случайной величины вероятность каждого отдельного ее значения равна 0. Это следует из свойства 4 функции распределения: $p(X = a) = F(a) - F(a) = 0$. Поэтому для такой случайной величины имеет смысл говорить только о вероятности ее попадания в некоторый интервал.

Вторым способом задания закона распределения непрерывной случайной величины является так называемая плотность распределения (плотность вероятности, дифференциальная функция).

Функция $f(x)$, называемая **плотностью распределения** непрерывной случайной величины, определяется по формуле:

$$f(x) = F'(x), \qquad (2)$$

то есть является производной функции распределения.

Свойства плотности распределения:

1) $f(x) \geq 0$, так как функция распределения является неубывающей.

2) $F(x) = \int_{-\infty}^{x} f(t)dt$, что следует из определения плотности распределения.

3) Вероятность попадания случайной величины в интервал (a, b) (на самом деле речь идет о попадании в полуинтервал [a,b); если функция F(x) непрерывна, то это различие несущественно) определяется формулой

$$p(a \le X < b) = \int_a^b f(x)dx.$$

Действительно,

$$p(a \le X < b) = F(b) - F(a) = \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)dx = \int_a^b f(x)dx.$$

4) $\int_{-\infty}^{+\infty} f(x)dx = 1$ (условие нормировки). Его справедливость следует из того, что

$$\int_{-\infty}^{+\infty} f(x)dx = F(+\infty), \quad a \quad \lim_{x \to +\infty} F(x) = 1.$$

5) $\lim_{x \to \pm\infty} f(x) = 0, \quad так \quad как \quad F(x) \to const \quad при \quad x \to \pm\infty.$

Таким образом, график плотности распределения представляет собой кривую, расположенную выше оси Оx, причем эта ось является ее горизонтальной асимптотой при $x \to \pm\infty$ (последнее справедливо только для случайных величин, множеством возможных значений которых является все множество действительных чисел). Площадь криволинейной трапеции, ограниченной графиком этой функции, равна единице.

Замечание. Если все возможные значения непрерывной случайной величины сосредоточены на интервале $[a, b]$, то все интегралы вычисляются в этих пределах, а вне интервала $[a, b]$ $f(x) \equiv 0$.

Пример 3.

Плотность распределения непрерывной случайной величины задана формулой

$$f(x) = \frac{C}{1+x^2}, \ -\infty < x < +\infty.$$

Найти: а) значение константы C; б) вид функции распределения; в) $p(-1 < x < 1)$.

Решение.

а) значение константы C найдем из свойства 4:

$$\int_{-\infty}^{+\infty} \frac{C}{1+x^2}dx = C arctg x \Big|_{-\infty}^{+\infty} = C\left(\frac{\pi}{2} + \frac{\pi}{2}\right) = C\pi = 1, \text{ откуда } C = \frac{1}{\pi}.$$

$$б)\ F(x) = \frac{1}{\pi}\int_{-\infty}^{x} \frac{1}{1+t^2}dt = \frac{1}{\pi} arctg\ t \Big|_{-\infty}^{x} = \frac{1}{\pi}\left(arctg x + \frac{\pi}{2}\right) = \frac{1}{\pi} arctg x + \frac{1}{2}.$$

$$в)\ p(-1<x<1)=\frac{1}{\pi}\int_{-1}^{1}\frac{1}{1+x^2}dx=\frac{1}{\pi}arctgx\Bigg|_{-1}^{1}=\frac{1}{\pi}\left(\frac{\pi}{4}+\frac{\pi}{4}\right)=0,5.$$

Пример 4.
Функция распределения непрерывной случайной величины имеет вид:

$$F(x)=\begin{cases}0, & x\le 2,\\ \dfrac{x-2}{2}, & 2<x\le 4,\\ 1, & x>4.\end{cases}$$

Найти плотность распределения.
Решение.

$$f(x)=\begin{cases}0', & x\le 2\\ \left(\dfrac{x-2}{2}\right)', & 2<x\le 4\\ 1', & x>4\end{cases}=\begin{cases}0, & x\le 2,\\ 0,5, & 2<x\le 4,\\ 0, & x>4.\end{cases}$$

ПРИМЕРЫ РЕШЕНИЯ ЗАДАЧ

Задача 1.
Случайная величина X – число стандартных деталей среди трех, выбранных из партии, состоящей из 7 стандартных деталей и 3 бракованных. Найти вероятность того, что $X < 2$.

Указание

Найдите вероятности того, что $X = 0$ и $X = 1$.

Решение

$$p(X<2)=p(X=0)+p(X=1).$$

$$p(X=0)=\frac{C_3^3}{C_7^3}=\frac{1}{35};\quad p(X=1)=\frac{C_3^2\cdot C_4^1}{C_7^3}=\frac{12}{35};$$

$$p(X<2)=\frac{1}{35}+\frac{12}{35}=\frac{13}{35}.$$

Ответ: $\frac{13}{35}$.

Задача 2.

Случайная величина X – число мишеней, пораженных после пяти выстрелов, если вероятность попадания при одном выстреле равна 0,4. Найти значение функции распределения $F(3)$.

Указание

$$F(3) = p(0) + p(1) + p(2).$$

Решение

По определению функции распределения

$$F(3) = p(X < 3) = p(0) + p(1) + p(2).$$

$$p(0) = 0,6^5 = 0,07776;$$

$$p(1) = C_5^1 \cdot 0,4 \cdot 0,6^4 = 0,2592;$$

$$p(2) = C_5^2 \cdot 0,4^2 \cdot 0,6^3 = 0,3456;$$

$$F(3) = 0,0776 + 0,2592 + 0,3456 = 0,6824.$$

Ответ: 0,6824.

Задача 3.

Функция распределения непрерывной случайной величины X задана формулами:

$$F(x) = \begin{cases} 0, & x \le 2, \\ x^2 - 4x + 4, & 2 < x < 3, \\ 1, & x \ge 3. \end{cases}$$

Найти плотность вероятности $f(x)$.

Указание

Используйте определение плотности вероятности: $f(x) = F'(x)$.

Решение

$$f(x) = F'(x) = \begin{cases} 0, & x \le 2, \\ 2x - 4, & 2 < x < 3, \\ 0, & x \ge 3. \end{cases}$$

Ответ: $f(x) = \begin{cases} 0, & x \le 2, \\ 2x - 4, & 2 < x < 3, \\ 0, & x \ge 3. \end{cases}$

Задача 4.

Функция распределения непрерывной случайной величины X задана формулами:

$$F(x)=\begin{cases} 0, & x\le 3,\\ \dfrac{x}{3}-1, & 3<x<6,\\ 1, & x\ge 6.\end{cases}$$

Найти вероятность того, что в четырех испытаниях случайная величина трижды примет значение из интервала (4; 5).

Указание

Вероятность того, что в одном испытании X примет значение из интервала (4; 5), равна $F(5) - F(4)$. Далее воспользуйтесь формулой Бернулли.

Решение

$$p(4<X<5)=F(5)-F(4)=\frac{5}{3}-1-\left(\frac{4}{3}-1\right)=\frac{1}{3};$$

$$p_4(3)=C_4^3\cdot\left(\frac{1}{3}\right)^3\cdot\frac{2}{3}=\frac{8}{81}.$$

Ответ: $\frac{8}{81}$.

Задача 5.

Плотность вероятности непрерывной случайной величины X задана формулой

$$f(x)=\frac{8C}{e^x+e^{-x}},\quad x\in(-\infty,\infty).$$

Найти значение константы C.

Указание

Воспользуйтесь условием нормировки:

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

Решение

Воспользуемся условием нормировки:

$$8C\int_{-\infty}^{+\infty}\frac{dx}{e^x+e^{-x}}=8C\int_{-\infty}^{+\infty}\frac{e^x dx}{e^{2x}+1}=8C\int_{-\infty}^{+\infty}\frac{de^x}{e^{2x}+1}=$$

$$=8C\,arctg\,e^x\Big|_{-\infty}^{+\infty}=8C\left(\frac{\pi}{2}-0\right)=4\pi C=1\Rightarrow C=\frac{1}{4\pi}.$$

Ответ: $\frac{1}{4\pi}$.

Задача 6.
Плотность вероятности непрерывной случайной величины X задана формулами:

$$f(x)=\begin{cases} 0, & x\le 2,\\ C(6x-8-x^2), & 2<x<4,\\ 0, & x\ge 4. \end{cases}$$

Найти значение функции распределения $F(3)$.

Указание

Найдите значение константы C из условия нормировки, а затем воспользуйтесь формулой

$$F(x)=\int_{-\infty}^{x} f(t)dt.$$

Решение

Найдем значение константы C из условия нормировки:

$$C\int_2^4 (6x-8-x^2)dx = C\left(3x^2-8x-\frac{x^3}{3}\right)\Bigg|_2^4 = 1 \Rightarrow$$

$$\Rightarrow \frac{4}{3}C=1 \Rightarrow C=\frac{3}{4}.$$

$$F(3)=\int_{-\infty}^{3} f(x)dx = \frac{3}{4}\int_2^3 (6x-8-x^2)dx =$$

$$=\frac{3}{4}\left(3x^2-8x-\frac{x^3}{3}\right)\Bigg|_2^3 = \frac{3}{4}\cdot\frac{2}{3}=\frac{1}{2}.$$

Ответ: $\frac{1}{2}$.

1.2.2. *Числовые характеристики дискретных и непрерывных случайных величин*

Закон распределения (функция распределения и ряд распределения или плотность вероятности) полностью описывают поведение случайной величины. Но в ряде задач достаточно знать только некоторые числовые характеристики исследуемой величины (например, ее среднее значение и возможное отклонение от него), чтобы ответить на поставленный вопрос. Рассмотрим основные числовые характеристики дискретных случайных величин.

Математическое ожидание

Определение. **Математическим ожиданием** дискретной случайной величины называется сумма произведений ее возможных значений на соответствующие им вероятности:

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

Если число возможных значений случайной величины бесконечно, то

$$M(X) = \sum_{i=1}^{\infty} x_i p_i,$$

если полученный ряд сходится абсолютно.

Замечание 1. Математическое ожидание называют иногда *взвешенным средним*, так как оно приближенно равно среднему арифметическому наблюдаемых значений случайной величины при большом числе опытов.

Замечание 2. Из определения математического ожидания следует, что его значение не меньше наименьшего возможного значения случайной величины и не больше наибольшего.

Замечание 3. Математическое ожидание дискретной случайной величины есть *неслучайная* (постоянная) величина. В дальнейшем увидим, что это же справедливо и для непрерывных случайных величин.

Пример 1.

Найдем математическое ожидание случайной величины X – числа стандартных деталей среди трех, отобранных из партии в 10 деталей, среди которых 2 бракованных. Составим ряд распределения для X. Из условия задачи следует, что X может принимать значения 1, 2, 3.

$$p(1) = \frac{C_8^1 \cdot C_2^2}{C_{10}^3} = \frac{1}{15}, \quad p(2) = \frac{C_8^2 \cdot C_2^1}{C_{10}^3} = \frac{7}{15}, \quad p(3) = \frac{C_8^3}{C_{10}^3} = \frac{7}{15}.$$

Тогда

$$M(X) = 1 \cdot \frac{1}{15} + 2 \cdot \frac{7}{15} + 3 \cdot \frac{7}{15} = 2,4.$$

Пример 2.

Определим математическое ожидание случайной величины X – числа бросков монеты до первого появления герба. Эта величина может принимать бесконечное число значений (множество возможных значений есть множество натуральных чисел). Ряд ее распределения имеет вид:

X	1	2	…	n	…
p	0,5	$(0,5)^2$	…	$(0,5)^n$	…

Тогда

$$M(X)=\sum_{n=1}^{\infty}\frac{n}{2^n}=\frac{1}{2}+2\cdot\left(\frac{1}{2}\right)^2+3\cdot\left(\frac{1}{2}\right)^3+\ldots+n\cdot\left(\frac{1}{2}\right)^n+\ldots=$$

$$=\sum_{n=1}^{\infty}\frac{1}{2^n}+\frac{1}{2}\sum_{n=1}^{\infty}\frac{1}{2^n}+\frac{1}{2^n}\sum_{n=1}^{\infty}\frac{1}{2^n}+\ldots=$$

$$=1\cdot\left(1+\frac{1}{2}+\frac{1}{4}+\ldots+\frac{1}{2^n}+\ldots\right)=1\cdot 2=2$$

(при вычислении дважды использовалась формула суммы бесконечно убывающей геометрической прогрессии:

$$S=\frac{b_1}{1-q},$$

откуда

$$\frac{1}{2}+\frac{1}{4}+\ldots+\frac{1}{2^n}+\ldots=1,\quad 1+\frac{1}{2}+\frac{1}{4}+\ldots+\frac{1}{2^n}+\ldots=2).$$

Свойства математического ожидания

1) Математическое ожидание постоянной равно самой постоянной:

$$M(C) = C.$$

Доказательство. Если рассматривать C как дискретную случайную величину, принимающую только одно значение C с вероятностью $p = 1$, то $M(C) = C\cdot 1 = C$.

2) Постоянный множитель можно выносит за знак математического ожидания:

$$M(CX) = C\,M(X).$$

Доказательство. Если случайная величина X задана рядом распределения

x_i	x_1	x_2	…	x_n
p_i	p_1	p_2	…	p_n

то ряд распределения для CX имеет вид:

Cx_i	Cx_1	Cx_2	…	Cx_n
p_i	p_1	p_2	…	p_n

Тогда $M(CX) = Cx_1p_1 + Cx_2p_2 + \ldots + Cx_np_n = C(x_1p_1 + x_2p_2 + \ldots + x_np_n) = CM(X)$.

Определение. Две случайные величины называются **независимыми**, если закон распределения одной из них не зависит от того, какие значения приняла другая. В противном случае случайные величины **зависимы**.

Определение. Назовем **произведением независимых случайных величин X и Y** случайную величину XY, возможные значения которой равны произведениям всех возможных значений X на все возможные значения Y, а соответствующие им вероятности равны произведениям вероятностей сомножителей.

3) Математическое ожидание произведения двух независимых случайных величин равно произведению их математических ожиданий:

$$M(XY) = M(X)M(Y).$$

Доказательство. Для упрощения вычислений ограничимся случаем, когда X и Y принимают только по два возможных значения:

x_i	x_1	x_2
p_i	p_1	p_2

y_i	y_1	y_2
g_i	g_1	g_2

Тогда ряд распределения для XY выглядит так:

XY	x_1y_1	x_2y_1	x_1y_2	x_2y_2
p	p_1g_1	$p_2\,g_1$	p_1g_2	p_2g_2

Следовательно, $M(XY) = x_1y_1 \cdot p_1g_1 + x_2y_1 \cdot p_2g_1 + x_1y_2 \cdot p_1g_2 + x_2y_2 \cdot p_2g_2 = y_1g_1(x_1p_1 + x_2p_2) + \quad + y_2g_2(x_1p_1 + x_2p_2) = (y_1g_1 + y_2g_2)\,(x_1p_1 + x_2p_2) = M(X) \cdot M(Y)$.

Замечание 1. Аналогично можно доказать это свойство для большего количества возможных значений сомножителей.

Замечание 2. Свойство 3 справедливо для произведения любого числа независимых случайных величин, что доказывается методом математической индукции.

Определение. Определим **сумму случайных величин X и Y** как случайную величину $X + Y$, возможные значения которой равны суммам каждого возможного значения X с каждым возможным значением Y; вероятности таких сумм равны произведениям вероятностей слагаемых (для зависимых случайных величин – произведениям вероятности одного слагаемого на условную вероятность второго).

4) Математическое ожидание суммы двух случайных величин (зависимых или независимых) равно сумме математических ожиданий слагаемых:

$$M\,(X + Y) = M\,(X) + M\,(Y).$$

Доказательство.

Вновь рассмотрим случайные величины, заданные рядами распределения, приведенными при доказательстве свойства 3. Тогда возможными значениями $X + Y$ являются $x_1 + y_1$, $x_1 + y_2$, $x_2 + y_1$, $x_2 + y_2$. Обозначим их вероятности соответственно как p_{11}, p_{12}, p_{21} и p_{22}. Найдем $M(\ X + Y\ ) = (x_1 + y_1)p_{11} + (x_1 + y_2)p_{12} + (x_2 + y_1)p_{21} + (x_2 + y_2)p_{22} =$

$= x_1(p_{11} + p_{12}) + x_2(p_{21} + p_{22}) + y_1(p_{11} + p_{21}) + y_2(p_{12} + p_{22})$.

Докажем, что $p_{11} + p_{22} = p_1$. Действительно, событие, состоящее в том, что $X + Y$ примет значения $x_1 + y_1$ или $x_1 + y_2$ и вероятность которого равна $p_{11} + p_{22}$, совпадает с событием, заключающемся в том, что $X = x_1$ (его вероятность –

p_1). Аналогично доказывается, что $p_{21} + p_{22} = p_2$, $p_{11} + p_{21} = g_1$, $p_{12} + p_{22} = g_2$. Значит,

$$M(X + Y) = x_1p_1 + x_2p_2 + y_1g_1 + y_2g_2 = M(X) + M(Y).$$

Замечание. Из свойства 4 следует, что математическое ожидание суммы любого числа случайных величин равно сумме математических ожиданий слагаемых.

Пример 3.

Найти математическое ожидание суммы числа очков, выпавших при броске пяти игральных костей.

Найдем математическое ожидание числа очков, выпавших при броске одной кости:

$M(X_1) = (1 + 2 + 3 + 4 + 5 + 6)\cdot\frac{1}{6} = \frac{7}{2}$. Тому же числу равно математическое ожидание числа очков, выпавших на любой кости. Следовательно, по свойству 4

$$M(X) = 5\cdot\frac{1}{6} = \frac{5}{6}.$$

Дисперсия

Для того, чтобы иметь представление о поведении случайной величины, недостаточно знать только ее математическое ожидание. Рассмотрим две случайные величины: X и Y, заданные рядами распределения вида

Y	0	100
p	0,5	0,5

X	49	50	51
p	0,1	0,8	0,1

Найдем $M(X) = 49\cdot0{,}1 + 50\cdot0{,}8 + 51\cdot0{,}1 = 50$, $M(Y) = 0\cdot0{,}5 + 100\cdot0{,}5 = 50$. Как видно, математические ожидания обеих величин равны, но если для X величина $M(X)$ хорошо описывает поведение случайной величины, являясь ее наиболее вероятным возможным значением (причем остальные значения ненамного отличаются от 50), то значения Y существенно отстоят от $M(Y)$. Следовательно, наряду с математическим ожиданием желательно знать, насколько значения случайной величины отклоняются от него. Для характеристики этого показателя служит дисперсия.

Дисперсией (рассеянием) случайной величины называется математическое ожидание квадрата ее отклонения от ее математического ожидания:

$$\boxed{D(X) = M(X - M(X))^2. \qquad (2)}$$

Пример 4.

Найдем дисперсию случайной величины X (числа стандартных деталей среди отобранных) в примере 1 данной лекции. Вычислим значения квадрата отклонения каждого возможного значения от математического ожидания:

$$(1 - 2{,}4)^2 = 1{,}96;\ (2 - 2{,}4)^2 = 0{,}16;\ (3 - 2{,}4)^2 = 0{,}36.$$

Следовательно,

$$D(X) = 1{,}96 \cdot \frac{1}{15} + 0{,}16 \cdot \frac{7}{15} + 0{,}36 \cdot \frac{7}{15} = \frac{28}{75} \approx 0{,}373.$$

Замечание 1. В определении дисперсии оценивается не само отклонение от среднего, а его квадрат. Это сделано для того, чтобы отклонения разных знаков не компенсировали друг друга.

Замечание 2. Из определения дисперсии следует, что эта величина принимает только неотрицательные значения.

Замечание 3. Существует более удобная для расчетов формула для вычисления дисперсии, справедливость которой доказывается в следующей теореме:

Теорема 1.

$$D(X) = M(X^2) - M^2(X). \qquad (3)$$

Доказательство.

Используя то, что $M(X)$ – постоянная величина, и свойства математического ожидания, преобразуем формулу (7.6) 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.

Пример 5.

Вычислим дисперсии случайных величин X и Y, рассмотренных в начале этого раздела.

$$M(X) = (49^2 \cdot 0{,}1 + 50^2 \cdot 0{,}8 + 51^2 \cdot 0{,}1) - 50^2 = 2500{,}2 - 2500 = 0{,}2.$$

$$M(Y) = (0^2 \cdot 0{,}5 + 100^2 \cdot 0{,}5) - 50^2 = 5000 - 2500 = 2500.$$

Итак, дисперсия второй случайной величины в несколько тысяч раз больше дисперсии первой. Таким образом, даже не зная законов распределения этих величин, по известным значениям дисперсии мы можем утверждать, что X мало отклоняется от своего математического ожидания, в то время как для Y это отклонение весьма существенно.

Свойства дисперсии

1) Дисперсия постоянной величины C равна нулю:

$$D\,(C) = 0.$$

Доказательство.

$$D(C) = M((C - M(C))^2) = M((C - C)^2) = M(0) = 0.$$

2) Постоянный множитель можно выносить за знак дисперсии, возведя его в квадрат:

$$D(CX) = C^2D(X).$$

Доказательство.

$D(CX) = M((CX - M(CX))^2) = M((CX - CM(X))^2) = M(C^2(X - M(X))^2) = C^2D(X).$

3) Дисперсия суммы двух независимых случайных величин равна сумме их дисперсий:

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

Доказательство.

$D(X + Y) = M(X^2 + 2XY + Y^2) - (M(X) + M(Y))^2 = M(X^2) + 2M(X)M(Y) +$
$+ M(Y^2) - M^2(X) - 2M(X)M(Y) - M^2(Y) = (M(X^2) - M^2(X)) + (M(Y^2) - M^2(Y)) = D(X) + D(Y).$

Следствие 1. Дисперсия суммы нескольких взаимно независимых случайных величин равна сумме их дисперсий.

Следствие 2. Дисперсия суммы постоянной и случайной величин равна дисперсии случайной величины.

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

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

Доказательство.

$$D(X - Y) = D(X) + D(-Y) = D(X) + (-1)^2D(Y) = D(X) + D(X).$$

Дисперсия дает среднее значение квадрата отклонения случайной величины от среднего; для оценки самого отклонения служит величина, называемая средним квадратическим (или квадратичным) отклонением.

Определение 7.6. **Средним квадратическим отклонением** σ случайной величины X называется квадратный корень из дисперсии:

$$\sigma = \sqrt{D(X)}. \qquad (4)$$

Пример 6.

В предыдущем примере средние квадратические отклонения X и Y равны соответственно

$$\sigma_x = \sqrt{0,2} \approx 0,447; \quad \sigma_y = \sqrt{2500} = 50.$$

Числовые характеристики непрерывных случайных величин

Распространим определения числовых характеристик случайных величин на непрерывные случайные величины, для которых плотность распределения служит в некотором роде аналогом понятия вероятности.

Математическим ожиданием непрерывной случайной величины называется

$$M(X)=\int_{-\infty}^{+\infty} xf(x)dx. \qquad (5)$$

Общее определение дисперсии сохраняется для непрерывной случайной величины таким же, как и для дискретной, а формула для ее вычисления имеет вид:

$$D(X)=\int_{-\infty}^{+\infty} x^2 f(x)dx - M^2(X). \qquad (6)$$

Среднее квадратическое отклонение вычисляется по формуле (5).
Если все возможные значения непрерывной случайной величины не выходят за пределы интервала $[a, b]$, то интегралы в формулах (5) и (6) вычисляются в этих пределах.

Пример 7.
Плотность распределения случайной величины X имеет вид:

$$f(x)=\begin{cases} 0, & x<2, \\ -\frac{3}{4}(x^2-6x+8), & 2\le x\le 4, \\ 0, & x>4. \end{cases}$$

Найти $M(X)$, $D(X)$, σ.
Решение.

$$M(X)=-\frac{3}{4}\int_2^4 x(x^2-6x+8)dx=-\frac{3}{4}\left(\frac{x^4}{4}-2x^3+4x^2\right)\Bigg|_2^4=3;$$

$$D(X)=-\frac{3}{4}\int_2^4 x^2(x^2-6x+8)dx-9=$$

$$=-\frac{3}{4}\left(\frac{x^5}{5}-\frac{3x^4}{2}+\frac{8x^3}{3}\right)\Bigg|_2^4-9=9,2-9=0,2;$$

$$\sigma=\sqrt{0,2}\approx 0,447.$$

Начальные и центральные моменты

Начальным моментом порядка *k* случайной величины X называется математическое ожидание величины X^k:

$$\nu_k = M(X^k).$$

В частности, $\nu_1 = M(X)$, $\nu_2 = M(X^2)$. Следовательно, дисперсия $D(X) = \nu_2 - \nu_1^2$.

Центральным моментом порядка *k* случайной величины X называется математическое ожидание величины $(X - M(X))^k$:

$$\mu_k = M((X - M(X))^k).$$

В частности, $\mu_1 = M(X - M(X)) = 0$, $\mu_2 = M((X - M(X))^2) = D(X)$.

Можно получить соотношения, связывающие начальные и центральные моменты:

$$\mu_2 = \nu_2 - \nu_1^2, \quad \mu_3 = \nu_3 - 3\nu_2\nu_1 + 2\nu_1^2, \quad \mu_4 = \nu_4 - 4\nu_3\nu_1 + 6\nu_2\nu_1^2 - 3\nu_1^4.$$

Мода и медиана

Такая характеристика случайной величины, как математическое ожидание, называется иногда *характеристикой положения*, так как она дает представление о положении случайной величины на числовой оси. Другими характеристиками положения являются мода и медиана.

Модой *M* дискретной случайной величины называется ее наиболее вероятное значение, **модой *M*** непрерывной случайной величины – значение, в котором плотность вероятности максимальна.

Пример 8.

Если ряд распределения дискретной случайной величины X имеет вид:

X	1	2	3	4
p	0,1	0,7	0,15	0,05

то $M = 2$.

Пример 9.

Для непрерывной случайной величины, заданной плотностью распределения

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

модой является абсцисса точки максимума: $M = 0$.

Замечание 1. Если кривая распределения имеет больше одного максимума, распределение называется **полимодальным**, если эта кривая не имеет максимума, но имеет минимум – **анти-модальным**.

Замечание 2. В общем случае мода и математическое ожидание не совпадают. Но, если распределение является симметричным и модальным (то есть кривая распределения симметрична относительно прямой $x = M$) и имеет математическое ожидание, оно совпадает с модой.

Медианой *Me* непрерывной случайной величины называют такое ее значение, для которого

$$p(X < Me) = p(X > Me).$$

Графически прямая $x = Me$ делит площадь фигуры, ограниченной кривой распределения, на две равные части.
Замечание. Для симметричного модального распределения медиана совпадает с математическим ожиданием и модой.

Для случайной величины X с функцией распределения $F(X)$ **квантилью порядка *p*** $(0 < p < 1)$ называется число K_p такое, что $F(K_p) \le p$, $F(K_p + 0) \ge p$. В частности, если $F(X)$ строго монотонна, K_p: $F(K_p) = p$.

Асимметрия и эксцесс

Если распределение не является симметричным, можно оценить асимметрию кривой распределения с помощью центрального момента 3-го порядка. Действительно, для симметричного распределения все нечетные центральные моменты равны 0 (как интегралы от нечетных функций в симметричных пределах), поэтому выбран нечетный момент наименьшего порядка, не тождественно равный 0. Чтобы получить безразмерную характеристику, его делят на σ^3 (так как μ_3 имеет размерность куба случайной величины).

Определение. **Коэффициентом асимметрии** случайной величины называется

$$S_k = \frac{\mu_3}{\sigma^3}.$$

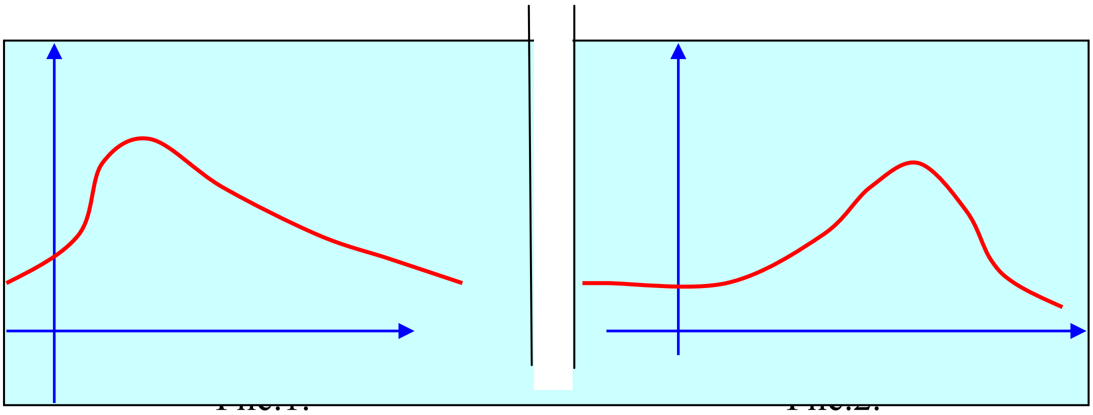


Рис.1. Рис.2.

В частности, для кривой, изображенной на рис.1, $S_k > 0$, а на рис.2 - $S_k < 0$.

Для оценки поведения кривой распределения вблизи точки максимума (для определения того, насколько «крутой» будет его вершина) применяется центральный момент 4-го порядка.

Определение. **Эксцессом** случайной величины называется величина

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

ПРИМЕРЫ РЕШЕНИЯ ЗАДАЧ

Задача 1.
Волейбольный матч заканчивается, когда одна из команд выиграет три партии. Пусть случайная величина X – число партий, сыгранных в матче, если вероятность выигрыша каждой партии одной из команд равна 0,7. Найти математическое ожидание $M(X)$ (ответ округлить до первого знака после запятой).

Указание

Составьте ряд распределения X. Возможные значения X: 3,4,5, причем пятая партия играется только в том случае, если счет в матче после четырех партий 2:2.

Решение

Составим ряд распределения X. Ее наименьшее возможное значение равно 3 (при этом одна из команд выигрывает три партии подряд). Если $X = 4$, одна из команд одержала две победы в первых трех партиях, а затем эта же команда выиграла четвертую партию. Если же $X = 5$, то счет в матче после четырех партий должен быть 2:2, а результат последней партии не важен – в любом случае одна из команд набирает последнее, третье очко.

$$p(3)=0,7^3+0,3^3=0,37;$$
$$p(4)=C_3^2\cdot 0,7^2\cdot 0,3\cdot 0,7+C_3^2\cdot 0,3^2\cdot 0,7\cdot 0,3=0,3654;$$
$$p(5)=C_4^2\cdot 0,7^2\cdot 0,3^2=0,2646.$$

x_i	3	4	5
p_i	0,37	0,3654	0,2646

$$M(X)=3\cdot 0,37+4\cdot 0,3654+5\cdot 0,2646=3,8946\approx 3,9.$$

Ответ: 3,9.

Задача 2.

Найти дисперсию случайной величины X – числа шаров, извлеченных без возвращения из урны, содержащей 3 белых и 4 черных шара, до появления белого шара.

Указание

Составьте ряд распределения X (возможные значения – от 1 до 5) и воспользуйтесь формулой

$$D(X)=M(X^2)-M^2(X).$$

Решение

Составим ряд распределения X. Если $X = k$, то из урны вынуты подряд $k-1$ черных шаров, а затем – белый шар.

$$p(1)=\frac{3}{7};\quad p(2)=\frac{4}{7}\cdot\frac{3}{6}=\frac{2}{7};\quad p(3)=\frac{4}{7}\cdot\frac{3}{6}\cdot\frac{3}{5}=\frac{6}{35};$$
$$p(4)=\frac{4}{7}\cdot\frac{3}{6}\cdot\frac{2}{5}\cdot\frac{3}{4}=\frac{3}{35};\quad p(5)=\frac{4}{7}\cdot\frac{3}{6}\cdot\frac{2}{5}\cdot\frac{1}{4}\cdot 1=\frac{1}{35}.$$

x_i	1	2	3	4	5
p_i	$\frac{3}{7}$	$\frac{2}{7}$	$\frac{6}{35}$	$\frac{3}{35}$	$\frac{1}{35}$

$$M(X)=1\cdot\frac{3}{7}+2\cdot\frac{2}{7}+3\cdot\frac{6}{35}+4\cdot\frac{3}{35}+5\cdot\frac{1}{35}=2;$$

$$M(X^2)=1\cdot\frac{3}{7}+4\cdot\frac{2}{7}+9\cdot\frac{6}{35}+16\cdot\frac{3}{35}+25\cdot\frac{1}{35}=\frac{26}{5};$$

$$D(X)=\frac{26}{5}-2^2=\frac{6}{5}=1,2.$$

Ответ: 1,2.

Задача 3.
Функция распределения непрерывной случайной величины X имеет вид

$$F(x)=\begin{cases}0, & x\le-\pi,\\ \cos\frac{x}{2}, & -\pi<x<0,\\ 1, & x\ge 0.\end{cases}$$

Найти $M(X)$.

Указание

Найдите плотность вероятности и воспользуйтесь формулой

$$M(X)=\int_{-\infty}^{+\infty}xf(x)dx.$$

Решение

$$f(x)=F'(x)=\begin{cases}0, & x<-\pi,\\ -\frac{1}{2}\sin\frac{x}{2}, & -\pi<x<0,\\ 0, & x\ge 0.\end{cases}$$

$$M(X)=-\frac{1}{2}\int_{-\pi}^{0}x\sin\frac{x}{2}dx=x\cos\frac{x}{2}\Bigg|_{-\pi}^{0}-\int_{-\pi}^{0}\cos\frac{x}{2}dx=$$

$$=-2\sin\frac{x}{2}\Bigg|_{-\pi}^{0}=-2.$$

Ответ: -2.

Задача 4.
Плотность вероятности непрерывной случайной величины X имеет вид

$$f(x)=\begin{cases}0, & x\le-2,\\ \frac{C}{\sqrt{4-x^2}}, & -2<x<2,\\ 0, & x\ge 2.\end{cases}$$

Найти дисперсию $D(X)$.

Указание

Найдите значение константы C из условия нормировки, а затем воспользуйтесь формулой

$$D(X)=\int_{-\infty}^{+\infty} x^2 f(x)dx - M^2(X).$$

Решение

$$C\int_{-2}^{2}\frac{dx}{\sqrt{4-x^2}} = C\arcsin\frac{x}{2}\Big|_{-2}^{2} = C\pi = 1 \Rightarrow C=\frac{1}{\pi}.$$

$$M(X)=\frac{1}{\pi}\int_{-2}^{2}\frac{xdx}{\sqrt{4-x^2}}=0,$$

так как определенный интеграл от нечетной функции в пределах, симметричных относительно нуля, равен нулю.

$$D(X)=\frac{1}{\pi}\int_{-2}^{2}\frac{x^2dx}{\sqrt{4-x^2}} \overset{x=2\sin t}{=} \frac{1}{\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\frac{4\sin^2 t\cdot 2\cos t dt}{2\cos t}=$$

$$=\frac{2}{\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}(1-\cos 2t)dt=\frac{2}{\pi}\left(t-\frac{1}{2}\sin 2t\right)\Bigg|_{-\frac{\pi}{2}}^{\frac{\pi}{2}}=2.$$

Ответ: 2.

Задача 5.

Плотность вероятности непрерывной случайной величины X имеет вид

$$f(x)=\begin{cases} 0, & x\le 0,\\ 4(x-x^3), & 0<x<1,\\ 0, & x\ge 1.\end{cases}$$

Найти моду.

Указание

Найдите точку максимума функции $f(x)$.

Решение

$$f'(x)=4(1-3x^2)=0\Rightarrow$$

$$\Rightarrow x_1=-\frac{1}{\sqrt{3}}\notin[0;1],\quad x_2=\frac{1}{\sqrt{3}}.$$

Исследование знака производной показывает, что x_2 – точка максимума плотности вероятности, то есть по определению является модой.

Ответ: $\dfrac{1}{\sqrt{3}}$.

Задача 6.
Плотность вероятности непрерывной случайной величины X имеет вид

$$f(x)=\begin{cases} 0, & x\le 0,\\ \dfrac{2x}{9}, & 0<x<3,\\ 0, & x\ge 3.\end{cases}$$

Найти медиану.

Указание

Найдите такое x_0, для которого

$$\int_0^{x_0} f(x)dx=\frac{1}{2}.$$

Решение

Медиана – это значение x_0, для которого прямая $x = x_0$ делит площадь области, ограниченной графиком функции $y = f(x)$, пополам. Следовательно,

$$\int_0^{x_0}\frac{2x}{9}dx=\left.\frac{x^2}{9}\right|_0^{x_0}=\frac{x_0^2}{9}=\frac{1}{2}\Rightarrow x_0=\frac{3}{\sqrt{2}}=\frac{3\sqrt{2}}{2}.$$

Ответ: $\dfrac{3\sqrt{2}}{2}$.

1.2.3. *Стандартные законы распределения*

Биномиальное распределение

Вернемся к схеме независимых испытаний и найдем закон распределения случайной величины X – числа появлений события A в серии из n испытаний. Возможные значения A: 0, 1, …, n. Соответствующие им вероятности можно вычислить по формуле Бернулли:

$$p(X=k)=C_n^k p^k q^{n-k} \qquad (1)$$

(p – вероятность появления A в каждом испытании).

Такой закон распределения называют **биномиальным**, поскольку правую часть равенства (1) можно рассматривать как общий член разложения бинома Ньютона:

$$(p+q)^n = C_n^n p^n + C_n^{n-1} p^{n-1} q + \ldots + C_n^k p^k q^{n-k} + \ldots + C_n^0 q^n.$$

Пример 1.

Составим ряд распределения случайной величины X – числа попаданий при 5 выстрелах, если вероятность попадания при одном выстреле равна 0,8.

$p(X=0) = 1\cdot(0,2)^5 = 0,00032$; $p(X=1) = 5\cdot 0,8\cdot(0,2)^4 = 0,0064$; $p(X=2) = 10\cdot(0,8)^2\cdot(0,2)^3 = 0,0512$; $p(X=3) = 10\cdot(0,8)^3\cdot(0,2)^2 = 0,2048$; $p(X=4) = 5\cdot(0,8)^4\cdot 0,2 = 0,4096$; $p(X=5) = 1\cdot(0,8)^5 = 0,32768$. Таким образом, ряд распределения имеет вид:

x	0	1	2	3	4	5
p	0,00032	0,0064	0,0512	0,2048	0,4096	0,32728

Числовые характеристики

Для случайной величины, распределенной по биномиальному закону, математическое ожидание $M(X)$ можно найти, используя свойство 4 математического ожидания. Пусть X_1 – число появлений A в первом испытании, X_2 – во втором и т.д. При этом каждая из случайных величин X_i задается рядом распределения вида

X_i	0	1
p_i	q	p

Следовательно, $M(X_i) = p$. Тогда

$$M(X) = \sum_{i=1}^{n} M(X_i) = \sum_{i=1}^{n} p = np.$$

Аналогичным образом вычислим дисперсию:

$$D(X_i) = 0^2\cdot q + 1^2\cdot p - p^2 = p - p^2 = p(1-p),$$

откуда по свойству 4 дисперсии

$$D(X) = \sum_{i=1}^{n} D(X_i) = np(1-p) = npq.$$

Распределение Пуассона

Рассмотрим дискретную случайную величину X, принимающую только целые неотрицательные значения $(0, 1, 2,\ldots, m,\ldots)$, последовательность которых не ограничена. Такая случайная величина называется распределенной **по закону Пуассона**, если вероятность того, что она примет значение m, выражается формулой:

$$p(X=m) = \frac{a^m}{m!} e^{-a}, \qquad (2)$$

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

Покажем, что сумма всех вероятностей равна 1:

$$\sum_{m=0}^{\infty} p(X=m) = e^{-a}\sum_{m=0}^{\infty}\frac{a^m}{m!} = e^{-a}\cdot e^{a} = 1$$

(использовано разложение в ряд Тейлора функции e^x).

Рассмотрим типичную задачу, приводящую к распределению Пуассона. Пусть на оси абсцисс случайным образом распределяются точки, причем их распределение удовлетворяет следующим условиям:

1) Вероятность попадания некоторого количества точек на отрезок длины l зависит только от длины отрезка и не зависит от его расположения на оси (то есть точки распределены с одинаковой средней плотностью);
2) Точки распределяются независимо друг от друга (вероятность попадания какого-либо числа точек на данный отрезок не зависит от количества точек, попавший на любой другой отрезок);
3) Практическая невозможность совпадения двух или более точек.

Тогда случайная величина X – число точек, попадающих на отрезок длины l – распределена по закону Пуассона, где a – среднее число точек, приходящееся на отрезок длины l.

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

Числовые характеристики

Если

$$p(X=m) = \frac{a^m}{m!}e^{-a},$$

то

$$M(X) = \sum_{m=1}^{\infty} m\frac{a^m}{m!}e^{-a} = ae^{-a}\sum_{m=1}^{\infty}\frac{a^{m-1}}{(m-1)!} = ae^{-a}e^{a} = a$$

(использовалось разложение в ряд Тейлора функции e^x).

Для определения дисперсии найдем вначале

$$M(X^2) = \sum_{m=1}^{\infty} m^2\frac{a^m}{m!}e^{-a} = a\sum_{m=1}^{\infty} m\frac{a^{m-1}}{(m-1)!}e^{-a} =$$

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

$$= a\left(\sum_{m=1}^{\infty}(m-1)\frac{a^{m-1}}{(m-1)!}e^{-a} + \sum_{m=1}^{\infty}\frac{a^{m-1}}{(m-1)!}e^{-a}\right) = a(a+1).$$

Поэтому $D(X) = a^2 + a - a^2 = a$.

Замечание. Таким образом, обнаружено интересное свойство распределения Пуассона: математическое ожидание равно дисперсии (и равно единственному параметру a, определяющему распределение).

Равномерный закон распределения

Закон распределения непрерывной случайной величины называется **равномерным**, если на интервале, которому принадлежат все возможные значения случайной величины, плотность распределения сохраняет постоянное значение

($f(x)$ = const при $a \le x \le b$, $f(x) = 0$ при $x < a, x > b$).
Найдем значение, которое принимает $f(x)$ при $x \in [a, b]$. Из условия нормировки следует, что

$$\int_a^b f(x)dx = \int_a^b c dx = c(b-a) = 1,$$

откуда

$$f(x) = c = \frac{1}{b-a}.$$

Вероятность попадания равномерно распределенной случайной величины на интервал $[\alpha, \beta]$ $(a \le \alpha < \beta \le b)$ равна при этом

$$\int_\alpha^\beta \frac{1}{b-a} dx = \frac{\beta - \alpha}{b-a}.$$

Вид функции распределения для нормального закона:

$$F(x) = \begin{cases} 0, \ x < a, \\ \dfrac{x-a}{b-a}, \ a \le x \le b, \\ 1, \ x > b. \end{cases} \qquad (3)$$

Пример 2.
Автобусы некоторого маршрута идут с интервалом 5 минут. Найти вероятность того, что пришедшему на остановку пассажиру придется ожидать автобуса не более 2 минут.
Решение.
Время ожидания является случайной величиной, равномерно распределенной в интервале [0, 5]. Тогда $f(x) = \frac{1}{5}$, $p(0 \le x \le 2) = \frac{2}{5} = 0,4.$

Числовые характеристики

Для равномерно распределенной на отрезке $[a,\ b]$ непрерывной случайной величины

$$M(X)=\int_a^b x\frac{1}{b-a}dx=\left.\frac{x^2}{2(b-a)}\right|_a^b=\frac{b^2-a^2}{2(b-a)}=\frac{a+b}{2},$$

то есть математическое ожидание равномерно распределенной случайной величины равно абсциссе середины отрезка $[a,\ b]$.

Дисперсия

$$D(X)=\int_a^b x^2\frac{1}{b-a}dx-\frac{(a+b)^2}{4}=\frac{b^3-a^3}{3(b-a)}-\frac{(a+b)^2}{4}=$$

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

Нормальный закон распределения

Непрерывная случайная величина называется распределенной по **нормальному закону**, если ее плотность распределения имеет вид:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}. \qquad (4)$$

Замечание. Таким образом, нормальное распределение определяется двумя параметрами: a и σ.

График плотности нормального распределения называют **нормальной кривой (кривой Гаусса)**. Выясним, какой вид имеет эта кривая, для чего исследуем функцию (4).

1) Область определения этой функции: $(-\infty,\ +\infty)$.

2) $f(x) > 0$ при любом x (следовательно, весь график расположен выше оси Ox).

3) $\lim\limits_{|x|\to\infty} f(x)=0,$ то есть ось Ox служит горизонтальной асимптотой графика при $x\to\pm\infty$.

4) $f'(x)=-\dfrac{x-a}{\sigma^3\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}=0$ при $x = a$; $f'(x)>0$ при $x > a$, $f'(x)<0$ при $x < a$. Следовательно, $\left(a,\dfrac{1}{\sigma\sqrt{2\pi}}\right)$ - точка максимума.

5) $f(x-a) = f(a-x)$, то есть график симметричен относительно прямой $x = a$.

6) $f''(x) = -\frac{1}{\sigma^3\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}} \left(1 - \frac{(x-a)^2}{\sigma^2}\right) = 0$ при $x = a \pm \sigma$, то есть точки $\left(a \pm \sigma, \frac{1}{\sigma\sqrt{2\pi e}}\right)$ являются точками перегиба.

Примерный вид кривой Гаусса изображен на рис.7.

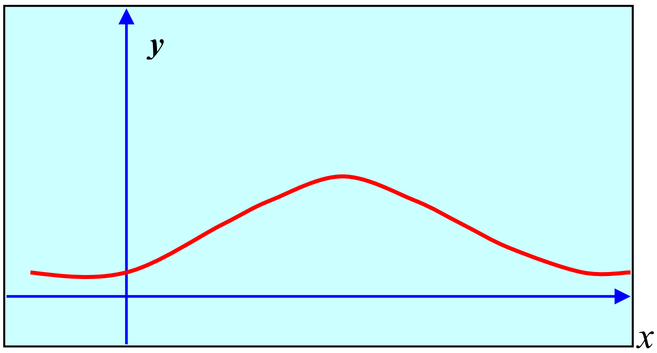


Рис.1.

Найдем вид функции распределения для нормального закона:

$$F(x) = \int_{-\infty}^{x} f(t)dt = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-a)^2}{2\sigma^2}} dt. \quad (5)$$

Перед нами так называемый «неберущийся» интеграл, который невозможно выразить через элементарные функции. Поэтому для вычисления значений $F(x)$ приходится пользоваться таблицами. Они составлены для случая, когда $a = 0$, а $\sigma = 1$.

Нормальное распределение с параметрами $a = 0$, $\sigma = 1$ называется **нормированным**, а его функция распределения

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt \quad - \quad (6)$$

функцией Лапласа.

Замечание. Функцию распределения для произвольных параметров можно выразить через функцию Лапласа, если сделать замену:

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

Найдем вероятность попадания нормально распределенной случайной величины на заданный интервал:

$$p(\alpha \le X < \beta) = F(\beta) - F(\alpha) = \Phi\left(\frac{\beta - a}{\sigma}\right) - \Phi\left(\frac{\alpha - a}{\sigma}\right). \quad (7)$$

Пример 3.

Случайная величина X имеет нормальное распределение с параметрами $a = 3$, σ = 2. Найти вероятность того, что она примет значение из интервала (4, 8).

Решение.

$$p(4 < x < 8) = F(8) - F(4) = \Phi\left(\frac{8-3}{2}\right) - \Phi\left(\frac{4-3}{2}\right) =$$
$$= \Phi(2,5) - \Phi(0,5) = 0,9938 - 0,6915 = 0,3023.$$

Правило «трех сигм»

Найдем вероятность того, что нормально распределенная случайная величина примет значение из интервала (a - 3σ, a + 3σ):

$$p(a - 3\sigma < x < a + 3\sigma) = \Phi(3) - \Phi(-3) = 0,9986 - 0,0014 = 0.9973.$$

Следовательно, вероятность того, что значение случайной величины окажется *вне* этого интервала, равна 0,0027, то есть составляет 0,27% и может считаться пренебрежимо малой. Таким образом, на практике можно считать, что *все* возможные значения нормально распределенной случайной величины лежат в интервале (a - 3σ, a + 3σ).

Полученный результат позволяет сформулировать **правило «трех сигм»**: *если случайная величина распределена нормально, то модуль ее отклонения от x = a не превосходит 3σ (точнее, не превосходит практически никогда).*

Числовые характеристики

Для вычисления математического ожидания нормально распределенной случайной величины воспользуемся тем, что *интеграл Пуассона*

$$\int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}} dz = \sqrt{2\pi}.$$

$$M(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} x e^{-\frac{(x-a)^2}{2\sigma^2}} dx = (z = \frac{x-a}{\sigma}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (\sigma z + a) e^{-\frac{z^2}{2}} dz =$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \sigma z e^{-\frac{z^2}{2}} dz + \frac{a}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}} dz = 0 + \frac{a}{\sqrt{2\pi}} \sqrt{2\pi} = a$$

(первое слагаемое равно 0, так как подынтегральная функция нечетна, а пределы интегрирования симметричны относительно нуля).

$$D(X)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty}(x-a)^2 e^{-\frac{(x-a)^2}{2\sigma^2}}dx=$$

$$=\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} z\cdot ze^{-\frac{z^2}{2}}dz=(u=z, dv=ze^{-\frac{z^2}{2}})=$$

$$=\frac{\sigma^2}{\sqrt{2\pi}}\left(-z\cdot e^{-\frac{z^2}{2}}\Bigg|_{-\infty}^{+\infty}+\int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}}dz\right)=\frac{\sigma^2}{\sqrt{2\pi}}\left(-0+\sqrt{2\pi}\right)=\sigma^2.$$

.

Следовательно, параметры нормального распределения (a и σ) равны соответственно математическому ожиданию и среднему квадратическому отклонению исследуемой случайной величины.

Замечание. Можно показать, что для нормального распределения $\frac{\mu_4}{\sigma^4}=3$, и, соответственно, эксцесс $Ex = 0$. Для кривых с более острой вершиной $Ex > 0$, в случае более плоской вершины $Ex < 0$.

Показательное распределение

Показательным (экспоненциальным) называют распределение вероятностей непрерывной случайной величины X, которое описывается плотностью

$$f(x)=\begin{cases}0, & x<0\\ \lambda e^{-\lambda x}, & x\ge 0.\end{cases} \qquad (8)$$

В отличие от нормального распределения, показательный закон определяется только одним параметром λ. В этом его преимущество, так как обычно параметры распределения заранее не известны и их приходится оценивать приближенно. Понятно, что оценить один параметр проще, чем несколько.

Найдем функцию распределения показательного закона:

$$F(x)=\int_{-\infty}^{x} f(t)dt=\int_{-\infty}^{0} 0\cdot dt+\lambda\int_{0}^{x} e^{-\lambda t}dt=1-e^{-\lambda x}.$$

Следовательно,

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

Теперь можно найти вероятность попадания показательно распределенной случайной величины в интервал (a, b):

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

Значения функции e^{-x} можно найти из таблиц.

Функция надежности

Пусть *элемент* (то есть некоторое устройство) начинает работать в момент времени $t_0 = 0$ и должен проработать в течение периода времени t. Обозначим за T непрерывную случайную величину – время безотказной работы элемента, тогда функция $F(t) = p(T > t)$ определяет вероятность отказа за время t. Следовательно, вероятность безотказной работы за это же время равна

$$R(t) = p(T > t) = 1 - F(t).$$

Эта функция называется **функцией надежности**.

Показательный закон надежности

Часто длительность безотказной работы элемента имеет показательное распределение, то есть

$$F(t) = 1 - e^{-\lambda t}.$$

Следовательно, функция надежности в этом случае имеет вид:

$$R(t) = 1 - F(t) = 1 - (1 - e^{-\lambda t}) = e^{-\lambda t}.$$

Определение. **Показательным законом надежности** называют функцию надежности, определяемую равенством

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

где λ – интенсивность отказов.

Пример 4.

Пусть время безотказной работы элемента распределено по показательному закону с плотностью распределения $f(t) = 0{,}1\, e^{-0{,}1t}$ при $t \ge 0$. Найти вероятность того, что элемент проработает безотказно в течение 10 часов.

Решение.

Так как $\lambda = 0{,}1$, $R(10) = e^{-0{,}1 \cdot 10} = e^{-1} = 0{,}368$.

Числовые характеристики

Для показательного распределения

$$M(X) = \int_0^\infty \lambda x e^{-\lambda x} dx = -x e^{-\lambda x}\Big|_0^\infty + \int_0^\infty e^{-\lambda x} dx =$$

$$= -\frac{1}{\lambda} e^{-\lambda x}\Big|_0^\infty = \frac{1}{\lambda};$$

$$D(X) = \int_0^\infty \lambda x^2 e^{-\lambda x} dx - \frac{1}{\lambda^2} = -x^2 e^{-\lambda x}\Big|_0^\infty + 2\int_0^\infty x e^{-\lambda x} dx - \frac{1}{\lambda^2} =$$

$$= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}; \quad \sigma = \frac{1}{\lambda}.$$

ПРИМЕРЫ РЕШЕНИЯ ЗАДАЧ

Задача 1.
Стрелок делает 6 выстрелов по мишени. Вероятность попадания при каждом выстреле равна 0,7. Для случайной величины X – числа попаданий – найти вероятность того, что $X < 3$.

Указание

Случайная величина X распределена по биномиальному закону. Найдите вероятности того, что X равна 0, 1 и 2.

Решение

Случайная величина X распределена по биномиальному закону, поэтому

$$p(X = k) = C_6^k \cdot 0,7^k \cdot 0,3^{6-k} \Rightarrow$$

$$\Rightarrow p(0) = 0,3^6 = 0,000729;$$

$$p(1) = 6 \cdot 0,7 \cdot 0,3^5 = 0,010206;$$

$$p(2) = C_6^2 \cdot 0,49 \cdot 0,3^4 = 0,059535;$$

$$p(X < 3) = p(0) + p(1) + p(2) = 0,07047.$$

Ответ: 0,07047.

Задача 2.
Стрелок делает 6 выстрелов по мишени. Вероятность попадания при каждом выстреле равна 0,7. Для случайной величины X – числа попаданий – найти дисперсию.

Указание

Случайная величина X распределена по биномиальному закону, поэтому $D(X) = npq$.

Решение

$$D(X) = npq = 6 \cdot 0,7 \cdot 0,3 = 1,26.$$

Ответ: 1,26.

Задача 3.
Вероятность повреждения изделия при перевозке составляет 2%. Случайная величина X – число изделий из партии в 100 штук, поврежденных при перевозке. Найти вероятность того, что $X \leq 2$ (ответ округлить до третьего знака после запятой).

Указание

Можно считать, что X распределена по закону Пуассона, то есть

$$p(X=k)=\frac{\lambda^k e^{-\lambda}}{k!}, \quad \lambda = np$$

(здесь n – число изделий, p – вероятность повреждения одного изделия).

Решение

Воспользуемся формулой Пуассона:

$$p(X=k)=\frac{\lambda^k e^{-\lambda}}{k!}, \quad \lambda = np \Rightarrow$$

$$\Rightarrow \lambda = 2, \quad p(X=k)=\frac{2^k e^{-2}}{k!}.$$

$$p(X=0)=e^{-2}, \quad p(X=1)=2e^{-2}, \quad p(X=2)=2e^{-2},$$

$$p(X\le 2)=p(X=0)+p(X=1)+p(X=2)=5e^{-2}\approx 0,677.$$

Ответ: 0,677.

Задача 4.

Цена деления шкалы измерительного прибора равна 0,2. Показания прибора округляют до ближайшего числа на шкале. Полагая, что ошибка измерения X распределена по равномерному закону, найти дисперсию $D(X)$.

Указание

Воспользуйтесь формулой для вычисления дисперсии равномерно распределенной случайной величины:

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

Решение

$$a=0, \quad b=0,2, \quad D(X)=\frac{(b-a)^2}{12}=\frac{0,04}{12}=\frac{1}{300}.$$

Ответ: $\frac{1}{300}$.

Задача 5.

Измерительный прибор имеет систематическую ошибку 5 м, а случайные ошибки распределены по нормальному закону со средним квадратическим отклонением 75 м. Найти вероятность того, что ошибка измерения не превысит по модулю 5 м.

Указание

Из условия задачи следует, что параметры нормального закона a = 5, σ = 75. Для определения искомой вероятности воспользуйтесь формулой

$$p(\alpha < X < \beta) = \Phi\left(\frac{\beta - a}{\sigma}\right) - \Phi\left(\frac{\alpha - a}{\sigma}\right).$$

Решение

Параметры нормального закона a = 5, σ = 75. Тогда

$$p(-5 < X < 5) = \Phi\left(\frac{5-5}{75}\right) - \Phi\left(\frac{-5-5}{75}\right) =$$
$$= \Phi(0) - \Phi(-0{,}133) = 0 + 0{,}053 = 0{,}053.$$

Ответ: 0,053.

Задача 6.

Какой наименьшей ширины должно быть поле допуска, чтобы с вероятностью не более 0,0027 получалась деталь с контролируемым размером вне поля допуска, если случайные отклонения размера от середины поля допуска распределены по нормальному закону с параметрами a = 0, σ = 5?

Указание

Используйте «правило трех сигм».

Решение

Если вероятность выхода размера детали из поля допуска равна 0,0027, то вероятность того, что размер попадает в поле допуска, равна 1 – 0,0027 = 0,9973. Именно такова вероятность попадания нормально распределенной случайной величины в интервал (a - 3σ, a + 3σ). Следовательно, размеры поля допуска: (0 – 15, 0 + 15) = (-15, 15), а ширина поля допуска – не менее 30.

Ответ: 30.

Задача 7.

Среднее время безотказной работы прибора равно 80 часам. Полагая, что время безотказной работы распределено по показательному закону, найти вероятность того, что в течение 100 часов прибор не выйдет из строя.

Указание

Найдите значение параметра λ из условия, что $M(X) = 1/\lambda$, и воспользуйтесь формулой

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

Решение

$$\lambda = \frac{1}{M(X)} = \frac{1}{80} = 0,0125;$$

$$R(100) = e^{-0,0125\cdot 100} = e^{-1,25} \approx 0,286.$$

Ответ: 0,286.

1.2.4. *Двумерные случайные величины. Коррелированность случайных величин*

Наряду с одномерными случайными величинами, возможные значения которых определяются одним числом, теория вероятностей рассматривает и многомерные случайные величины. Каждое возможное значение такой величины представляет собой упорядоченный набор нескольких чисел. Геометрической иллюстрацией этого понятия служат точки n-мерного пространства, каждая координата которых является случайной величиной (дискретной или непрерывной), или n-мерные векторы. Поэтому многомерные случайные величины называют еще случайными векторами.

Дискретные двумерные случайные величины

Закон распределения дискретной двумерной случайной величины (X,Y) имеет вид таблицы с двойным входом, задающей перечень возможных значений каждой компоненты и вероятности $p(x_i, y_j)$, с которыми величина 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	$p(x_1, y_m)$	$p(x_2, y_m)$	…	$p(x_i, y_m)$	…	$p(x_n, y_m)$

При этом сумма вероятностей, стоящих во всех клетках таблицы, равна 1.

Зная закон распределения двумерной случайной величины, можно найти законы распределения ее составляющих. Действительно, событие $X = x_1$ представляется собой сумму несовместных событий $(X = x_1, Y = y_1)$, $(X = x_1, Y = y_2)$,…, $(X = x_1, Y = y_m)$, поэтому

$$p(X = x_1) = p(x_1, y_1) + p(x_1, y_2) +…+ p(x_1, y_m)$$

(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

Непрерывные двумерные случайные величины

Функцией распределения $F(x, y)$ двумерной случайной величины (X, Y) называется вероятность того, что $X < x$, а $Y < y$:

$$F(x,y) = p(X < x, Y < y). \qquad (1)$$

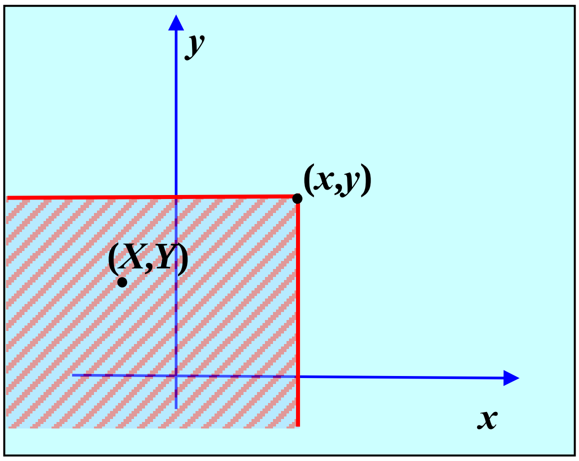


Рис.1

Это означает, что точка (X, Y) попадет в область, заштрихованную на рис. 1, если вершина прямого угла располагается в точке (x, y).

Замечание. Определение функции распределения справедливо как для непрерывной, так и для дискретной двумерной случайной величины.

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:

$$F(x_2, y) \ge F(x_1, y), \text{ если } x_2 > x_1;$$
$$F(x, y_2) \ge F(x, y_1), \text{ если } y_2 > y_1.$$

Доказательство.

$$F(x_2, y) = p(X < x_2, Y < y) = p(X < x_1, Y < y) + p(x_1 \le X < x_2, Y < y) \ge$$
$$\ge p(X < x_1, Y < y) = F(x_1, y).$$

Аналогично доказывается и второе утверждение.

3) Имеют место предельные соотношения:

$$a)\quad F(-\infty, y) = 0;$$
$$b)\quad F(x, +\infty) = 0;$$
$$c)\quad F(-\infty, -\infty) = 0;$$
$$d) F(\infty, \infty) = 1.$$

Доказательство.

События a), b) и c) невозможны (так как невозможно событие

$$X < -\infty \quad или \quad Y < -\infty),$$

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.

Плотностью совместного распределения вероятностей (двумерной плотностью вероятности) непрерывной двумерной случайной величины называется смешанная частная производная 2-го порядка от функции распределения:

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}. \qquad (2)$$

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) $F(x,y)=\int\limits_{-\infty}^{y}\int\limits_{-\infty}^{x} f(x,y)dxdy$

(поскольку это вероятность того, что точка попадет на плоскость Оxy, то есть достоверного события).

Вероятность попадания случайной точки в произвольную область

Пусть в плоскости Оxy 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

$$p((X,Y)\subset D)=\iint\limits_{D} f(x,y)dxdy. \qquad (3)$$

Отыскание плотностей вероятности составляющих двумерной случайной величины

Выше было сказано, как найти функцию распределения каждой составляющей, зная двумерную функцию распределения. Тогда по определению плотности распределения

$$f_1(x)=\frac{dF_1(x)}{dx}=\frac{dF(x,\infty)}{dx}=\frac{d\left(\int\limits_{-\infty}^{x}\int\limits_{-\infty}^{\infty} f(x,y)\right)}{dx}=\int\limits_{-\infty}^{\infty} f(x,y)dy. \qquad (4)$$

Аналогично находится

$$f_2(y)=\int\limits_{-\infty}^{\infty} f(x,y)dx. \qquad (4')$$

Условные законы распределения составляющих дискретной двумерной случайной величины

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 $X = x_1$, $X = x_2, \ldots, X = x_n$, а событием A – событие $Y = y_1$. При такой постановке задачи нам требуется найти условные вероятности гипотез при условии, что A произошло. Следовательно,

$$p_{y_1}(x_i) = \frac{p(x_i, y_1)}{p(y_1)}.$$

Таким же образом можно найти вероятности возможных значений X при условии, что Y принимает любое другое свое возможное значение:

$$p_{y_j}(x_i) = \frac{p(x_i, y_j)}{p(y_j)}. \qquad (5)$$

Аналогично находят условные законы распределения составляющей Y:

$$p_{x_i}(y_j) = \frac{p(x_i, y_j)}{p(x_i)}. \qquad (5')$$

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.

$$p(x_1 / y_1) = \frac{0,1}{0,5} = \frac{1}{5} = 0,2; \quad p(x_2 / y_1) = \frac{0,3}{0,5} = \frac{3}{5} = 0,6;$$

$$p(x_3 / y_1) = \frac{0,1}{0,5} = \frac{1}{5} = 0,2.$$

$$p(y_1 / x_2) = \frac{0,3}{0,55} = \frac{6}{11}; \quad p(y_2 / x_2) = \frac{0,25}{0,55} = \frac{5}{11}.$$

Условные законы распределения составляющих непрерывной двумерной случайной величины

Условной плотностью $\varphi_y(x)$ распределения составляющих X при данном значении $Y = y$ называется

$$\varphi_y(x) = \frac{f(x,y)}{f_2(y)} = \frac{f(x,y)}{\int\limits_{-\infty}^{\infty} f(x,y)dx}. \qquad (6)$$

Аналогично определяется условная плотность вероятности Y при $X = x$:

$$\psi_x(y) = \frac{f(x,y)}{f_1(x)} = \frac{f(x,y)}{\int\limits_{-\infty}^{\infty} f(x,y)dy}. \qquad (6')$$

Равномерное распределение на плоскости

Система двух случайных величин называется **равномерно распределенной на плоскости**, если ее плотность вероятности $f(x, y)$ = const внутри некоторой области и равна 0 вне ее. Пусть данная область – прямоугольник вида

$$a \le x \le b, \quad c \le y \le d.$$

Тогда из свойств $f(x, y)$ следует, что

$$f(x,y)=\begin{cases} \dfrac{1}{S_{пр}}=\dfrac{1}{(b-a)(d-c)} & \textit{внутри прямоугольника,} \\ 0 & \textit{вне его.} \end{cases}$$

Найдем двумерную функцию распределения:

$$F(x,y)=\frac{1}{(b-a)(d-c)}\int_c^y\int_a^x dxdy=\frac{(x-a)(y-c)}{(b-a)(d-c)}$$

при $a < x < b, c < y < d$, $F(x, y) = 0$ 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:

$$F_1(x)=\frac{x-a}{b-a}, \quad F_2(y)=\frac{y-c}{d-c}.$$

Корреляционный момент и коэффициент корреляции. Коррелированность и зависимость случайных величин

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:

$$\alpha_{k,s} = M(X^kY^s).$$

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

$$\mu_{k,s} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x - M(X))^k (y - M(Y))^s f(x,y)dxdy.$$

При этом $M(X) = \alpha_{1,0}$, $M(Y) = \alpha_{0,1}$, $D(X) = \mu_{2,0}$, $D(Y) = \mu_{0,2}$.

Корреляционным моментом системы двух случайных величин называется второй смешанный центральный момент:

$$K_{xy} = \mu_{1,1} = M\big((X - M(X))(Y - M(Y))\big). \quad (7)$$

Для дискретных случайных величин

$$K_{xy} = \sum_i \sum_j (x_i - M(X))(y_j - M(Y))p_{ij},$$

для непрерывных случайных величин

$$K_{xy} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x - M(X))(y - M(Y))f(x,y)dxdy.$$

Безразмерной характеристикой коррелированности двух случайных величин является **коэффициент корреляции**

$$r_{xy} = \frac{K_{xy}}{\sigma_x \sigma_y}. \qquad (8)$$

Корреляционный момент описывает связь между составляющими двумерной случайной величины. Действительно, убедимся, что для независимых X и Y $K_{xy} = 0$. В этом случае $f(x,y) = \ =f_1(x)f_2(y)$, тогда

$$K_{xy} = \int_{-\infty}^{\infty}(x - M(X))f_1(x)dx \int_{-\infty}^{\infty}(y - M(Y))f_2(y)dy = \mu_1(x)\mu_2(y) = 0.$$

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 коэффициент корреляции характеризует не всякую зависимость, а только линейную. В частности, если $Y = aX + b$, то $r_{xy} = \underline{+}1$.

Найдем возможные значения коэффициента корреляции.

Теорема.

$$|r_{xy}| \le 1.$$

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

$$\left|\frac{K_{xy}}{\sigma_x\sigma_y}\right| = \left|r_{xy}\right| \le 0,$$

что и требовалось доказать.

ПРИМЕРЫ РЕШЕНИЯ ЗАДАЧ

Задача 1.
Закон распределения дискретной двумерной случайной величины (X,Y) имеет вид:

y_i \ x_i	1	2	3	4
0	0,1	0,1	0,3	0,1
2	0	0,2	0,1	0,1

Найти вероятность того, что $X + Y = 4$.

Указание

Сумма составляющих равна 4 в двух случаях: $X = 2$, $Y = 2$ или $X = 4$, $Y = 0$.

Решение

$$p(X+Y=4) = p(X=2, Y=2) + p(X=4, Y=0) =$$
$$= 0,2 + 0,1 = 0,3.$$

Ответ: 0,3.

Задача 2.
Закон распределения дискретной двумерной случайной величины (X,Y) имеет вид:

y_i \ x_i	1	2	3	4
0	0,1	0,1	0,3	0,1
2	0	0,2	0,1	0,1

Найти вероятность того, что $X = 3$ при условии, что $Y = 0$.

Указание

Составьте условный закон распределения X при условии, что $Y = 0$.

Решение

Составим условный закон распределения X при условии, что $Y = 0$.

$$p(Y=0)=0,1+0,1+0,3+0,1=0,6.$$

$$p_{Y=0}(X=1)=\frac{0,1}{0,6}=\frac{1}{6}, \quad p_{Y=0}(X=2)=\frac{0,1}{0,6}=\frac{1}{6},$$

$$p_{Y=0}(X=3)=\frac{0,3}{0,6}=\frac{1}{2}, \quad p_{Y=0}(X=4)=\frac{0,1}{0,6}=\frac{1}{6}.$$

Следовательно,

$$p_{Y=0}(X=3)=\frac{0,3}{0,6}=\frac{1}{2}=0,5.$$

Ответ: 0,5.

Задача 3.

Двумерная плотность вероятности системы двух случайных величин (X,Y) имеет вид:

$$f(x,y)=\begin{cases} 0,5\sin(x+y), & 0\le x\le\dfrac{\pi}{2}, 0\le y\le\dfrac{\pi}{2}, \\ 0, & x\notin\left[0,\dfrac{\pi}{2}\right] \quad \textit{или} \quad y\notin\left[0,\dfrac{\pi}{2}\right]. \end{cases}$$

Найти $M(X)$.

Указание

Найдите плотность вероятности X по формуле

$$f_1(x)=\int_{-\infty}^{+\infty} f(x,y)dy,$$

тогда

$$M(X)=\int_{-\infty}^{+\infty} xf_1(x)dx.$$

Решение

$$f_1(x)=\int\limits_{-\infty}^{+\infty} f(x,y)dy=0,5\int\limits_0^{\frac{\pi}{2}}\sin(x+y)dy=$$

$$=-0,5\cos(x+y)\Bigg|_{y=0}^{y=\frac{\pi}{2}}=0,5\cos x-0,5\cos\left(x+\frac{\pi}{2}\right)=$$

$$=0,5(\sin x+\cos x);$$

$$M(X)=0,5\int\limits_0^{\frac{\pi}{2}} x(\sin x+\cos x)dx=$$

$$=0,5x(\sin x-\cos x)\Bigg|_0^{\frac{\pi}{2}}-\int\limits_0^{\frac{\pi}{2}}(\sin x-\cos x)dx=$$

$$=\frac{\pi}{4}+(\sin x+\cos x)\Bigg|_0^{\frac{\pi}{2}}=\frac{\pi}{4}+1-1=\frac{\pi}{4}.$$

Ответ: $\frac{\pi}{4}$.

Задача 4.

Двумерная плотность вероятности системы двух случайных величин (X,Y) имеет вид:

$$f(x,y)=\begin{cases}0,5\sin(x+y), & 0\le x\le\frac{\pi}{2}, 0\le y\le\frac{\pi}{2},\\ 0, & x\notin\left[0,\frac{\pi}{2}\right] \quad \textit{или} \quad y\notin\left[0,\frac{\pi}{2}\right].\end{cases}$$

Найти двумерную функцию распределения в области

$$0\le x\le\frac{\pi}{2}, 0\le y\le\frac{\pi}{2}.$$

Указание

Воспользуйтесь формулой

$$F(x,y)=\int\limits_{-\infty}^{x}\int\limits_{-\infty}^{y} f(x,y)dxdy.$$

Решение

$$F(x,y)=0,5\int_0^x dx\int_0^y \sin(x+y)dy=$$

$$=-0,5\int_0^x dx\left(\cos(x+y)\Big|_0^y\right)=0,5\int_0^x(\cos x-\cos(x+y)dx=$$

$$=0,5(\sin x-\sin(x+y))\Big|_0^y=0,5(\sin x+\sin y-\sin(x+y)).$$

Ответ: $0,5(\sin x+\sin y-\sin(x+y))$.

Задача 5.

Дискретная двумерная случайная величина (X,Y) задана законом распределения

y_i \ x_i	1	3	7
2	0,2	0	0,1
3	0	0,7	0

Найти корреляционный момент K_{xy}.

Указание

Воспользуйтесь формулой

$$K_{xy}=\sum_i\sum_j(x_i-M(X))(y_j-M(Y))p_{ij},$$

вычислив предварительно $M(X)$ и $M(Y)$.

Решение

Найдем законы распределения X и Y:

x_i	1	3	7
p_i	0,2	0,7	0,1

y_i	2	3
p_i	0,3	0,7

$$M(X)=1\cdot 0,2+3\cdot 0,7+7\cdot 0,1=3;$$

$$M(Y)=2\cdot 0,3+3\cdot 0,7=2,7;$$

$$K_{xy}=(1-3)(2-2,7)\cdot 0,2+(3-3)(3-2,7)\cdot 0,7+$$

$$+(7-3)(2-2,7)\cdot 0,1=0.$$

Ответ: 0.

Задача 6.

Плотность вероятности системы случайных величин

$$f(x,y)=\begin{cases} \dfrac{2}{3}\left(x+3y^2\right), & 0\le x\le 1, 0\le y\le 1,\\ 0 \quad \textit{вне} \quad \textit{квадрата} & 0\le x\le 1, 0\le y\le 1. \end{cases}$$

Найти коэффициент корреляции r_{xy} (ответ округлить до третьего знака после запятой).

Указание

Требуется последовательно найти функции распределения составляющих, их математические ожидания и дисперсии и корреляционный момент

$$K_{xy}=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-M(X))(y-M(Y))f(x,y)dxdy.$$

Затем воспользуйтесь формулой

$$r_{xy}=\frac{K_{xy}}{\sigma_x\sigma_y}.$$

Решение

$$f_1(x)=\frac{2}{3}\int_0^1(x+3y^2)dy=\frac{2}{3}(xy+y^3)\Big|_0^1=\frac{2}{3}(x+1);$$

$$f_2(y)=\frac{2}{3}\int_0^1(x+3y^2)dx=\frac{2}{3}\left(\frac{x^2}{2}+3xy^2\right)\Bigg|_0^1=\frac{1}{3}+2y^2;$$

$$M(X)=\frac{2}{3}\int_0^1 x(x+1)dx=\frac{2}{3}\left(\frac{x^3}{3}+\frac{x^2}{2}\right)\Bigg|_0^1=\frac{5}{9};$$

$$M(Y)=\int_0^1 y\left(\frac{1}{3}+2y^2\right)dy=\left(\frac{y^2}{6}+\frac{y^4}{2}\right)\Bigg|_0^1=\frac{2}{3};$$

$$D(X)=\frac{2}{3}\int_0^1 x^2(x+1)dx-\frac{25}{81}=\frac{7}{18}-\frac{25}{81}=\frac{13}{81};$$

$$D(Y)=\int_0^1\left(\frac{y^2}{3}+2y^4\right)dy-\frac{4}{9}=\frac{23}{45}-\frac{4}{9}=\frac{1}{15};$$

$$\sigma_x=\frac{\sqrt{13}}{9};\quad \sigma_y=\frac{1}{\sqrt{15}};$$

$$K_{xy}=\frac{2}{3}\int_0^1\left(x-\frac{5}{9}\right)dx\int_0^1\left(y-\frac{2}{3}\right)\left(x+3y^2\right)dy=$$

$$=\frac{1}{18}\int_0^1\left(x-\frac{5}{9}\right)(7-8x)\,dx=-\frac{5}{108};$$

$$r_{xy}=\frac{K_{xy}}{\sigma_x\sigma_y}=-\frac{5\sqrt{15}\cdot 9}{108\sqrt{13}}\approx -0,448.$$

Ответ: -0,448.

1.2.5. *Функции от случайных величин*

При решении задач часто удобно бывает представить исследуемую случайную величину как функцию других случайных величин с известными законами распределения, что помогает установить и закон распределения заданной случайной величины.

Если каждому возможному значению случайной величины X соответствует одно возможное значение случайной величины Y, то Y называют **функцией случайного аргумента** X**:**

$$Y = \varphi(X).$$

Выясним, как найти закон распределения функции по известному закону распределения аргумента.

1) Пусть аргумент X – дискретная случайная величина, причем различным значениям X соответствуют различные значения Y. Тогда вероятности соответствующих значений X и Y равны.

Пример 1.

Ряд распределения для X имеет вид:

X	5	6	7	8
p	0,1	0,2	0,3	0,4

Найдем закон распределения функции $Y = 2X^2 - 3$:

Y	47	69	95	125
p	0,1	0,2	0,3	0,4

(при вычислении значений Y в формулу, задающую функцию, подставляются возможные значения X).

2) Если разным значениям X могут соответствовать одинаковые значения Y, то вероятности значений аргумента, при которых функция принимает одно и то же значение, складываются.

Пример 2.

Ряд распределения для X имеет вид:

X	0	1	2	3
p	0,1	0,2	0,3	0,4

Найдем закон распределения функции $Y = X^2 - 2X$:

Y	-1	0	3
p	0,2	0,4	0,4

(так как $Y = 0$ при $X = 0$ и $X = 2$, то $p(Y = 0) = p(X = 0) + p(X = 2) = 0{,}1 + 0{,}3 = 0{,}4$).

3) Если X – непрерывная случайная величина, $Y = \varphi(X)$, $\varphi(x)$ – монотонная и дифференцируемая функция, а $\psi(y)$ – функция, обратная к $\varphi(x)$, то плотность распределения $g(y)$ случайной функции Y равна:

$$g(y) = f(\psi(y))\,|\psi'(y)|. \qquad (1)$$

Пример 3.

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad Y = x^3.$$

Тогда

$$\psi(y) = \sqrt[3]{y}, \quad g(y) = \frac{1}{\pi(1+y^{\frac{2}{3}})} \cdot \left(\frac{1}{3} y^{-\frac{2}{3}}\right) = \frac{1}{3\pi y^{\frac{2}{3}}(1+y^{\frac{2}{3}})}.$$

Математическое ожидание функции одного случайного аргумента

Пусть $Y = \varphi(X)$ – функция случайного аргумента X, и требуется найти ее математическое ожидание, зная закон распределения X.

1) Если X – дискретная случайная величина, то

$$M(Y) = M(\varphi(x)) = \sum_{i=1}^{n} \varphi(x_i) p_i. \qquad (2)$$

Пример 4.

Найдем $M(Y)$ для примера 1: $M(Y) = 47\cdot 0{,}1 + 69\cdot 0{,}2 + 95\cdot 0{,}3 + 125\cdot 0{,}4 = 97$.

2) Если X – непрерывная случайная величина, то $M(Y)$ можно искать по-разному. Если известна плотность распределения $g(y)$, то

$$M(Y) = \int_{-\infty}^{\infty} y g(y) dy. \qquad (3)$$

Если же $g(y)$ найти сложно, то можно использовать известную плотность распределения $f(x)$:

$$M(Y) = \int_{-\infty}^{\infty} \varphi(x) f(x) dx. \qquad (4)$$

В частности, если все значения X принадлежат промежутку $(a,\ b)$, то

$$M(Y) = \int_{a}^{b} \varphi(x) f(x) dx. \qquad (4')$$

Функция двух случайных величин. Распределение суммы независимых слагаемых

Если каждой паре возможных значений случайных величин X и Y соответствует одно возможное значение случайной величины Z, то Z называют **функцией двух случайных аргументов** X и Y: $Z = \varphi(X, Y)$.

Рассмотрим в качестве такой функции сумму $X + Y$. В некоторых случаях можно найти ее закон распределения, зная законы распределения слагаемых.

1) Если X и Y – дискретные *независимые* случайные величины, то для определения закона распределения $Z = X + Y$ нужно найти все возможные значения Z и соответствующие им вероятности.

Пример 5.

Рассмотрим дискретные случайные величины X и Y, законы распределения которых имеют вид:

X	-2	1	3
p	0,3	0,4	0,3

Y	0	1	2
p	0,2	0,5	0,3

Найдем возможные значения Z:

-2 + 0 = -2 (p = 0,3·0,2 = 0,06), -2 + 1 = -1 (p = 0,3·0,5 = 0,15), -2 + 2 = 0 (p = 0,3·0,3 = 0,09), 1 + 0 = 1 (p = 0,4·0,2 = 0,08), 1 + 1 = 2 (p = 0,4·0,5 = 0,2), 1 + 2 = 3 (p = 0,4·0,3 = 0,12), 3 + 0 = 3 (p = 0,3·0,2 = 0,06), 3 + 1 = 4 (p = 0,3·0,5 = 0,15), 3 + 2 = 5 (p = 0,3·0,3 = 0,09).

Сложив вероятности повторившегося дважды значения Z = 3, составим ряд распределения для Z:

Z	-2	-1	0	1	2	3	4	5
p	0,06	0,15	0,09	0,08	0,2	0,18	0,15	0,09

2) Если X и Y – непрерывные *независимые* случайные величины, то, если плотность вероятности хотя бы одного из аргументов задана на $(-\infty, \infty)$ одной формулой, то плотность суммы $g(z)$ можно найти по формулам

$$g(z) = \int_{-\infty}^{\infty} f_1(x) f_2(z-x)dx = \int_{-\infty}^{\infty} f_1(z-y) f_2(y)dy, \quad (5)$$

где $f_1(x)$, $f_2(y)$ – плотности распределения слагаемых. Если возможные значения аргументов неотрицательны, то

$$g(z) = \int_{0}^{z} f_1(x) f_2(z-x)dx = \int_{0}^{z} f_1(z-y) f_2(y)dy. \quad (6)$$

Замечание. Плотность распределения суммы двух независимых случайных величин называют **композицией**.

Устойчивость нормального распределения

Закон распределения вероятностей называется **устойчивым**, если композиция таких законов есть тот же закон (возможно, отличающийся другими значениями параметров).

В частности, свойством устойчивости обладает нормальный закон распределения: композиция нормальных законов тоже имеет нормальное распределение, причем ее математическое ожидание и дисперсия равны суммам соответствующих характеристик слагаемых.

ПРИМЕРЫ РЕШЕНИЯ ЗАДАЧ

Задача 1.

Случайная величина X задана рядом распределения

x_i	$\frac{\pi}{6}$	$\frac{\pi}{4}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$	$\frac{2\pi}{3}$	$\frac{3\pi}{4}$
p_i	0,05	0,1	0,15	0,2	0,2	0,3

Найти вероятность того, что случайная величина $Y = \sin^2 2x$ примет значение 0,75.

Указание

Составьте ряд распределения Y и сложите вероятности тех значений X, при которых Y принимает одинаковые значения.

Решение

Найдем значения Y, соответствующие возможным значениям X:

$$X = \frac{\pi}{6} \Rightarrow Y = \sin^2 \frac{\pi}{3} = \left(\frac{\sqrt{3}}{2} \right)^2 = \frac{3}{4} = 0{,}75;$$

$$X = \frac{\pi}{4} \Rightarrow Y = \sin^2 \frac{\pi}{2} = 1^2 = 1;$$

$$X=\frac{\pi}{3}\Rightarrow Y=\sin^2\frac{2\pi}{3}=\left(\frac{\sqrt{3}}{2}\right)^2=\frac{3}{4}=0,75;$$

$$X=\frac{\pi}{2}\Rightarrow Y=\sin^2\pi=0;$$

$$X=\frac{2\pi}{3}\Rightarrow Y=\sin^2\frac{4\pi}{3}=\left(-\frac{\sqrt{3}}{2}\right)^2=\frac{3}{4}=0,75;$$

$$X=\frac{3\pi}{4}\Rightarrow Y=\sin^2\frac{3\pi}{2}=(-1)^2=1.$$

Таким образом, ряд распределения *Y* имеет вид:

y_i	0	0,75	1
p_i	0,2	0,4	0,4

Следовательно, $p(Y=0,75)=0,4$.
Ответ: 0,4.

Задача 2.
Случайная величина *X* – число появлений события *A* в серии из 6 независимых испытаний, в каждом из которых вероятность появления *A* равна 0,7. Найти математическое ожидание случайной величины $Y=e^{2x}$.

Указание

Случайная величина *X* распределена по биномиальному закону. Воспользуйтесь формулой

$$M(Y)=\sum_{i=0}^{6}e^{2x_i}\cdot p_i$$

и формулой бинома Ньютона:

$$(a+b)^n=\sum_{k=0}^{n}C_n^k a^k b^{n-k}.$$

Решение

$$p(X=k)=C_6^k\cdot 0,7^k\cdot 0,3^{6-k}\Rightarrow$$

$$\Rightarrow M(Y)=\sum_{k=0}^{6}e^{2k}\cdot C_6^k\cdot 0,7^k\cdot 0,3^{6-k}=$$

$$=\sum_{k=0}^{6}C_6^k\cdot(0,7e^2)^k\cdot 0,3^{6-k}=\left(0,7e^2+0,3\right)^6.$$

Ответ: $\left(0,7e^2+0,3\right)^6$.

Задача 3.

Случайная величина X равномерно распределена на отрезке [-1;1]. Найти плотность распределения случайной величины Y = -ln(2 + X).

Указание

Задайте плотность вероятности $f(x)$ для равномерного распределения, найдите функцию $\psi(y)$, обратную к Y = -ln(2 + X), а затем воспользуйтесь формулой

$$g(y) = f(\psi(y))\,|\psi'(y)|.$$

Решение

$$f(x) = \begin{cases} \dfrac{1}{2}, & x \in [-1;1], \\ 0, & x \notin [-1;1]. \end{cases}$$

$$y = -\ln(2+x) \Rightarrow x = e^{-y} - 2 = \psi(y).$$

$$x = -1 \Rightarrow y = -\ln(2-1) = 0; \quad x = 1 \Rightarrow y = -\ln 3,$$

то есть

$$x \in [-1;1] \Rightarrow y \in [-\ln 3; 0].$$

Функция $\psi(y)$ монотонно убывает, поэтому

$$\psi'(y) < 0 \Rightarrow |\psi'(y)| = -\psi'(y) = e^{-y};$$

$$g(y) = \begin{cases} \dfrac{1}{2}e^{-y}, & y \in [-\ln 3; 0], \\ 0, & y \notin [-\ln 3; 0]. \end{cases}$$

Ответ: $g(y) = \begin{cases} \dfrac{1}{2}e^{-y}, & y \in [-\ln 3; 0], \\ 0, & y \notin [-\ln 3; 0]. \end{cases}$

Задача 4.

Независимые случайные величины X и Y заданы рядами распределения:

x_i	-1	0	1
p_i	0,2	0,6	0,2

y_i	0	1	2
p_i	0,5	0,3	0,2

Найти закон распределения случайной величины $Z = X + Y$.

Указание

Найдите возможные значения суммы $X + Y$. Если сумма для различных пар слагаемых совпадает, сложите соответствующие вероятности.

Решение

Найдем возможные значения Z и соответствующие им вероятности:

$$-1+0=-1; \quad p=0,2\cdot 0,5=0,1;$$

$$-1+1=0; \quad p=0,2\cdot 0,3=0,06;$$

$$-1+2=1; \quad p=0,2\cdot 0,2=0,04;$$

$$0+0=0; \quad p=0,6\cdot 0,5=0,3;$$

$$0+1=1; \quad p=0,6\cdot 0,3=0,18;$$

$$0+2=2; \quad p=0,6\cdot 0,2=0,12;$$

$$1+0=1; \quad p=0,2\cdot 0,5=0,1;$$

$$1+1=2; \quad p=0,2\cdot 0,3=0,06;$$

$$1+2=3; \quad p=0,2\cdot 0,2=0,04.$$

Итак, Z принимает 5 различных значений:-1,0,1,2 и 3. Сложив вероятности, соответствующие одинаковым значениям Z, получим ряд распределения:

z_i	-1	0	1	2	3
p_i	0,1	0,36	0,32	0,18	0,04

Ответ:

z_i	-1	0	1	2	3
p_i	0,1	0,36	0,32	0,18	0,04

Задача 5.

X и Y – независимые случайные величины, плотности вероятности которых

$$f_1(x)=\begin{cases}\frac{1}{2}e^{-\frac{x}{2}}, & 0\le x<\infty,\\ 0, & x<0;\end{cases}$$

$$f_2(y)=\begin{cases}\frac{1}{7}e^{-\frac{y}{7}}, & 0\le y<\infty,\\ 0, & y<0.\end{cases}$$

Найти плотность вероятности случайной величины $Z = X + Y$.

Указание

Воспользуйтесь формулой

$$g(z)=\int_0^z f_1(x)f_2(z-x)dx.$$

Решение

Поскольку возможные значения аргументов неотрицательны, плотность суммы $g(z)$ можно найти по формуле

$$g(z)=\int_0^z f_1(x)f_2(z-x)dx=\frac{1}{14}\int_0^z e^{-\frac{x}{2}}\cdot e^{-\frac{z-x}{7}}dx=$$

$$=\frac{1}{14}\int_0^z e^{-\frac{5x+z}{14}}dx=-\frac{1}{5}e^{-\frac{5x+z}{14}}\Bigg|_0^z=-\frac{1}{5}\left(e^{-\frac{6z}{14}}-e^{-\frac{z}{14}}\right)=$$

$$=\frac{1}{5}e^{-\frac{z}{14}}\left(1-e^{-\frac{5z}{14}}\right).$$

Ответ: $\frac{1}{5}e^{-\frac{z}{14}}\left(1-e^{-\frac{5z}{14}}\right).$

1.2.6. *Нормальный закон распределения на плоскости. Линейная регрессия. Линейная корреляция. Распределения «хи-квадрат», Стьюдента и Фишера*

Нормальным законом распределения на плоскости называют распределение вероятностей двумерной случайной величины (X, Y), если

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-r_{xy}^2}}e^{-\frac{1}{2(1-r_{xy}^2)}\left(\frac{(x-a_1)^2}{\sigma_x^2}+\frac{(y-a_2)^2}{\sigma_y^2}-2r_{xy}\frac{x-a_1}{\sigma_x}\frac{y-a_2}{\sigma_y}\right)} \qquad (1)$$

Таким образом, нормальный закон на плоскости определяется 5 параметрами: a_1, a_2, σ_x, σ_y, r_{xy}, где a_1, a_2 – математические ожидания, σ_x, σ_y – средние квадратические отклонения, r_{xy} – коэффициент корреляции X и Y.

Предположим, что r_{xy} = 0, то есть X и Y некоррелированы. Тогда из (1) получим:

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y}\cdot e^{-0,5\left(\frac{(x-a_1)^2}{\sigma_x^2}+\frac{(y-a_2)^2}{\sigma_y^2}\right)}=$$

$$=\frac{1}{\sigma_x\sqrt{2\pi}}e^{-\frac{(x-a_1)^2}{2\sigma_x^2}}\cdot\frac{1}{\sigma_y\sqrt{2\pi}}e^{-\frac{(y-a_2)^2}{2\sigma_y^2}}=f_1(x)f_2(y).$$

Следовательно, из некоррелированности составляющих нормально распределенной двумерной случайной величины следует их независимость, то есть для них понятия независимости и некоррелированности равносильны.

Линейная регрессия

Пусть составляющие X и Y двумерной случайной величины (X, Y) зависимы. Будем считать, что одну из них можно приближенно представить как линейную функцию другой, например

$$Y \approx g(X) = \alpha + \beta X, \qquad (2)$$

и определим параметры α и β с помощью метода наименьших квадратов.

Функция $g(X) = \alpha + \beta X$ называется **наилучшим приближением** Y в смысле метода наименьших квадратов, если математическое ожидание $M(Y - g(X))^2$ принимает наименьшее возможное значение; функцию $g(X)$ называют **среднеквадратической регрессией** Y на X.

Теорема 1. Линейная средняя квадратическая регрессия Y на X имеет вид:

$$g(X) = m_y + r\frac{\sigma_y}{\sigma_x}(X - m_x), \qquad (3)$$

где

$$m_x = M(X), \quad m_y = M(Y), \quad \sigma_x = \sqrt{D(X)}, \quad \sigma_y = \sqrt{D(Y)},$$

$$r = \frac{K_{xy}}{\sigma_x \sigma_y} \quad -$$

коэффициент корреляции X и Y.

Доказательство. Рассмотрим функцию

$$F(\alpha, \beta) = M(Y - \alpha - \beta X)^2$$

и преобразуем ее, учитывая соотношения

$M(X - m_x) = M(Y - m_y) = 0$, $M((X - m_x)(Y - m_y)) = K_{xy} = r\sigma_x\sigma_y$:

$$F(\alpha, \beta) = \sigma_y^2 + \beta^2\sigma_x^2 - 2r\sigma_x\sigma_y\beta + (m_y - \alpha - \beta m_x)^2. \quad (4)$$

Найдем стационарные точки полученной функции, решив систему

$$\begin{cases} \dfrac{\partial F}{\partial \alpha} = -2(m_y - \alpha - \beta m_x) = 0, \\ \dfrac{\partial F}{\partial \beta} = 2\beta\sigma_x^2 - 2r\sigma_x\sigma_y = 0. \end{cases}$$

Решением системы будет

$$\beta = r\frac{\sigma_y}{\sigma_x}, \quad \alpha = m_y - r\frac{\sigma_y}{\sigma_x}m_x.$$

Можно проверить, что при этих значениях функция $F(\alpha, \beta)$ имеет минимум, что доказывает утверждение теоремы.

Коэффициент $\beta = r\frac{\sigma_y}{\sigma_x}$ называется **коэффициентом регрессии *Y* на *X***, а прямая

$$y - m_y = r\frac{\sigma_y}{\sigma_x}(x - m_x) \quad - \qquad (5)$$

прямой среднеквадратической регрессии *Y* на *X*.

Подставив координаты стационарной точки в равенство (4), можно найти минимальное значение функции $F(\alpha, \beta)$, равное $\sigma_y^2(1-r^2)$. Эта величина называется **остаточной дисперсией** *Y* относительно *X* и характеризует величину ошибки, допускаемой при замене *Y* на $g(X) = \alpha+\beta X$. При $r = \pm 1$ остаточная дисперсия равна 0, то есть равенство (2) является не приближенным, а точным. Следовательно, при $r = \pm 1$ *Y* и *X* связаны *линейной* функциональной зависимостью. Аналогично можно получить прямую среднеквадратической регрессии *X* на *Y*:

$$x - m_x = r\frac{\sigma_x}{\sigma_y}(y - m_y) \qquad (6)$$

и остаточную дисперсию *X* относительно *Y*. При $r = \pm 1$ обе прямые регрессии совпадают. Решив систему из уравнений (5) и (6), можно найти точку пересечения прямых регрессии – точку с координатами (m_x, m_y), называемую **центром совместного распределения величин *X* и *Y*.**

Линейная корреляция

Для двумерной случайной величины (X, Y) можно ввести так называемое **условное математическое ожидание *Y*** при $X = x$. Для дискретной случайной величины оно определяется как

$$M(Y \mid X = x) = \sum_{j=1}^{m} y_j p(y_j / x),$$

для непрерывной случайной величины –

$$M(Y \mid X = x) = \int_{-\infty}^{\infty} y\psi(y / x)dy.$$

Функцией регрессии *Y* на *X* называется условное математическое ожидание

$$M(Y / x) = f(x).$$

Аналогично определяется условное математическое ожидание *X* и функция регрессии *X* на *Y*.

Если обе функции регрессии *X* на *Y* и *Y* на *X* линейны, то говорят, что *X* и *Y* связаны **линейной корреляционной зависимостью**.

При этом графики линейных функций регрессии являются прямыми линиями, причем можно доказать, что эти линии совпадают с прямыми среднеквадратической регрессии.

Теорема 2. Если двумерная случайная величина (*X, Y*) распределена нормально, то *X* и *Y* связаны линейной корреляционной зависимостью.
Доказательство. Найдем условный закон распределения *Y* при $X = x$

$$\left(\psi(y/x) = \frac{f(x,y)}{f_1(x)}\right),$$

используя формулу двумерной плотности вероятности нормального распределения (1) и формулу плотности вероятности *X*:

$$f_1(x) = \frac{1}{\sigma_x\sqrt{2\pi}} e - \frac{(x-a_1)^2}{2\sigma_x^2}.$$

Сделаем замену

$$u = \frac{x-a_1}{\sigma_x},\ v = \frac{y-a_2}{\sigma_y}.$$

Тогда

$$\psi(y/x) = \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{(v-ru)^2}{2(1-r^2)}} = \frac{1}{\sigma_y\sqrt{1-r^2}\sqrt{2\pi}} e^{-\frac{\left(y-\left(a_2+r\frac{\sigma_y}{\sigma_x}(x-a_1)\right)\right)^2}{2\sigma_y^2(1-r^2)}}.$$

Полученное распределение является нормальным, а его математическое ожидание

$$M(Y/x) = a_2 + r\frac{\sigma_y}{\sigma_x}(x-a_1)$$

есть функция регрессии *Y* на *X*. Аналогично можно получить функцию регрессии *X* на *Y*:

$$M(X/y) = a_1 + r\frac{\sigma_x}{\sigma_y}(y-a_2).$$

Обе функции регрессии линейны, поэтому корреляция между *X* и *Y* линейна, что и требовалось доказать. При этом уравнения прямых регрессии имеют вид

$$y - a_2 = r\frac{\sigma_y}{\sigma_x}(x-a_1), \quad x - a_1 = r\frac{\sigma_x}{\sigma_y}(y-a_2),$$

то есть совпадают с уравнениями прямых среднеквадратической регрессии (см. формулы (5), (6)).

Распределения «хи-квадрат», Стьюдента и Фишера. Связь этих распределений с нормальным распределением

Рассмотрим некоторые распределения, связанные с нормальным и широко применяющиеся в математической статистике.

Распределение «хи-квадрат»

Пусть имеется несколько нормированных нормально распределенных случайных величин: $X_1, X_2, \ldots, X_n$ ($a_i = 0$, $\sigma_i = 1$). Тогда сумма их квадратов

$$\chi^2 = \sum_{i=1}^{n} X_i^2$$

является случайной величиной, распределенной по так называемому **закону «хи-квадрат»** с $k = n$ степенями свободы; если же слагаемые связаны каким-либо соотношением (например, $\sum X_i = n\bar{X}$), то число степеней свободы $k = n - 1$.

Плотность этого распределения

$$f(x) = \begin{cases} 0, & x \le 0; \\ \dfrac{1}{2^{\frac{k}{2}}\Gamma\left(\dfrac{k}{2}\right)} e^{-\frac{x}{2}} x^{\frac{k}{2}-1}, & x > 0. \end{cases}$$

Здесь $\Gamma(x) = \int\limits_0^{\infty} t^{x-1} e^{-t} dt$ – гамма-функция (*относится к числу специальных функций); в частности, $\Gamma(n+1) = n!$.

Следовательно, распределение «хи-квадрат» определяется одним параметром – числом степеней свободы k.

Замечание 1. С увеличением числа степеней свободы распределение «хи-квадрат» постепенно приближается к нормальному.

Замечание 2. С помощью распределения «хи-квадрат» определяются многие другие распределения, встречающиеся на практике, например, распределение случайной величины $\sqrt{\chi^2}$ – длины случайного вектора ($X_1, X_2, \ldots, X_n$), координаты которого независимы и распределены по нормальному закону.

Распределение Стьюдента

Рассмотрим две независимые случайные величины: Z, имеющую нормальное распределение и нормированную (то есть $M(Z) = 0$, $\sigma(Z) = 1$), и V, распределенную по закону «хи-квадрат» с k степенями свободы. Тогда величина

$$T = \frac{Z}{\sqrt{V/k}}$$

имеет распределение, называемое ***t* – распределением или распределением Стьюдента** с k степенями свободы.

С возрастанием числа степеней свободы распределение Стьюдента быстро приближается к нормальному.

Распределение F Фишера – Снедекора

Рассмотрим две независимые случайные величины U и V, распределенные по закону «хи-квадрат» со степенями свободы k_1 и k_2 и образуем из них новую величину

$$F = \frac{U / k_1}{V / k_2}.$$

Ее распределение называют **распределением F Фишера – Снедекора** со степенями свободы k_1 и k_2. Плотность его распределения имеет вид

$$f(x) = \begin{cases} 0, \quad x \le 0; \\ C_0 \dfrac{x^{\frac{k_1-2}{2}}}{(k_2 + k_1 x)^{\frac{k_1+k_2}{2}}}, \quad x > 0, \end{cases}$$

где

$$C_0 = \frac{\Gamma\left(\dfrac{k_1 + k_2}{2}\right) k_1^{\frac{k_1}{2}} k_2^{\frac{k_2}{2}}}{\Gamma\left(\dfrac{k_1}{2}\right)\Gamma\left(\dfrac{k_2}{2}\right)}.$$

Таким образом, распределение Фишера определяется двумя параметрами – числами степеней свободы.

ПРИМЕРЫ РЕШЕНИЯ ЗАДАЧ

Задача 1.

Для двумерной случайной величины

i	1	2	3	4	5	6	7	8	9	10
x_i	1,2	1,5	1,8	2,1	2,3	3,0	3,6	4,2	5,7	6,3
y_i	5,6	6,8	7,8	9,4	10,3	11,4	12,9	14,8	15,2	18,5

составить уравнение прямой линии регрессии Y на X.

Указание

Решение

$$M(X) = \frac{1,2 + 1,5 + \ldots + 6,3}{10} = 3,17; \quad M(Y) = \frac{5,6 + 6,8 + .. + 18,5}{10} = 11,27.$$

$$D(X) = 0,1(1,2^2 + 1,5^2 + \ldots + 6,3^2) - 3,17^2 = 2,7921;$$

$$\sigma_X = \sqrt{2,7921} = 1,671.$$

$$D(Y)=0,1(5,6^2+6,8^2+...+18,5^2)-11,27^2=15,146;$$
$$\sigma_Y=\sqrt{15,146}=3,892.$$
Для определения коэффициента корреляции вычислим предварительно
$$\sum_{i=1}^{10}x_iy_i=1,2\cdot 5,6+1,5\cdot 6,8+...+6,3\cdot 18,5=420,38.$$
Тогда
$$r_{xy}=\frac{420,38-10\cdot 3,17\cdot 11,27}{10\cdot 1,671\cdot 3,892}=0,97.$$
Уравнение прямой линии регрессии Y на X имеет вид:
$$y-11,27=0,97\cdot\frac{3,892}{1,671}(x-3,17),$$
или
$$y=2,26x-4,104.$$
Ответ: $y=2,26x-4,104.$

1.2.7. *Закон больших чисел. Центральная предельная теорема*

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**.

Неравенство Чебышева

Неравенство Чебышева, используемое для доказательства дальнейших теорем, справедливо как для непрерывных, так и для дискретных случайных величин. Докажем его для дискретных случайных величин.

Теорема 1(неравенство Чебышева).

$$p\left(|X-M(X)|<\varepsilon\right)\geq 1-\frac{D(X)}{\varepsilon^2}. \qquad (1)$$

Доказательство.
Пусть X задается рядом распределения

X	x_1	x_2	…	x_n
p	p_1	p_2	…	p_n

Так как события $|X - M(X)| < \varepsilon$ и $|X - M(X)| \geq \varepsilon$ противоположны, то $p\ (|X - M(X)| < \varepsilon) + {} + p\ (|X - M(X)| \geq \varepsilon) = 1$, следовательно, $p\ (|X - M(X)| < \varepsilon) = 1 - p\ (|X - M(X)| \geq \varepsilon)$. Найдем $p\ (|X - M(X)| \geq \varepsilon)$.

$D(X) = (x_1 - M(X))^2 p_1 + (x_2 - M(X))^2 p_2 + \ldots + (x_n - M(X))^2 p_n$. Исключим из этой суммы те слагаемые, для которых $|X - M(X)| < \varepsilon$. При этом сумма может только уменьшиться, так как все входящие в нее слагаемые неотрицательны. Для определенности будем считать, что отброшены первые k слагаемых. Тогда

$D(X) \geq (x_{k+1} - M(X))^2 p_{k+1} + (x_{k+2} - M(X))^2 p_{k+2} + \ldots + (x_n - M(X))^2 p_n \geq \varepsilon^2 (p_{k+1} + p_{k+2} + \ldots + p_n)$.

Отметим, что $p_{k+1} + p_{k+2} + \ldots + p_n$ есть вероятность того, что $|X - M(X)| \geq \varepsilon$, так как это сумма вероятностей всех возможных значений X, для которых это неравенство справедливо. Следовательно, $D(X) \geq \varepsilon^2 p(|X - M(X)| \geq \varepsilon)$, или $p\ (|X - M(X)| \geq \varepsilon) \leq D(X) / \varepsilon^2$. Тогда вероятность противоположного события $p(\ |X - M(X)| < \varepsilon) \geq 1 - D(X) / \varepsilon^2$, что и требовалось доказать.

Теоремы Чебышева и Бернулли

Теорема 2 (теорема Чебышева). Если $X_1, X_2, \ldots, X_n$ – попарно независимые случайные величины, дисперсии которых равномерно ограничены ($D(X_i) \leq C$), то для сколь угодно малого числа ε вероятность неравенства

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

будет сколь угодно близка к 1, если число случайных величин достаточно велико.

Замечание. Иначе говоря, при выполнении этих условий

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

Доказательство. Рассмотрим новую случайную величину

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

и найдем ее математическое ожидание. Используя свойства математического ожидания, получим, что

$$M\left(\frac{X_1 + X_2 + \ldots + X_n}{n} \right) = \frac{M(X_1) + M(X_2) + \ldots + M(X_n)}{n}.$$

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

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

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

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

$$D\left(\frac{X_1+X_2+\ldots+X_n}{n}\right)=$$

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

Используя этот результат, представим предыдущее неравенство в виде:

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

$$\geq 1-\frac{C}{n\varepsilon^2}.$$

Перейдем к пределу при $n\to\infty$:

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

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

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

Теорема доказана.

Следствие.

Если X_1, X_2, …, 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 утрачивает характер случайной 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 Бернулли

Теорема 3 (теорема Бернулли). Если в каждом из n независимых опытов вероятность p появления события A постоянна, то при достаточно большом числе испытаний вероятность того, что модуль отклонения относительной частоты появлений A в n опытах от p будет сколь угодно малым, как угодно близка к 1:

$$\lim_{n\to\infty} p\left(\left|\frac{m}{n}-p\right|<\varepsilon\right)=1. \qquad (2)$$

Доказательство. Введем случайные величины X_1, X_2, …, X_n, где X_i – 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 (так как $D(X_i) = pq$, $p + q = 1$, откуда $pq \le ¼$). Следовательно, к ним можно применить теорему Чебышева при $M_i = p$:

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

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. Из теоремы Бернулли *не следует*, что $\lim\limits_{n\to\infty}\dfrac{m}{n}=p.$ Речь идет лишь о *вероятности* того, что разность относительной частоты и вероятности по модулю может стать сколь угодно малой. Разница заключается в следующем: при обычной сходимости, рассматриваемой в математическом анализе, для всех n, начиная с некоторого значения, неравенство $\left|\dfrac{m}{n}-p\right|<\varepsilon$ выполняется всегда; в нашем случае могут найтись такие значения n, при которых это неравенство неверно. Этот вид сходимости называют **сходимостью по вероятности**.

Центральная предельная теорема Ляпунова. *Предельная теорема Муавра-Лапласа*

Закон больших чисел не исследует вид предельного закона распределения суммы случайных величин. Этот вопрос рассмотрен в группе теорем, называемых **центральной предельной теоремой.** Они утверждают, что закон распределения суммы случайных величин, каждая из которых может иметь различные распределения, приближается к нормальному при достаточно большом числе слагаемых. Этим объясняется важность нормального закона для практических приложений.

Характеристические функции

Для доказательства центральной предельной теоремы используется метод характеристических функций.

Характеристической функцией случайной величины X называется функция

$$g(t) = M(e^{itX}).$$

Таким образом, $g(t)$ представляет собой математическое ожидание некоторой комплексной случайной величины $U = e^{itX}$, связанной с величиной X. В частности, если X – дискретная случайная величина, заданная рядом распределения, то

$$g(t)=\sum_{k=1}^{n}e^{itx_k}p_k.$$

Для непрерывной случайной величины с плотностью распределения $f(x)$

$$g(t)=\int_{-\infty}^{+\infty}e^{itx}f(x)dx. \qquad (1)$$

Пример 1.

Пусть X – число выпадений 6 очков при одном броске игральной кости. Тогда

$$g(t)=e^{it\cdot 0}\cdot\frac{5}{6}+e^{it\cdot 1}\cdot\frac{1}{6}=\frac{5+e^{it}}{6}.$$

Пример 2.
Найдем характеристическую функцию для нормированной непрерывной случайной величины, распределенной по нормальному закону

$$\left(f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\right).$$

При этом

$$g(t)=\int_{-\infty}^{+\infty}e^{itx}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{itx-\frac{x^2}{2}}dx=e^{-\frac{t^2}{2}}$$

(использовалась формула

$$\int_{-\infty}^{+\infty}e^{-Ax^2\pm 2Bx-C}dx=\sqrt{\frac{\pi}{A}}e^{-\frac{AC-B^2}{A}}$$

и то, что $i^2 = -1$).

Свойства характеристических функций

1. Функцию $f(x)$ можно найти по известной функции $g(t)$ по формуле

$$f(x)=\frac{1}{2\pi}\int_{-\infty}^{+\infty}e^{-itx}g(t)dt. \qquad (2)$$

(преобразование (1) называется *преобразованием Фурье*, а преобразование (2) – *обратным преобразованием Фурье*).
2. Если случайные величины X и Y связаны соотношением $Y = aX$, то их характеристические функции связаны соотношением

$$g_y(t) = g_x(at).$$

3. Характеристическая функция суммы независимых случайных величин равна произведению характеристических функций слагаемых: для $Y=\sum_{k=1}^{n}X_k$

$$g_y(t)=g_{x_1}(t)\cdot g_{x_2}(t)\cdot\ldots\cdot g_{x_n}(t).$$

Теорема 4 (центральная предельная теорема для одинаково распределенных слагаемых).

Если $X_1, X_2,\ldots, X_n,\ldots$ - независимые случайные величины с одинаковым законом распределения, математическим ожиданием m и дисперсией σ^2, то при неограниченном увеличении n закон распределения суммы

$$Y_n=\sum_{k=1}^{n}X_k$$

неограниченно приближается к нормальному.

Доказательство.

Докажем теорему для непрерывных случайных величин X_1, $X_2,\ldots$, X_n (доказательство для дискретных величин аналогично). Согласно условию теоремы, характеристические функции слагаемых одинаковы:

$$g_x(t) = \int_{-\infty}^{+\infty} e^{itx} f(x)dx.$$

Тогда по свойству 3 характеристическая функция суммы Y_n буде

$$g_{y_n}(t) = g_x^n(t).$$

Разложим функцию $g_x(t)$ в ряд Маклорена:

$$g_x(t) = g_x(0) + g_x'(0)t + \left(\frac{g_x''(0)}{2} + \alpha(t)\right)t^2,$$

где

$$\alpha(t) \to 0 \quad при \quad t \to 0.$$

Найдем

$$g_x(0) = \int_{-\infty}^{+\infty} f(x)dx = 1,$$

$$g_x'(0) = \int_{-\infty}^{+\infty} ixe^{itx} f(x)dx \Bigg|_{t=0} = i\int_{-\infty}^{+\infty} xe^{itx} f(x)dx \Bigg|_{t=0} =$$

$$= i\int_{-\infty}^{+\infty} xf(x)dx = im.$$

Если предположить, что $m = 0$ (то есть перенести начало отсчета в точку m), то $g_x'(0) = 0$.

$$g_x''(0) = -\int_{-\infty}^{+\infty} x^2 e^{itx} f(x)dx \Bigg|_{t=0} = -\int_{-\infty}^{+\infty} x^2 f(x)dx = -\sigma^2$$

(так как $m = 0$). Подставив полученные результаты в формулу Маклорена, найдем, что

$$g_x(t) = 1 - \left(\frac{\sigma^2}{2} - \alpha(t)\right)t^2.$$

Рассмотрим новую случайную величину

$$Z_n = \frac{Y_n}{\sigma\sqrt{n}},$$

отличающуюся от Y_n тем, что ее дисперсия при любом n равна 0. Так как Y_n и Z_n связаны линейной зависимостью, достаточно доказать, что Z_n распределена по нормальному закону, или, что то же самое, что ее характеристическая функция приближается к характеристической функции нормального закона (см. пример 2). По свойству характеристических функций

$$g_{z_n}(t)=g_{y_n}\left(\frac{t}{\sigma\sqrt{n}}\right)=\left(g_x\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n=$$

$$=\left(1-\left(\frac{\sigma^2}{2}-\alpha\left(\frac{t}{\sigma\sqrt{n}}\right)\right)\frac{t^2}{n\sigma^2}\right)^n.$$

Прологарифмируем полученное выражение:

$$\ln g_{z_n}(t)=n\ln(1-k),$$

где

$$k=\left(\frac{\sigma^2}{2}-\alpha\left(\frac{t}{\sigma\sqrt{n}}\right)\right)\frac{t^2}{n\sigma^2},\quad \lim_{n\to\infty}k=0.$$

Разложим $\ln(1-k)$ в ряд при $n\to\infty$, ограничившись двумя членами разложения, тогда

$$\ln(1-k)\approx -k.$$

Отсюда

$$\lim_{n\to\infty}\ln g_{z_n}(t)=\lim_{n\to\infty}n\cdot(-k)=$$

$$=\lim_{n\to\infty}\left(-\frac{t^2}{2}+\alpha\left(\frac{t}{\sigma\sqrt{n}}\right)\frac{t^2}{\sigma^2}\right)=-\frac{t^2}{2}+\lim_{n\to\infty}\frac{t^2}{\sigma^2}\alpha\left(\frac{t}{\sigma\sqrt{n}}\right),$$

где последний предел равен 0, так как

$$\alpha(t)\to 0\quad \textit{при}\quad t\to 0.$$

Следовательно,

$$\lim_{n\to\infty}\ln g_{z_n}(t)=-\frac{t^2}{2},$$

то есть

$$\lim_{n\to\infty}g_{z_n}(t)=e^{-\frac{t^2}{2}}\quad -$$

характеристическая функция нормального распределения. Итак, при неограниченном увеличении числа слагаемых характеристическая функция величины Z_n неограниченно приближается к характеристической функции нормального закона; следовательно, закон распределения Z_n (и Y_n) неограниченно приближается к нормальному. Теорема доказана.

А.М.Ляпунов доказал центральную предельную теорему для условий более общего вида:

Теорема 5 (теорема Ляпунова). Если случайная величина X представляет собой сумму очень большого числа взаимно независимых случайных величин, для которых выполнено условие:

$$\lim_{n\to\infty}\frac{\sum_{k=1}^{n}b_k}{\left(\sum_{k=1}^{n}D_k\right)^{\frac{3}{2}}},$$

где b_k – третий абсолютный центральный момент величины $X_к$, а D_k – ее дисперсия, то X имеет распределение, близкое к нормальному (условие Ляпунова означает, что влияние каждого слагаемого на сумму ничтожно мало).

Практически можно использовать центральную предельную теорему при достаточно небольшом количестве слагаемых, так как вероятностные расчеты требуют сравнительно малой точности. Опыт показывает, что для суммы даже десяти и менее слагаемых закон их распределения можно заменить нормальным.

Частным случаем центральной предельной теоремы для дискретных случайных величин является теорема Муавра-Лапласа.

Теорема 6 (теорема Муавра-Лапласа). Если производится n независимых опытов, в каждом из которых событие A появляется с вероятностью p, то справедливо соотношение:

$$p\left(\alpha \leq \frac{Y-np}{\sqrt{npq}} < \beta\right) = \Phi(\beta) - \Phi(\alpha), \quad (3)$$

где Y – число появлений события A в n опытах, $q = 1 - p$.

Доказательство.

Будем считать, что

$$Y = \sum_{i=1}^{n} X_i,$$

где X_i – число появлений события A в i-м опыте. Тогда случайную величину

$$Z = \frac{Y - m_y}{\sigma_y}$$

(см. теорему 4) можно считать распределенной по нормальному закону и нормированной, следовательно, вероятность ее попадания в интервал (α, β) можно найти по формуле

$$p(\alpha \leq Z < \beta) = \Phi(\beta) - \Phi(\alpha).$$

Поскольку Y имеет биномиальное распределение,

$$m_y = np, \quad D_y = npq, \quad \sigma_y = \sqrt{npq}..$$

Тогда

$$Z = \frac{Y - np}{\sqrt{npq}}.$$

Подставляя это выражение в предыдущую формулу, получим равенство (3).

Следствие.

В условиях теоремы Муавра-Лапласа вероятность $p_n(k)$ того, что событие A

появится в n опытах ровно k раз, при большом количестве опытов можно найти по формуле:

$$p_n(k) \approx \frac{1}{\sqrt{npq}} \cdot \varphi(x),$$

где

$$x = \frac{k - np}{\sqrt{npq}}, \quad a \quad \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

(значения этой функции приводятся в специальных таблицах).

Пример 3.

Найти вероятность того, что при 100 бросках монеты число выпадений герба окажется в пределах от 40 до 60.

Применим формулу (3), учитывая, что $n = 0,5$. Тогда $np = 100 \cdot 0,5 = 50$,

$$\sqrt{npq} = \sqrt{100 \cdot 0,5 \cdot (1 - 0,5)} = 5.$$

Тогда, если

$$40 < Y < 60, \quad -2 < \frac{Y - 50}{5} < 2.$$

Следовательно,

$$p(40 < Y < 60) = p\left(-2 < \frac{Y - 50}{5} < 2\right) =$$

$$= \Phi(2) - \Phi(-2) = 0,9772 - 0,0228 = 0,9544.$$

Пример 4.

В условиях предыдущего примера найти вероятность того, что выпадет 45 гербов.

Найдем

$$x = \frac{45 - 50}{5} = -1,$$

тогда

$$p_{100}(45) \approx \frac{1}{5} \cdot \varphi(-1) = \frac{1}{5} \cdot \varphi(1) = \frac{1}{5} \cdot 0,2420 = 0,0484.$$

ПРИМЕРЫ РЕШЕНИЯ ЗАДАЧ

Задача 1.

Вероятность того, что акции, переданные на депозит, будут востребованы, равна 0,08. Оценить с помощью неравенства Чебышева вероятность того, что среди 1000 клиентов от 70 до 90 востребуют свои акции.

Указание

Количество клиентов, востребовавших акции, – случайная величина, распределенная по биномиальному закону. Найдите ее математическое ожидание и дисперсию, определите из условия задачи значение ε и примените неравенство Чебышева:

$$p\left(\left|X-M(X)\right|<\varepsilon\right)\geq 1-\frac{D(X)}{\varepsilon^2}.$$

Решение

Количество клиентов, востребовавших акции, – случайная величина, распределенная по биномиальному закону, n = 1000, p = 0,08. Тогда

$$M(X)=np=80,\quad D(X)=npq=73,6,$$

$$p(70\leq X\leq 90)=p\left(-10\leq X-M(X)\leq 10\right)=$$

$$=p\left(\left|X-M(X)\right|\leq 10\right)\Rightarrow \varepsilon=10,\quad \frac{D(X)}{\varepsilon^2}=0,736,$$

$$p\left(\left|X-M(X)\right|\leq 10\right)\geq 1-0,736=0,264.$$

Ответ: $p\geq$ 0,264.

Задача 2.

Дискретная случайная величина X задана законом распределения

x_i	3	5	8
p_i	0,1	0,5	0,4

Используя неравенство Чебышева, оценить вероятность того, что

$$|X-M(X)|<2,5.$$

Указание

Найдите дисперсию заданной случайной величины и примените неравенство Чебышева в виде

$$p\left(\left|X-M(X)\right|<2,5\right)\geq 1-\frac{D(X)}{6,25}.$$

Решение

Найдем дисперсию заданной случайной величины:

$$M(X)=3\cdot 0,1+5\cdot 0,5+8\cdot 0,4=6;$$

$$D(X)=9\cdot 0,1+25\cdot 0,5+64\cdot 0,4-36=3.$$

$$p\left(\left|X-M(X)\right|<2,5\right)\geq 1-\frac{3}{6,25}=0,48.$$

Ответ: $p\geq$ 0,48.

Задача 3.
Используя локальную теорему Лапласа, найти вероятность того, что событие *A* наступит ровно 70 раз в 100 испытаниях, если вероятность его появления в каждом испытании равна 0,64.

Указание

Используйте табличное значение

$$\varphi(x)=\varphi\left(\frac{k-np}{\sqrt{npq}}\right)=\varphi\left(\frac{70-100\cdot 0,64}{\sqrt{100\cdot 0,64\cdot 0,36}}\right)=\varphi(1,25)=0,1826.$$

Решение

Найдем

$$x=\frac{k-np}{\sqrt{npq}}=\frac{70-100\cdot 0,64}{\sqrt{100\cdot 0,64\cdot 0,36}}=1,25;$$

$$p_{100}(70)=\frac{1}{\sqrt{npq}}\varphi(x)=\frac{1}{4,8}\cdot 0,1826=0,038.$$

Ответ: 0,038.

Задача 4.
Используя интегральную теорему Лапласа, найти вероятность того, что событие *A* появится от 50 до 80 раз в 144 испытаниях, если вероятность его появления в каждом испытании равна 0,8.

Указание

Воспользуйтесь табличными значениями функции Лапласа:

$$\Phi(x_1)=\Phi\left(\frac{k_1-np}{\sqrt{npq}}\right)=\Phi\left(\frac{80-144\cdot 0,8}{\sqrt{144\cdot 0,8\cdot 0,2}}\right)=\Phi(-7,33)=-0,5;$$

$$\Phi(x_2)=\Phi\left(\frac{k_2-np}{\sqrt{npq}}\right)=\Phi\left(\frac{120-144\cdot 0,8}{\sqrt{144\cdot 0,8\cdot 0,2}}\right)=\Phi(1)=0,3413.$$

Решение

Вычислим значения функции Лапласа для аргументов, используемых в интегральной теореме Лапласа:

$$\Phi(x_1)=\Phi\left(\frac{k_1-np}{\sqrt{npq}}\right)=\Phi\left(\frac{80-144\cdot 0,8}{\sqrt{144\cdot 0,8\cdot 0,2}}\right)=\Phi(-7,33)=-0,5;$$

$$\Phi(x_2)=\Phi\left(\frac{k_2-np}{\sqrt{npq}}\right)=\Phi\left(\frac{120-144\cdot 0,8}{\sqrt{144\cdot 0,8\cdot 0,2}}\right)=\Phi(1)=0,3413.$$

$$p_{100}(80,120)=\Phi(x_2)-\Phi(x_1)=0,3413+0,5=0,8413.$$

Ответ: 0,8413.

2. МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

2.1. ОБРАБОТКА РЕЗУЛЬТАТОВ ОПЫТА

2.1.1. *Основные понятия математической статистики. Числовые характеристики статистического распределения*

Математическая статистика занимается установлением закономерностей, которым подчинены массовые случайные явления, на основе обработки статистических данных, полученных в результате наблюдений. Двумя основными задачами математической статистики являются:

- определение способов сбора и группировки этих статистических данных;
- разработка методов анализа полученных данных в зависимости от целей исследования, к которым относятся:

а) оценка неизвестной вероятности события; оценка неизвестной функции распределения; оценка параметров распределения, вид которого известен; оценка зависимости от других случайных величин и т.д.;

б) проверка статистических гипотез о виде неизвестного распределения или о значениях параметров известного распределения.

Для решения этих задач необходимо выбрать из большой совокупности однородных объектов ограниченное количество объектов, по результатам изучения которых можно сделать прогноз относительно исследуемого признака этих объектов.

Определим основные понятия математической статистики.

Генеральная совокупность – все множество имеющихся объектов.

Выборка – набор объектов, случайно отобранных из генеральной совокупности.

Объем генеральной совокупности N и объем выборки n – число объектов в рассматриваемой совокупности.

Виды выборки:

Повторная – каждый отобранный объект перед выбором следующего возвращается в генеральную совокупность;

Бесповторная – отобранный объект в генеральную совокупность не возвращается.

Замечание. Для того, чтобы по исследованию выборки можно было сделать выводы о поведении интересующего нас признака генеральной совокупности, нужно, чтобы выборка правильно представляла пропорции генеральной совокупности, то есть была **репрезентативной** (представительной). Учитывая закон больших чисел, можно утверждать, что это условие выполняется, если каждый объект выбран случайно, причем для любого объекта вероятность попасть в выборку одинакова.

Первичная обработка результатов

Пусть интересующая нас случайная величина X принимает в выборке значение x_1 n_1 раз, x_2 – n_2 раз, …, $x_к$ – $n_к$ раз, причем $\sum_{i=1}^{k} n_k = n,$ где n – объем выборки. Тогда наблюдаемые значения случайной величины x_1, x_2,…, $x_к$ называют **вариантами**, а n_1, n_2,…, $n_к$ – **частотами**. Если разделить каждую частоту на объем выборки, то получим **относительные частоты**

$$w_i = \frac{n_i}{n}.$$

Последовательность вариант, записанных в порядке возрастания, называют **вариационным** рядом, а перечень вариант и соответствующих им частот или относительных частот – **статистическим рядом**:

x_i	x_1	x_2	…	x_k
n_i	n_1	n_2	…	n_k
w_i	w_1	w_2	…	w_k

Пример 1.

При проведении 20 серий из 10 бросков игральной кости число выпадений шести очков оказалось равным 1,1,4,0,1,2,1,2,2,0,5,3,3,1,0,2,2,3,4,1.Составим вариационный ряд: 0,1,2,3,4,5. Статистический ряд для абсолютных и относительных частот имеет вид:

x_i	0	1	2	3	4	5
n_i	3	6	5	3	2	1
w_i	0,15	0,3	0,25	0,15	0,1	0,05

Если исследуется некоторый непрерывный признак, то вариационный ряд может состоять из очень большого количества чисел. В этом случае удобнее использовать **группированную выборку**. Для ее получения интервал, в котором заключены все наблюдаемые значения признака, разбивают на несколько равных частичных интервалов длиной h, а затем находят для каждого частичного интервала n_i – сумму частот вариант, попавших в i-й интервал. Составленная по этим результатам таблица называется **группированным статистическим рядом**:

Номера интервалов	1	2	…	k
Границы интервалов	$(a, a + h)$	$(a + h, a + 2h)$	…	$(b – h, b)$
Сумма частот вариант, попавших в интервал	n_1	n_2	…	n_k

Полигон частот. Выборочная функция распределения и гистограмма

Для наглядного представления о поведении исследуемой случайной величины в выборке можно строить различные графики. Один из них – **полигон частот**: ломаная, отрезки которой соединяют точки с координатами (x_1, n_1), (x_2, n_2),…, (x_k, n_k), где x_i откладываются на оси абсцисс, а n_i – на оси ординат. Если на оси ординат откладывать не абсолютные (n_i), а относительные (w_i) частоты, то получим **полигон относительных частот** (рис.1)**.**

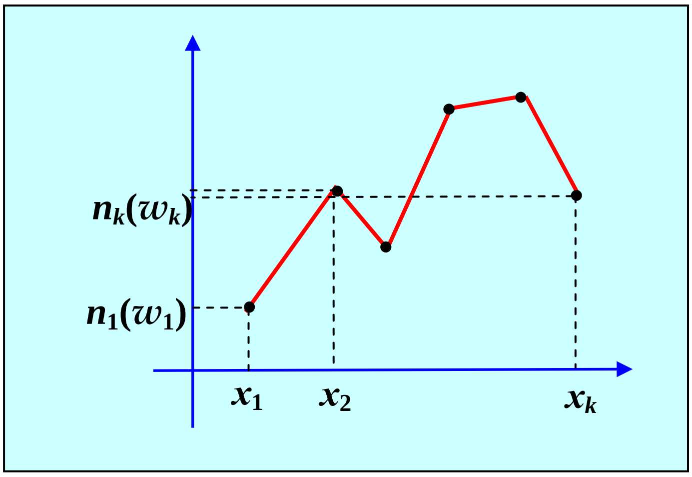


Рис. 1.

По аналогии с функцией распределения случайной величины можно задать некоторую функцию, относительную частоту события $X < x$.

Выборочной (эмпирической) функцией распределения называют функцию $F^*(x)$, определяющую для каждого значения x относительную частоту события $X < x$. Таким образом,

$$F^*(x) = \frac{n_x}{n},$$

где n_x – число вариант, меньших x, n – объем выборки.

Замечание. В отличие от эмпирической функции распределения, найденной опытным путем, функцию распределения $F(x)$ генеральной совокупности называют *теоретической функцией распределения*. $F(x)$ определяет вероятность события $X < x$, а $F^*(x)$ – его относительную частоту. При достаточно больших n, как следует из теоремы Бернулли, $F^*(x)$ стремится по вероятности к $F(x)$.

Из определения эмпирической функции распределения видно, что ее свойства совпадают со свойствами $F(x)$, а именно:

1) $0 \le F^*(x) \le 1$.
2) $F^*(x)$ – неубывающая функция.
3) Если x_1 – наименьшая варианта, то $F^*(x) = 0$ при $x \le x_1$; если $x_к$ – наибольшая варианта, то $F^*(x) = 1$ при $x > x_к$.

Для непрерывного признака графической иллюстрацией служит **гистограмма**, то есть ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат частичные интервалы длиной h, а высотами – отрезки длиной n_i/h (гистограмма частот) или w_i/h (гистограмма относительных частот). В первом случае площадь гистограммы равна объему выборки, во втором – единице (рис.2).

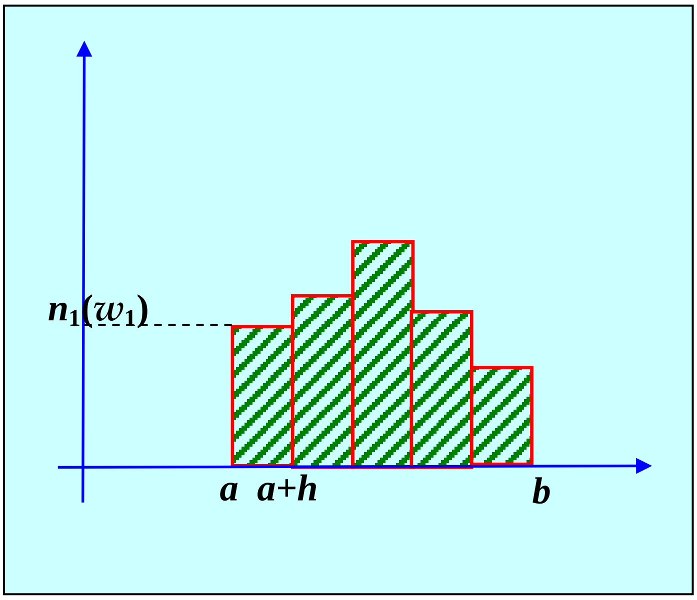


Рис.2.

Числовые характеристики статистического распределения

Одна из задач математической статистики: по имеющейся выборке оценить значения числовых характеристик исследуемой случайной величины.

Выборочным средним называется среднее арифметическое значений случайной величины, принимаемых в выборке:

$$\bar{x}_B = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{n_1x_1 + n_2x_2 + \ldots + n_kx_k}{n} = \frac{\sum_{i=1}^{k} n_i x_i}{n},$$

где x_i – варианты, n_i - частоты.

Замечание. Выборочное среднее служит для оценки математического ожидания исследуемой случайной величины. В дальнейшем будет рассмотрен вопрос, насколько точной является такая оценка.

Выборочной дисперсией называется

$$D_B = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_B)^2}{n} = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x}_B)^2}{n},$$

а **выборочным средним квадратическим отклонением** –

$$\sigma_B = \sqrt{D_B}.$$

Так же, как в теории случайных величин, можно доказать, что справедлива следующая формула для вычисления выборочной дисперсии:

$$D = \overline{x^2} - (\bar{x})^2.$$

Пример 2.

Найдем числовые характеристики выборки, заданной статистическим рядом

x_i	2	5	7	8
n_i	3	8	7	2

$$\bar{x}_B = \frac{2\cdot 3+5\cdot 8+7\cdot 7+8\cdot 2}{20} = 5{,}55;$$

$$D_B = \frac{4\cdot 3+25\cdot 8+49\cdot 7+64\cdot 2}{20} - 5{,}55^2 = 3{,}3475;$$

$$\sigma_B = \sqrt{3{,}3475} = 1{,}83.$$

Другими характеристиками вариационного ряда являются:

- **мода** $\boldsymbol{M_0}$ – варианта, имеющая наибольшую частоту (в предыдущем примере $M_0 = 5$).

- **медиана** $\boldsymbol{m_e}$ - варианта, которая делит вариационный ряд на две части, равные по числу вариант. Если число вариант нечетно ($n = 2k + 1$), то $m_e = x_{k+1}$, а при четном $n = 2k$

$$m_e = \frac{x_k + x_{k+1}}{2}.$$

В частности, в примере 1

$$m_e = \frac{5+7}{2} = 6.$$

Оценки начальных и центральных моментов (так называемые эмпирические моменты) определяются аналогично соответствующим теоретическим моментам:

- **начальным эмпирическим моментом порядка** $\boldsymbol{k}$ называется

$$M_k = \frac{\sum n_i x_i^k}{n}.$$

В частности,

$$M_1 = \frac{\sum n_i x_i}{n} = \bar{x}_B,$$

то есть начальный эмпирический момент первого порядка равен выборочному среднему.

- **центральным эмпирическим моментом порядка** $\boldsymbol{k}$ называется

$$m_k = \frac{\sum n_i (x_i - \bar{x}_B)^k}{n}.$$

В частности,

$$m_2 = \frac{\sum n_i (x_i - \bar{x}_B)^2}{n} = D_B,$$

то есть центральный эмпирический момент второго порядка равен выборочной дисперсии.

Статистическое описание и вычисление характеристик двумерного случайного вектора

При статистическом исследовании двумерных случайных величин основной задачей является обычно выявление связи между составляющими.
Двумерная выборка представляет собой набор значений случайного вектора: (x_1, y_1), (x_2, y_2), …, (x_n, y_n). Для нее можно определить выборочные средние составляющих:

$$\bar{x}_B = \frac{\sum x_i}{n}, \quad \bar{y}_B = \frac{\sum y_i}{n}$$

и соответствующие выборочные дисперсии и средние квадратические отклонения. Кроме того, можно вычислить **условные средние**: $\bar{y}_x$ – среднее арифметическое наблюдавшихся значений Y, соответствующих $X = x$, и $\bar{x}_y$ – среднее значение наблюдавшихся значений X, соответствующих $Y = y$.
Если существует зависимость между составляющими двумерной случайной величины, она может иметь разный вид: функциональная зависимость, если каждому возможному значению X соответствует одно значение Y, и статистическая, при которой изменение одной величины приводит к изменению распределения другой. Если при этом в результате изменения одной величины меняется среднее значение другой, то статистическую зависимость между ними называют корреляционной.

ПРИМЕРЫ РЕШЕНИЯ ЗАДАЧ

Задача 1.
Дана выборка: 1,7,8,3,3,7,3,1,1,3. Найти выборочное среднее.

Указание

Составьте статистический ряд и воспользуйтесь формулой:

$$\bar{x}_B = \frac{\sum_{i=1}^{k} n_i x_i}{n}.$$

Решение

Составим для данной выборки статистический ряд:

x_i	1	3	7	8
n_i	3	4	2	1

Объем выборки, то есть количество входящих в нее чисел, равен 10. Тогда

$$\bar{x}_B = \frac{\sum_{i=1}^{k} n_i x_i}{n} = \frac{1\cdot 3 + 3\cdot 4 + 7\cdot 2 + 8\cdot 1}{10} = 3{,}7.$$

Ответ: 3,7.

Задача 2.
Дана выборка, вариационный ряд которой имеет вид:
10,8; 11,1; 11,7; 12,2; 13,1; 13,4; 13,9; 14,3; 14,3; 14,4; 14,8; 16,5; 17,7; 18,2; 19,9; 20,0; 20,3; 20,8; 23,1; 24,2; 25,1; 25,1; 25,7; 28,4; 28,5; 29,3; 29,8; 29,9; 30,2; 30,4.
Составить статистический ряд распределения абсолютных и относительных частот, состоящий из пяти интервалов, и найти выборочное среднее.

Указание

Объем выборки $n = 30$. Выберите в качестве границ интервала $a = 10{,}5$ и $b = 30{,}5$. Тогда

$$h = \frac{30{,}5 - 10{,}5}{5} = 4,$$

и $(a,\ b)$ разбивается на части (10,5; 14,5), (14,5; 18,5), (18,5; 22,5), (22,5; 26,5) и (26,5; 30,5).

Решение

Объем выборки $n = 30$. Выберем в качестве границ интервала $a = 10{,}5$ и $b = 30{,}5$. Тогда $h = \frac{30{,}5 - 10{,}5}{5} = 4,$ и $(a,\ b)$ разбивается на части (10,5; 14,5), (14,5; 18,5), (18,5; 22,5), (22,5; 26,5) и (26,5; 30,5). Статистический ряд при этом имеет вид:

Номер интервала	Границы интервала	Середина интервала	Абсолютные частоты
1	10,5 14,5	12,5	10
2	14,5 18,5	16,5	4
3	18,5 22,5	20,5	4
4	22,5 26,5	24,5	5
5	26,5 30,5	28,5	7

Найдем выборочное среднее, принимая в качестве вариант середины полученных интервалов:

$$\overline{x}_B = \frac{12{,}5\cdot 10 + 16{,}5\cdot 4 + 20{,}5\cdot 4 + 24{,}5\cdot 5 + 28{,}5\cdot 7}{30} =$$

$$= \frac{119}{6} \approx 19{,}833.$$

Ответ: $\overline{x}_B = 19{,}833.$

Задача 3.
Дана выборка, состоящая из чисел: 3,2; 4,1; 8,1; 8,1; 6,7; 4,4; 4,4; 3,2; 5,0; 6,7; 6,7; 7,5; 3,2; 4,4; 6,7; 6,7; 5,0; 5,0; 4,4; 8,1. Найти выборочную дисперсию.

Указание

Составьте статистический ряд, найдите выборочное среднее и воспользуйтесь формулой

$$D_B = \sum_{i=1}^{k} \frac{(x_i - \overline{x}_B)^2}{n}.$$

Решение

Составим статистический ряд:

x_i	3,2	4,4	5,0	6,7	7,5	8,1
n_i	3	5	3	5	1	3

Тогда

$$\overline{x}_B = \frac{3,2\cdot 3 + 4,4\cdot 5 + 5\cdot 3 + 6,7\cdot 5 + 7,5\cdot 1 + 8,1\cdot 3}{20} = 5,595;$$

$$D_B = \frac{(3,2-5,595)^2\cdot 3 + \ldots + (8-5,595)^2\cdot 3}{20} = 2,698.$$

Ответ: 2,698.

Задача 4.

Дана выборка значений случайной величины: 2, 3, 3, 4, 2, 5, 5, 5, 6, 3, 6, 3, 4, 4, 4, 6, 5, 7, 3, 5. Найти исправленное выборочное среднее квадратическое отклонение.

Указание

Воспользуйтесь формулами для исправленной выборочной дисперсии и исправленного выборочного среднего квадратического отклонения:

$$s^2 = \frac{n}{n-1} D_B, \quad s = \sqrt{s^2}.$$

Решение

Составим статистический ряд:

x_i	2	3	4	5	6	7
n_i	2	5	4	5	3	1

Тогда

$$n = 20, \quad \overline{x}_B = \frac{2\cdot 2 + 3\cdot 5 + 4\cdot 4 + 5\cdot 5 + 6\cdot 3 + 7\cdot 1}{20} = 4,25;$$

$$s^2 = \frac{(2-4,25)^2\cdot 2 + (3-4,25)^2\cdot 5 + \ldots + (7-4,25)^2\cdot 1}{19} = 1,987;$$

$$s = \sqrt{1,987} = 1,41.$$

Ответ: 1,41.

2.1.2. *Основные свойства статистических характеристик параметров распределения. Точечные и интервальные оценки*

Получив статистические оценки параметров распределения (выборочное среднее, выборочную дисперсию и т.д.), нужно убедиться, что они в достаточной степени служат приближением соответствующих характеристик генеральной совокупности. Определим требования, которые должны при этом выполняться.

Пусть Θ* - статистическая оценка неизвестного параметра Θ теоретического распределения. Извлечем из генеральной совокупности несколько выборок одного и того же объема *n* и вычислим для каждой из них оценку параметра Θ: $\Theta_1^*, \Theta_2^*, ..., \Theta_k^*$. Тогда оценку Θ* можно рассматривать как случайную величину, принимающую возможные значения $\Theta_1^*, \Theta_2^*, ..., \Theta_k^*$. Если математическое ожидание Θ* не равно оцениваемому параметру, мы будем получать при вычислении оценок систематические ошибки одного знака (с избытком, если $M(\Theta^*) > \Theta$, и с недостатком, если $M(\Theta^*) < \Theta$). Следовательно, необходимым условием отсутствия систематических ошибок является требование $M(\Theta^*) = \Theta$.

Статистическая оценка Θ* называется **несмещенной**, если ее математическое ожидание равно оцениваемому параметру Θ при любом объеме выборки:

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

Смещенной называют оценку, математическое ожидание которой не равно оцениваемому параметру.

Однако несмещенность не является достаточным условием хорошего приближения к истинному значению оцениваемого параметра. Если при этом возможные значения Θ* могут значительно отклоняться от среднего значения, то есть дисперсия Θ* велика, то значение, найденное по данным одной выборки, может значительно отличаться от оцениваемого параметра. Следовательно, требуется наложить ограничения на дисперсию.

Статистическая оценка называется **эффективной**, если она при заданном объеме выборки *n* имеет наименьшую возможную дисперсию.

При рассмотрении выборок большого объема к статистическим оценкам предъявляется еще и требование состоятельности.

Состоятельной называется статистическая оценка, которая при $n\to\infty$ стремится по вероятности к оцениваемому параметру

(если эта оценка несмещенная, то она будет состоятельной, если при $n\to\infty$ ее дисперсия стремится к 0).

Убедимся, что $\bar{x}_B$ представляет собой несмещенную оценку математического ожидания $M(X)$.

Будем рассматривать $\bar{x}_B$ как случайную величину, а $x_1, x_2,\ldots, x_n$, то есть значения исследуемой случайной величины, составляющие выборку, – как независимые, одинаково распределенные случайные величины $X_1, X_2,\ldots, X_n$, имеющие математическое ожидание a. Из свойств математического ожидания следует, что

$$M(\bar{X}_B)=M\left(\frac{X_1+X_2+\ldots+X_n}{n}\right)=a.$$

Но, поскольку каждая из величин $X_1, X_2,\ldots, X_n$ имеет такое же распределение, что и генеральная совокупность, $a = M(X)$, то есть

$$M\left(\bar{X}_B\right)=M(X),$$

что и требовалось доказать. Выборочное среднее является не только несмещенной, но и состоятельной оценкой математического ожидания. Если предположить, что $X_1, X_2,\ldots, X_n$ имеют ограниченные дисперсии, то из теоремы Чебышева следует, что их среднее арифметическое, то есть $\bar{X}_B$, при увеличении n стремится по вероятности к математическому ожиданию a каждой их величин, то есть к $M(X)$. Следовательно, выборочное среднее есть состоятельная оценка математического ожидания.

В отличие от выборочного среднего, выборочная дисперсия является смещенной оценкой дисперсии генеральной совокупности. Можно доказать, что

$$M(D_B)=\frac{n-1}{n}D_Г,$$

где $D_Г$ – истинное значение дисперсии генеральной совокупности. Можно предложить другую оценку дисперсии – **исправленную дисперсию** s^2, вычисляемую по формуле

$$s^2=\frac{n}{n-1}D_B=\frac{\sum_{i=1}^{k}n_i(x_i-\bar{x}_B)^2}{n-1}.$$

Такая оценка будет являться несмещенной. Ей соответствует **исправленное среднее квадратическое отклонение**

$$s=\sqrt{s^2}=\sqrt{\frac{\sum_{i=1}^{k} n_i(x_i-\overline{x}_B)^2}{n-1}}.$$

Оценка некоторого признака называется **асимптотически несмещенной**, если для выборки $x_1, x_2, \ldots, x_n$

$$\lim_{n\to\infty}\frac{x_1+x_2+\ldots+x_n}{n}=X,$$

где X – истинное значение исследуемой величины.

Способы построения оценок

1. Метод наибольшего правдоподобия

Пусть X – дискретная случайная величина, которая в результате n испытаний приняла значения x_1, x_2, …, x_n. Предположим, что нам известен закон распределения этой величины, определяемый параметром Θ, но неизвестно численное значение этого параметра. Найдем его точечную оценку.

Пусть $p(x_i, \Theta)$ – вероятность того, что в результате испытания величина X примет значение x_i. Назовем **функцией правдоподобия** дискретной случайной величины X функцию аргумента Θ, определяемую по формуле:

$$L\ (x_1, x_2, \ldots, x_n; \Theta) = p(x_1,\Theta)p(x_2,\Theta)\ldots p(x_n,\Theta).$$

Тогда в качестве точечной оценки параметра Θ принимают такое его значение

$$\Theta^* = \Theta(x_1, x_2, \ldots, x_n),$$

при котором функция правдоподобия достигает максимума. Оценку Θ* называют **оценкой наибольшего правдоподобия**.

Поскольку функции L и ln$L$ достигают максимума при одном и том же значении Θ, удобнее искать максимум ln L – **логарифмической функции правдоподобия**. Для этого нужно:

1) найти производную $\dfrac{d\ln L}{d\Theta}$;

2) приравнять ее нулю (получим так называемое *уравнение правдоподобия*) и найти критическую точку;

3) найти вторую производную $\dfrac{d^2\ln L}{d\Theta^2}$; если она отрицательна в критической точке, то это – точка максимума.

Достоинства метода наибольшего правдоподобия: полученные оценки состоятельны (хотя могут быть смещенными), распределены асимптотически нормально при больших значениях n и имеют наименьшую дисперсию по сравнению с другими асимптотически нормальными оценками; если для оцениваемого параметра Θ существует эффективная оценка Θ*, то уравнение правдоподобия имеет единственное решение Θ*; метод наиболее полно

использует данные выборки и поэтому особенно полезен в случае малых выборок.
Недостаток метода наибольшего правдоподобия: сложность вычислений.

Для непрерывной случайной величины с известным видом плотности распределения $f(x)$ и неизвестным параметром Θ функция правдоподобия имеет вид:

$$L\,(x_1, x_2, \ldots, x_n; \Theta) = f(x_1,\Theta)f(x_2,\Theta)\ldots f(x_n,\Theta).$$

Оценка наибольшего правдоподобия неизвестного параметра проводится так же, как для дискретной случайной величины.

2. Метод моментов

Метод моментов основан на том, что начальные и центральные эмпирические моменты являются состоятельными оценками соответственно начальных и центральных теоретических моментов, поэтому можно приравнять теоретические моменты соответствующим эмпирическим моментам того же порядка.
Если задан вид плотности распределения $f(x, \Theta)$, определяемой одним неизвестным параметром Θ, то для оценки этого параметра достаточно иметь одно уравнение. Например, можно приравнять начальные моменты первого порядка:

$$\bar{x}_B = M(X) = \int_{-\infty}^{\infty} xf(x;\Theta)dx = \varphi(\Theta),$$

получив тем самым уравнение для определения Θ. Его решение Θ^* будет точечной оценкой параметра, которая является функцией от выборочного среднего и, следовательно, и от вариант выборки:

$$\Theta = \psi\,(x_1, x_2, \ldots, x_n).$$

Если известный вид плотности распределения $f(x, \Theta_1, \Theta_2)$ определяется двумя неизвестными параметрами Θ_1 и Θ_2, то требуется составить два уравнения, например

$$\nu_1 = M_1,\ \ \mu_2 = m_2.$$

Отсюда

$$\begin{cases} M(X) = \bar{x}_B \\ D(X) = D_B \end{cases} -$$

система двух уравнений с двумя неизвестными Θ_1 и Θ_2. Ее решениями будут точечные оценки Θ_1^* и Θ_2^* - функции вариант выборки:

$$\Theta_1 = \psi_1\,(x_1, x_2, \ldots, x_n),$$
$$\Theta_2 = \psi_2(x_1, x_2, \ldots, x_n).$$

3. Метод наименьших квадратов

Если требуется оценить зависимость величин y и x, причем известен вид связывающей их функции, но неизвестны значения входящих в нее коэффициентов, их величины можно оценить по имеющейся выборке с помощью метода наименьших квадратов. Для этого функция $y = \varphi(x)$ выбирается так, чтобы сумма квадратов отклонений наблюдаемых значений $y_1, y_2, \ldots, y_n$ от $\varphi(x_i)$ была минимальной:

$$\sum_{i=1}^{n}(y_i - \varphi(x_i))^2 = \min.$$

При этом требуется найти стационарную точку функции $\varphi(x;\ a,\ b,\ c\ldots)$, то есть решить систему:

$$\begin{cases} \sum\limits_{i=1}^{n}(y_i - \varphi(x_i; a, b, c\ldots))\left(\dfrac{\partial\varphi}{\partial a}\right)_i = 0 \\ \sum\limits_{i=1}^{n}(y_i - \varphi(x_i; a, b, c\ldots))\left(\dfrac{\partial\varphi}{\partial b}\right)_i = 0 \\ \sum\limits_{i=1}^{n}(y_i - \varphi(x_i; a, b, c\ldots))\left(\dfrac{\partial\varphi}{\partial c}\right)_i = 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{cases}$$

(решение, конечно, возможно только в случае, когда известен конкретный вид функции φ).

Рассмотрим в качестве примера подбор параметров линейной функции методом наименьших квадратов.

Для того, чтобы оценить параметры a и b в функции $y = ax + b$, найдем

$$\left(\frac{\partial\varphi}{\partial a}\right)_i = x_i; \quad \left(\frac{\partial\varphi}{\partial b}\right)_i = 1.$$

Тогда

$$\begin{cases} \sum\limits_{i=1}^{n}(y_i - (ax_i + b))x_i = 0 \\ \sum\limits_{i=1}^{n}(y_i - (ax_i + b)) = 0 \end{cases}.$$

Отсюда

$$\begin{cases} \sum\limits_{i=1}^{n}x_i y_i - a\sum\limits_{i=1}^{n}x_i^2 - b\sum\limits_{i=1}^{n}x_i = 0 \\ \sum\limits_{i=1}^{n}y_i - a\sum\limits_{i=1}^{n}x_i - bn = 0 \end{cases}.$$

Разделив оба полученных уравнения на n и вспомнив определения эмпирических моментов, можно получить выражения для a и b в виде:

$$a = \frac{(K_{xy})_B}{(D_x)_B}, \quad b = \overline{y}_B - \frac{(K_{xy})_B}{(D_x)_B}\overline{x}_B..$$

Следовательно, связь между x и y можно задать в виде:

$$y - \overline{y}_B = \frac{(K_{xy})_B}{(D_x)_B}(x - \overline{x}_B).$$

4. Байесовский подход к получению оценок

Пусть (Y, X) – случайный вектор, для которого известна плотность $p(y|x)$ условного распределения Y при каждом значении $X = x$. Если в результате эксперимента получены лишь значения Y, а соответствующие значения X неизвестны, то для оценки некоторой заданной функции $\varphi(x)$ в качестве ее приближенного значения предлагается искать условное математическое ожидание $M(\varphi(x)|Y)$, вычисляемое по формуле:

$$\psi(Y) = \frac{\int \varphi(x) p(Y \mid x) p(x) d\mu(x)}{q(Y)},$$

где

$$q(y) = \int p(y \mid x) p(x) d\mu(x),$$

$p(x)$ – плотность безусловного распределения X, $q(y)$ – плотность безусловного распределения Y. Задача может быть решена только тогда, когда известна $p(x)$. Иногда, однако, удается построить состоятельную оценку для $q(y)$, зависящую только от полученных в выборке значений Y.

Интервальное оценивание неизвестных параметров

При выборке малого объема точечная оценка может значительно отличаться от оцениваемого параметра, что приводит к грубым ошибкам. Поэтому в таком случае лучше пользоваться *интервальными оценками*, то есть указывать интервал, в который с заданной вероятностью попадает истинное значение оцениваемого параметра. Разумеется, чем меньше длина этого интервала, тем точнее оценка параметра. Поэтому, если для оценки Θ^* некоторого параметра Θ справедливо неравенство $|\Theta^* - \Theta| < \delta$, число $\delta > 0$ характеризует **точность оценки** (чем меньше δ, тем точнее оценка). Но статистические методы позволяют говорить только о том, что это неравенство выполняется с некоторой вероятностью.

Надежностью (доверительной вероятностью) оценки Θ^* параметра Θ называется вероятность γ того, что выполняется неравенство

$$|\Theta^* - \Theta| < \delta.$$

Если заменить это неравенство двойным неравенством

$$-\delta < \Theta^* - \Theta < \delta,$$

то получим:

$$p\ (\ \Theta^* - \delta < \Theta < \Theta^* + \delta\ ) = \gamma.$$

Таким образом, γ есть вероятность того, что Θ попадает в интервал (Θ* - δ, Θ* + δ).

Доверительным называется интервал, в который попадает неизвестный параметр с заданной надежностью γ.

Построение доверительных интервалов

1. Доверительный интервал для оценки математического ожидания нормального распределения при известной дисперсии.

Пусть исследуемая случайная величина X распределена по нормальному закону с известным средним квадратическим σ, и требуется по значению выборочного среднего $\bar{x}_B$ оценить ее математическое ожидание a. Будем рассматривать выборочное среднее $\bar{x}_B$ как случайную величину $\bar{X}$, а значения вариант выборки x_1, $x_2,\ldots$, x_n как одинаково распределенные независимые случайные величины X_1, $X_2,\ldots$, X_n, каждая из которых имеет математическое ожидание a и среднее квадратическое отклонение σ. При этом

$$M\left(\bar{X}\right) = a, \quad \sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

(используем свойства математического ожидания и дисперсии суммы независимых случайных величин). Оценим вероятность выполнения неравенства

$$|\,\bar{X} - a\,| < \delta.$$

Применим формулу для вероятности попадания нормально распределенной случайной величины в заданный интервал:

$$p\left(|\,X - a\,| < \delta\right) = 2\Phi\left(\frac{\delta}{\sigma}\right).$$

Тогда , с учетом того, что

$$\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}},$$

$$p\left(|\,X - a\,| < \delta\right) = 2\Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right) = 2\Phi(t),$$

$$где \quad t = \frac{\delta\sqrt{n}}{\sigma}.$$

Отсюда $\delta = \frac{t\sigma}{\sqrt{n}}$, и предыдущее равенство можно переписать так:

$$p\left(\bar{x}_B - \frac{t\sigma}{\sqrt{n}} < a < \bar{x}_B + \frac{t\sigma}{\sqrt{n}}\right) = 2\Phi(t) = \gamma.$$

Итак, значение математического ожидания a с вероятностью (надежностью) γ попадает в интервал

$$\left(\bar{x}_B - \frac{t\sigma}{\sqrt{n}}; \bar{x}_B + \frac{t\sigma}{\sqrt{n}}\right),$$

где значение t определяется из таблиц для функции Лапласа так, чтобы выполнялось равенство $2\Phi(t) = \gamma$.

Пример 1.

Найдем доверительный интервал для математического ожидания нормально распределенной случайной величины, если объем выборки n = 49, $\bar{x}_B = 2{,}8$, σ = 1,4, а доверительная вероятность γ = 0,9.

Определим t, при котором Ф(t) = 0,9:2 = 0,45: t = 1,645. Тогда

$$2{,}8 - \frac{1{,}645 \cdot 1{,}4}{\sqrt{49}} < a < 2{,}8 + \frac{1{,}645 \cdot 1{,}4}{\sqrt{14}},$$

или 2,471 < a < 3,129. Найден доверительный интервал, в который попадает a с надежностью 0,9.

2. Доверительный интервал для оценки математического ожидания нормального распределения при неизвестной дисперсии.

Если известно, что исследуемая случайная величина X распределена по нормальному закону с неизвестным средним квадратическим отклонением, то для поиска доверительного интервала для ее математического ожидания построим новую случайную величину

$$T = \frac{\bar{x}_B - a}{\frac{s}{\sqrt{n}}},$$

где $\bar{x}_B$ – выборочное среднее, s – исправленная дисперсия, n – объем выборки. Эта случайная величина, возможные значения которой будем обозначать t, имеет распределение Стьюдента с $k = n - 1$ степенями свободы.

Поскольку плотность распределения Стьюдента

$$s(t, n) = B_n \left(1 + \frac{t^2}{n-1}\right)^{-\frac{n}{2}},$$

где

$$B_n = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi(n-1)}\,\Gamma\left(\frac{n-1}{2}\right)},$$

явным образом не зависит от a и σ, можно задать вероятность ее попадания в некоторый интервал (- t_γ , t_γ), учитывая четность плотности распределения, следующим образом:

$$p\left(\left|\frac{\bar{x}_B - a}{\frac{s}{\sqrt{n}}}\right| < t_\gamma\right) = 2\int_0^{t_\gamma} s(t,n)dt = \gamma.$$

Отсюда получаем:

$$p\left(\bar{x}_B - \frac{t_\gamma s}{\sqrt{n}} < a < \bar{x}_B + \frac{t_\gamma s}{\sqrt{n}}\right) = \gamma.$$

Таким образом, получен доверительный интервал для a, где t_γ можно найти по соответствующей таблице при заданных n и γ.

Пример 2.

Пусть объем выборки $n = 25$, $\bar{x}_B = 3$, $s = 1{,}5$. Найдем доверительный интервал для a при $\gamma = 0{,}99$. Из таблицы находим, что t_γ ($n = 25$, $\gamma = 0{,}99$) = 2,797. Тогда

$$3 - \frac{2{,}797 \cdot 1{,}5}{\sqrt{25}} < a < 3 + \frac{2{,}797 \cdot 1{,}5}{\sqrt{25}},$$

или $2{,}161 < a < 3{,}839$ – доверительный интервал, в который попадает a с вероятностью 0,99.

3. Доверительные интервалы для оценки среднего квадратического отклонения нормального распределения.

Будем искать для среднего квадратического отклонения нормально распределенной случайной величины доверительный интервал вида $(s - \delta, s + \delta)$, где s – исправленное выборочное среднее квадратическое отклонение, а для δ выполняется условие: $p\ (|\sigma - s| < \delta) = \gamma$.

Запишем это неравенство в виде:

$$s\left(1 - \frac{\delta}{s}\right) < \sigma < s\left(1 + \frac{\delta}{s}\right)$$

или, обозначив $q = \dfrac{\delta}{s}$,

$$s(1-q) < \sigma < s(1+q). \qquad (1)$$

Рассмотрим случайную величину χ, определяемую по формуле

$$\chi = \frac{s}{\sigma}\sqrt{n-1},$$

которая распределена по закону «хи-квадрат» с n-1 степенями свободы. Плотность ее распределения

$$R(\chi, n) = \frac{\chi^{n-2} e^{-\frac{\chi^2}{2}}}{2^{\frac{n-3}{2}}\,\Gamma\left(\frac{n-1}{2}\right)}$$

не зависит от оцениваемого параметра σ, а зависит только от объема выборки n. Преобразуем неравенство (1) так, чтобы оно приняло вид $\chi_1 < \chi < \chi_2$.

Вероятность выполнения этого неравенства равна доверительной вероятности γ, следовательно,

$$\int_{\chi_1}^{\chi_2} R(\chi, n)d\chi = \gamma.$$

Предположим, что $q < 1$, тогда неравенство (1) можно записать так:

$$\frac{1}{s(1+q)} < \frac{1}{\sigma} < \frac{1}{s(1-q)},$$

или, после умножения на $s\sqrt{n-1}$,

$$\frac{\sqrt{n-1}}{1+q} < \frac{s\sqrt{n-1}}{\sigma} < \frac{\sqrt{n-1}}{1-q}.$$

Следовательно,

$$\frac{\sqrt{n-1}}{1+q} < \chi < \frac{\sqrt{n-1}}{1-q}.$$

Тогда

$$\int_{\frac{\sqrt{n-1}}{1+q}}^{\frac{\sqrt{n-1}}{1-q}} R(\chi, n)d\chi = \gamma.$$

Существуют таблицы для распределения «хи-квадрат», из которых можно найти q по заданным n и γ, не решая этого уравнения. Таким образом, вычислив по выборке значение s и определив по таблице значение q, можно найти доверительный интервал (18.4), в который значение σ попадает с заданной вероятностью γ.

Замечание. Если $q > 1$, то с учетом условия $\sigma > 0$ доверительный интервал для σ будет иметь границы

$$0 < \sigma < s(1+q).$$

Пример 3.

Пусть $n = 20$, $s = 1{,}3$. Найдем доверительный интервал для σ при заданной надежности $\gamma = 0{,}95$. Из соответствующей таблицы находим q ($n = 20$, $\gamma = 0{,}95$) = 0,37. Следовательно, границы доверительного интервала: 1,3(1-0,37) = 0,819 и 1,3(1+0,37) = 1,781. Итак, $0{,}819 < \sigma < 1{,}781$ с вероятностью 0,95.

ПРИМЕРЫ РЕШЕНИЯ ЗАДАЧ

Задача 1.

Плотность вероятности случайной величины X имеет вид:

$$f(x) = \frac{1}{\pi\left(1+(x-\theta)^2\right)}.$$

Оценить с помощью метода наибольшего правдоподобия значение параметра θ по выборке из двух значений: -1 и 1.

Указание

Найдите максимум функции правдоподобия

$$L(\theta)=\frac{1}{\pi\left(1+(x_1-\theta)^2\right)}\cdot\frac{1}{\pi\left(1+(x_2-\theta)^2\right)}.$$

Решение

Составим функцию правдоподобия

$$L(\theta)=\frac{1}{\pi\left(1+(x_1-\theta)^2\right)}\cdot\frac{1}{\pi\left(1+(x_2-\theta)^2\right)}=$$

$$\frac{1}{\pi^2\left(1+(\theta+1)^2\right)\left(1+(\theta-1)^2\right)}$$

и будем искать минимум функции

$$R(\theta)=\frac{1}{\pi^2 L(\theta)}=\left(1+(\theta+1)^2\right)\left(1+(\theta-1)^2\right)=\theta^4+1.$$

$$R_{\max}=R(0)=4 \quad при \quad \theta=0.$$

Ответ: 0.

Задача 2.

Имеется выборка значений случайной величины *X* – числа появлений события *A* в пяти испытаниях, распределенной по биномиальному закону. В таблице приведено число появлений *A* в 20 сериях по 5 испытаний в каждой:

x_i	0	1	2	3	4
n_i	2	6	5	4	3

Методом моментов найти точечную оценку параметра *p* биномиального распределения.

Указание

Воспользуйтесь тем, что математическое ожидание случайной величины, распределенной по биномиальному закону, равно *np*, и приравняйте его выборочному среднему.

Решение

Объем выборки *N* = 20, количество опытов в одной серии *n* =5.

$$\overline{x}_B=\frac{0\cdot2+1\cdot6+2\cdot5+3\cdot4+4\cdot3}{20}=2;$$

$$5p=2\Rightarrow p=0,4.$$

Ответ: 0,4.

Задача 3.

С помощью метода наименьших квадратов найти параметры зависимости между x и y для выборки

x_i	1,4	1,7	2,6	3,1	4,5	5,3
y_i	2,5	4,7	18,3	29,8	74,2	110,4

для случаев:

1) линейной зависимости $y = ax + b$;

2) квадратичной зависимости $y = (ax + b)^2$;

3) показательной зависимости $y = e^{ax + b}$.

Определить, какая из функций является лучшим приближением зависимости между x и y.

Указание

Для определения параметров каждого вида зависимости необходимо решить системы:

$$\begin{cases} \overline{x}_B \cdot a + b = \overline{y}_B \\ \dfrac{1}{n}\sum_{i=1}^{k} x_i^2 \cdot a + \overline{x}_B \cdot b = \dfrac{1}{n}\sum_{i=1}^{k} x_i y_i \end{cases} -$$

для линейной зависимости;

$$\begin{cases} \overline{x}_B \cdot a + b = \dfrac{1}{n}\sum_{i=1}^{k} \sqrt{y_i} \\ \dfrac{1}{n}\sum_{i=1}^{k} x_i^2 \cdot a + \overline{x}_B \cdot b = \dfrac{1}{n}\sum_{i=1}^{k} x_i \sqrt{y_i} \end{cases} -$$

для квадратичной зависимости;

$$\begin{cases} \overline{x}_B \cdot a + b = \dfrac{1}{n}\sum_{i=1}^{k} \ln y_i \\ \dfrac{1}{n}\sum_{i=1}^{k} x_i^2 \cdot a + \overline{x}_B \cdot b = \dfrac{1}{n}\sum_{i=1}^{k} x_i \cdot \ln y_i \end{cases} -$$

для показательной зависимости.

Лучший вид зависимости даст наименьшее значение выражения

$$\sum (y - y_i^*)^2,$$

где y^*_i – значение выбранной функции (линейной, квадратичной, показательной) с найденными коэффициентами при подстановке в нее значений x_i из выборки.

Решение

По виду выборки достаточно очевидно, что связь между x и y скорее всего не является линейной – y растет не пропорционально x. Проверим это предположение, найдя коэффициенты a и b для каждой из функций. Для этого вычислим предварительно

$$\overline{x}_B = 3,1; \quad \overline{y}_B = 40,0; \quad \frac{1}{n}\sum_{i=1}^{6} x_i^2 = 11,6;$$

$$\frac{1}{n}\sum_{i=1}^{6} x_i y_i = 178,4; \quad \frac{1}{n}\sum_{i=1}^{6} \sqrt{y_i} = 5,43;$$

$$\frac{1}{n}\sum_{i=1}^{6} x_i \sqrt{y_i} = 21,4; \quad \frac{1}{n}\sum_{i=1}^{6} \ln y_i = 2,96;$$

$$\frac{1}{n}\sum_{i=1}^{6} x_i \ln y_i = 11,05.$$

Теперь можно решать линейные системы для a и b:

1) $\begin{cases} 3,1a + b = 40 \\ 11,6a + 3,1b = 178,4 \end{cases} \Rightarrow a = 27,34, \quad b = -44,74,$

то есть линейная зависимость имеет вид: $y = 27,34x - 44,74$.

2) $\begin{cases} 3,1a + b = 5,43 \\ 11,6a + 3,1b = 21,4 \end{cases} \Rightarrow a = 2,29, \quad b = -1,68;$

квадратичная функция: $y = (2,29x - 1,68)^2$.

3) $\begin{cases} 3,1a + b = 2,96 \\ 11,6a + 3,1b = 11,05 \end{cases} \Rightarrow a = 0,94, \quad b = 0,04;$

показательная функция: $y = e^{0,94x + 0,04}$.

Вычислим значения

$$(y_i)_{лин} = 27,34x_i - 44,74; \quad (y_i)_{кв} = (2,29x_i - 1,68)^2;$$

$$(y_i)_{показ} = e^{0,94x_i + 0,04}:$$

:

y_i	2,5	4,7	18,3	29,8	74,2	110,4	$\sum(y - y_i^*)^2$
$(y_i)_{лин}$	-6,46	1,74	26,34	40,0	78,29	100,13	379,93
$(y_i)_{кв}$	2,33	4,9	18,27	29,37	74,4	109,35	1,397
$(y_i)_{показ}$	3,85	5,09	11,67	18,8	69,5	146,66	1503,81

Итак, наилучшим приближением является квадратичная функция

$$y = (2,29x - 1,68)^2.$$

Ответ: квадратичная, $y = (2,29x - 1,68)^2$.

Задача 4.

Дана выборка значений случайной величины, распределенной по нормальному закону со средним квадратическим отклонением $\sigma = 3$:

x_i	2	3	4	5	6
p_i	1	7	12	6	4

Найти границы доверительного интервала для $M(X)$ при доверительной вероятности $\gamma = 0,95$.

Указание

Найдите выборочное среднее и воспользуйтесь формулой

$$\overline{x}_B - \frac{t \cdot \sigma}{\sqrt{n}} < a < \overline{x}_B + \frac{t \cdot \sigma}{\sqrt{n}}, \quad \Phi(t) = \frac{\gamma}{2}.$$

Решение

$$n = 30; \quad \overline{x}_B = 4,167;$$

$$\Phi(t) = \frac{0,95}{2} = 0,475 \Rightarrow t = 1,96;$$

$$\frac{t \cdot \sigma}{\sqrt{n}} = \frac{1,96 \cdot 3}{\sqrt{30}} = 1,074;$$

$$4,167 - 1,074 < M(X) < 4,167 + 1,074;$$

$$3,093 < M(X) < 5,241.$$

Ответ: $(3,093; 5,241)$.

Задача 5.

Дана выборка значений нормально распределенной случайной величины: 2, 3, 3, 4, 2, 5, 5, 5, 6, 3, 6, 3, 4, 4, 4, 6, 5, 7, 3, 5. Найти с доверительной вероятностью γ = 0,95 границы доверительного интервала для математического ожидания.

Указание

Воспользуйтесь формулой

$$\overline{x}_B - \frac{t_\gamma \cdot s}{\sqrt{n}} < a < \overline{x}_B + \frac{t_\gamma \cdot s}{\sqrt{n}},$$

где

$$t_\gamma = t_\gamma(n, \gamma) -$$

критическая точка распределения Стьюдента.

Решение

Объем выборки n = 20. Найдем выборочное среднее и исправленное выборочное среднее квадратическое отклонение:

$$\overline{x}_B = 4,25; \quad s = 1,41.$$

По таблице значений критических точек распределения Стьюдента определим t (0,95; 20) = 2,093. Тогда

$$4,25-\frac{2,093\cdot 1,41}{\sqrt{20}}<a<4,25+\frac{2,093\cdot 1,41}{\sqrt{20}};\quad 3,59<a<4,91-$$

доверительный интервал для математического ожидания.

Ответ: $(3,59;4,91)$.

Задача 6.

Дана выборка значений нормально распределенной случайной величины: 2, 3, 3, 4, 2, 5, 5, 5, 6, 3, 6, 3, 4, 4, 4, 6, 5, 7, 3, 5. Найти с доверительной вероятностью γ = 0,95 границы доверительного интервала для среднего квадратического отклонения.

Указание

Воспользуйтесь формулой

$$s(1-q)<\sigma<s(1+q),$$

где s – исправленное выборочное среднее квадратическое отклонение, а $q = q\ (n, \gamma)$ – значение, определяемое из таблиц.

Решение

Объем выборки n = 20. Найдем $\bar{x}_B$ = 4,25, s = 1,41. По таблицам определим значение q (0,95; 20) = 0,37. Тогда

$$1,41(1-0,37)<\sigma<1,41(1+0,37);\quad 0,89<\sigma<1,93-$$

доверительный интервал для среднего квадратического отклонения.

Ответ: (0,89; 1,93).

2.2. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

2.2.1. *Проверка статистических гипотез. Критерии для проверки гипотез о вероятности события, о математическом ожидании, о сравнении двух дисперсий*

Статистической гипотезой называют гипотезу о виде неизвестного распределения генеральной совокупности или о параметрах известных распределений. **Нулевой (основной)** называют выдвинутую гипотезу H_0. **Конкурирующей (альтернативной)** называют гипотезу H_1, которая противоречит нулевой.

Пример 1.

Пусть H_0 заключается в том, что математическое ожидание генеральной совокупности a = 3. Тогда возможные варианты H_1: а) $a \neq 3$; б) $a > 3$; в) $a < 3$.

Простой называют гипотезу, содержащую только одно предположение, **сложной** – гипотезу, состоящую из конечного или бесконечного числа простых гипотез.

Пример 2.
Для показательного распределения гипотеза H_0: $\lambda = 2$ – простая, H_0: $\lambda > 2$ – сложная, состоящая из бесконечного числа простых (вида $\lambda = c$, где c – любое число, большее 2).

В результате проверки правильности выдвинутой нулевой гипотезы (такая проверка называется **статистической**, так как производится с применением методов математической статистики) возможны ошибки двух видов: **ошибка первого рода**, состоящая в том, что будет отвергнута правильная нулевая гипотеза, и **ошибка второго рода**, заключающаяся в том, что будет принята неверная гипотеза.
Замечание. Какая из ошибок является на практике более опасной, зависит от конкретной задачи. Например, если проверяется правильность выбора метода лечения больного, то ошибка первого рода означает отказ от правильной методики, что может замедлить лечение, а ошибка второго рода (применение неправильной методики) чревата ухудшением состояния больного и является более опасной.
Вероятность ошибки первого рода называется **уровнем значимости** α**.**

Основной прием проверки статистических гипотез заключается в том, что по имеющейся выборке вычисляется значение некоторой случайной величины, имеющей известный закон распределения.
Статистическим критерием называется случайная величина K с известным законом распределения, служащая для проверки нулевой гипотезы.
Критической областью называют область значений критерия, при которых нулевую гипотезу отвергают, **областью принятия гипотезы** – область значений критерия, при которых гипотезу принимают.

Итак, процесс проверки гипотезы состоит из следующих этапов:
1) выбирается статистический критерий K;
2) вычисляется его наблюдаемое значение $K_{набл}$ по имеющейся выборке;
3) поскольку закон распределения K известен, определяется (по известному уровню значимости α) **критическое значение** $k_{кр}$, разделяющее критическую область и область принятия гипотезы (например, если $p(K > k_{кр}) = \alpha$, то справа от $k_{кр}$ располагается критическая область, а слева – область принятия гипотезы);
4) если вычисленное значение $K_{набл}$ попадает в область принятия гипотезы, то нулевая гипотеза принимается, если в критическую область – нулевая гипотеза отвергается.

Различают разные виды критических областей:
- **правостороннюю** критическую область, определяемую неравенством $K > k_{кр}$ ($k_{кр} > 0$) (рис.1);

- **левостороннюю** критическую область, определяемую неравенством $K < k_{кр}$ ($k_{кр} < 0$) (рис.2);
- **двустороннюю** критическую область, определяемую неравенствами $K < k_1$, $K > k_2$ ($k_2 > k_1$) (рис.3).

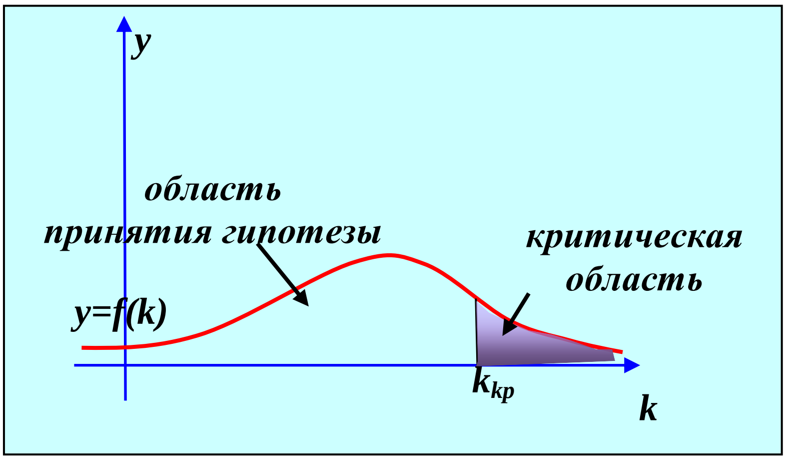


Рис. 1

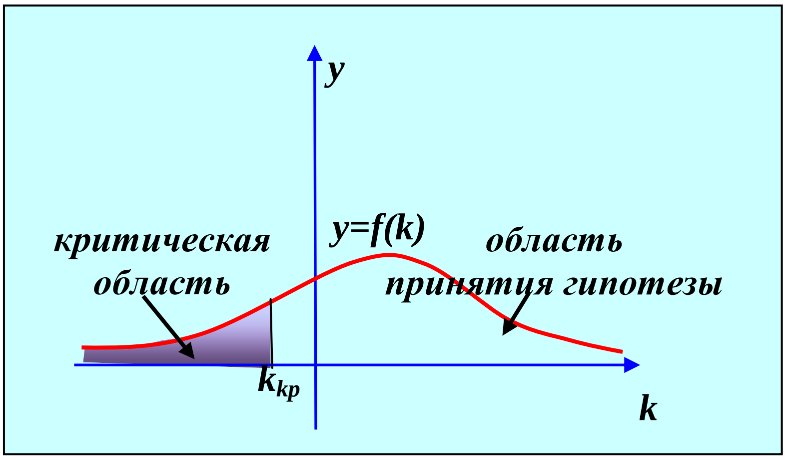


Рис. 2

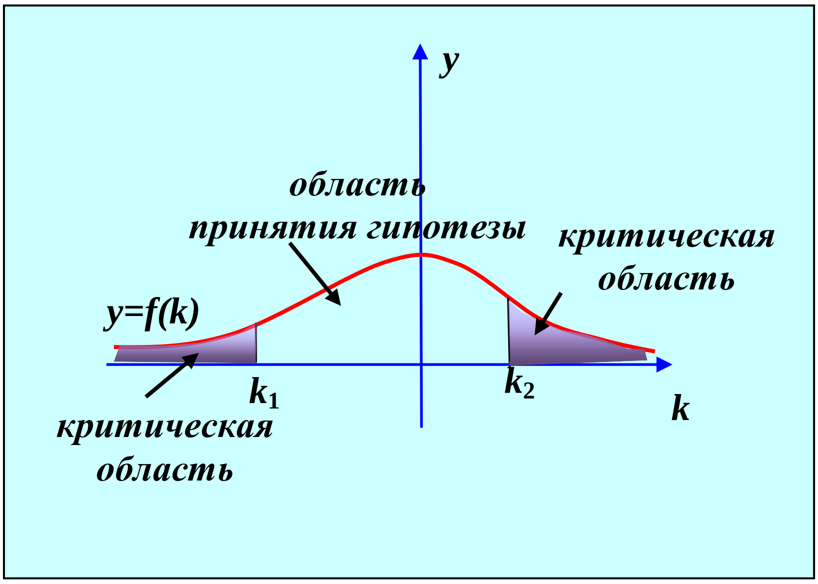


Рис. 3

Мощностью критерия называют вероятность попадания критерия в критическую область при условии, что верна конкурирующая гипотеза.

Если обозначить вероятность ошибки второго рода (принятия неправильной нулевой гипотезы) β, то мощность критерия равна $1 - \beta$. Следовательно, чем больше мощность критерия, тем меньше вероятность совершить ошибку второго рода. Поэтому после выбора уровня значимости следует строить критическую область так, чтобы мощность критерия была максимальной.

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

$$U = \frac{\left(\frac{M}{n} - p_0\right)\sqrt{n}}{\sqrt{p_0 q_0}},$$

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 $\frac{\alpha}{2}$. Поскольку U симметрична относительно оси Оy, вероятность ее попадания в интервалы $(-\infty;0)$ и $(0;+\infty)$ равна 0,5, следовательно, критическая область тоже должна быть симметрична относительно Оy. Поэтому $u_{кр}$ определяется по таблице значений функции Лапласа из условия

$$\Phi(u_{кр}) = \frac{1-\alpha}{2},$$

а критическая область имеет вид $(-\infty;-u_{кр}) \cup (u_{кр};+\infty)$.

Замечание. Предполагается, что используется таблица значений функции Лапласа, заданной в виде

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

где нижний предел интегрирования равен 0, а не $-\infty$. Функция Лапласа, заданная таким образом, является нечетной, а ее значения на 0,5 меньше, чем значения стандартной функции $\Phi(x)$.

Далее нужно вычислить наблюдаемое значение критерия:

$$U_{набл} = \frac{\left(\frac{m}{n} - p_0\right)\sqrt{n}}{\sqrt{p_0 q_0}}. \qquad (1)$$

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

$$p(0 < U < u_{кр}) = \frac{1}{2} - \alpha = \frac{1 - 2\alpha}{2}.$$

Следовательно, $u_{кр}$ можно найти по таблице значений функции Лапласа из условия, что

$$\Phi(u_{кр}) = \frac{1 - 2\alpha}{2}.$$

Вычислим наблюдаемое значение критерия по формуле (1).

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 так же, как в предыдущем случае.

Если $U_{набл} > - u_{кр}$, то нулевая гипотеза принимается.

Если $U_{набл} < - u_{кр}$, то нулевая гипотеза отвергается.

Пример 3.

Пусть проведено 50 независимых испытаний, и относительная частота появления события A оказалась равной 0,12. Проверим при уровне значимости $\alpha = 0{,}01$ нулевую гипотезу H_0: $p = 0{,}1$ при конкурирующей гипотезе H_1: $p > 0{,}1$. Найдем

$$U_{набл} = \frac{(0{,}12 - 0{,}1)\sqrt{50}}{\sqrt{0{,}1 \cdot 0{,}9}} = 0{,}471.$$

Критическая область является правосторонней, а $u_{кр}$ находим из равенства

$$\Phi\left(u_{кр}\right)\frac{1 - 2 \cdot 0{,}01}{2} = 0{,}49.$$

Из таблицы значений функции Лапласа определяем $u_{кр}$ = 2,33. Итак, $U_{набл} < u_{кр}$, и гипотеза о том, что $p = 0{,}1$, принимается.

Критерий для проверки гипотезы о математическом ожидании

Пусть генеральная совокупность X имеет нормальное распределение, и требуется проверить предположение о том, что ее математическое ожидание равно некоторому числу a_0. Рассмотрим две возможности.

1) Известна дисперсия σ^2 генеральной совокупности. Тогда по выборке объема n найдем выборочное среднее $\bar{x}_В$ и проверим нулевую гипотезу H_0: $M(X) = a_0$.

Учитывая, что выборочное среднее $\bar{X}$ является несмещенной оценкой $M(X)$, то есть

$$M(\bar{X}) = M(X),)$$

можно записать нулевую гипотезу так:

$$M(\bar{X}) = a_0.$$

Для ее проверки выберем критерий

$$U = \frac{\bar{X} - a_0}{\sigma(\bar{X})} = \frac{(\bar{X} - a_0)\sqrt{n}}{\sigma}.$$

Это случайная величина, имеющая нормальное распределение, причем, если нулевая гипотеза справедлива, то $M(U) = 0$, $\sigma(U) = 1$.

Выберем критическую область в зависимости от вида конкурирующей гипотезы:

- если

$$H_1 : M(\bar{X}) \neq a_0,$$

то

$$u_{кр} : \Phi(u_{кр}) = \frac{1-\alpha}{2},$$

критическая область двусторонняя,

$$U_{набл} = \frac{(\bar{x} - a_0)\sqrt{n}}{\sigma},$$

и, если $|U_{набл}| < u_{кр}$, то нулевая гипотеза принимается; если $|U_{набл}| > u_{кр}$, то нулевая гипотеза отвергается.

- если

$$H_1 : M(\bar{X}) > a_0,$$

то

$$u_{кр} : \Phi(u_{кр}) = \frac{1-2\alpha}{2},$$

критическая область правосторонняя, и, если $U_{набл} < u_{кр}$, то нулевая гипотеза принимается; если $U_{набл} > u_{кр}$, то нулевая гипотеза отвергается.

- если

$$H_1 : M(\bar{X}) < a_0,$$

то

$$u_{кр} : \Phi(u_{кр}) = \frac{1-2\alpha}{2},$$

критическая область левосторонняя, и, если $U_{набл} > -\, u_{кр}$, то нулевая гипотеза принимается; если $U_{набл} < -\, u_{кр}$, то нулевая гипотеза отвергается.
2) Дисперсия генеральной совокупности неизвестна.

В этом случае выберем в качестве критерия случайную величину

$$T = \frac{(\bar{X} - a_0)\sqrt{n}}{S},$$

где S – исправленное среднее квадратическое отклонение. Такая случайная величина имеет распределение Стьюдента с $k = n - 1$ степенями свободы. Рассмотрим те же, что и в предыдущем случае, конкурирующие гипотезы и соответствующие им критические области. Предварительно вычислим наблюдаемое значение критерия:

$$T_{набл} = \frac{(\bar{x}_B - a_0)\sqrt{n}}{S}.$$

- если

$$H_1 : M(\bar{X}) \neq a_0,$$

то критическая точка $t_{двуст.кр.}$ находится по таблице критических точек распределения Стьюдента по известным α и $k = n - 1$.
Если $|\, T_{набл} \,| < t_{двуст.кр.}$, то нулевая гипотеза принимается.
Если $|\, T_{набл} \,| > t_{двуст.кр.}$, то нулевая гипотеза отвергается.
- если

$$H_1 : M(\bar{X}) > a_0,$$

то по соответствующей таблице находят $t_{правост.кр.}(\alpha, k)$ – критическую точку правосторонней критической области. Нулевая гипотеза принимается, если $T_{набл} < t_{правост.кр.}$.
- при конкурирующей гипотезе

$$H_1 : M(\bar{X}) < a_0$$

критическая область является левосторонней, и нулевая гипотеза принимается при условии $T_{набл} > -\, t_{правост.кр.}$. Если $T_{набл} < -\, t_{правост.кр.}$, нулевую гипотезу отвергают.

Критерий для проверки гипотезы о сравнении двух дисперсий

Пусть имеются две нормально распределенные генеральные совокупности X и Y. Из них извлечены независимые выборки объемов соответственно n_1 и n_2, по которым вычислены исправленные выборочные дисперсии s_X^2 и s_Y^2. Требуется при заданном уровне значимости α проверить нулевую гипотезу H_0: $D(X) = D(Y)$ о равенстве дисперсий рассматриваемых генеральных совокупностей. Учитывая несмещенность исправленных выборочных дисперсий, можно записать нулевую гипотезу так:

$$H_0 : M\left(s_X^2\right) = M\left(s_Y^2\right).$$

Замечание. Конечно, исправленные дисперсии, вычисленные по выборкам, обычно оказываются различными. При проверке гипотезы выясняется, является ли это различие незначимым и обусловленным случайными причинами (в случае принятия нулевой гипотезы) или оно является следствием того, что сами генеральные дисперсии различны.

В качестве критерия примем случайную величину

$$F = \frac{S_б^2}{S_м^2} \quad -$$

отношение большей выборочной дисперсии к меньшей. Она имеет распределение Фишера-Снедекора со степенями свободы $k_1 = n_1 - 1$ и $k_2 = n_2 - 1$, где n_1 – объем выборки, по которой вычислена большая исправленная дисперсия, а n_2 – объем второй выборки. Рассмотрим два вида конкурирующих гипотез:

- пусть H_1: $D(X) > D(Y)$. Наблюдаемым значением критерия будет отношение большей из исправленных дисперсий к меньшей:

$$F_{набл} = \frac{s_б^2}{s_м^2}.$$

По таблице критических точек распределения Фишера-Снедекора можно найти критическую точку $F_{набл}(\alpha;\ k_1;\ k_2)$. При $F_{набл} < F_{кр}$ нулевая гипотеза принимается, при $F_{набл} > F_{кр}$ отвергается.

- если

$$H_1 : D(X) \neq D(Y),$$

то критическая область является двусторонней и определяется неравенствами $F < F_1$, $F > F_2$, где $p(F < F_1) = p(\ F > F_2) = \alpha/2$. При этом достаточно найти правую критическую точку

$$F_2 = F_{кр}\left(\frac{\alpha}{2}, k_1, k_2\right).$$

ПРИМЕРЫ РЕШЕНИЯ ЗАДАЧ

Задача 1.

Проведено 90 независимых испытаний, и относительная частота появления события A оказалась равной 0,27. Проверить при уровне значимости $\alpha = 0{,}02$ нулевую гипотезу H_0: $p = 0{,}3$ при конкурирующей гипотезе H_1: $p < 0{,}3$. В ответе указать значение $U_{набл}$ и сделать вывод о принятии или отклонении нулевой гипотезы.

Указание

Найдите $u_{кр}$ из условия

$$\Phi(u_{кр}) = \frac{1-2\alpha}{2},$$

а $U_{набл}$ – по формуле

$$U_{набл} = \frac{\left(\frac{m}{n} - p_0\right)\sqrt{n}}{\sqrt{p_0 q_0}}.$$

Критическая область в данном случае является левосторонней, поэтому если $U_{набл} > -u_{кр}$, то нулевая гипотеза принимается, а если $U_{набл} < -u_{кр}$, то нулевая гипотеза отвергается.

Решение

$$\Phi(u_{кр}) = \frac{1-2\alpha}{2} = \frac{1-0,04}{2} = 0,48 \Rightarrow u_{кр} = 2,06;$$

$$\frac{m}{n} = 0,27, \quad n = 90, \quad p_0 = 0,3 \Rightarrow$$

$$\Rightarrow U_{набл} = \frac{\left(\frac{m}{n} - p_0\right)\sqrt{n}}{\sqrt{p_0 q_0}} = \frac{(0,27-0,3)\sqrt{90}}{\sqrt{0,3 \cdot 0,7}} = -0,62.$$

$$U_{набл} > -u_{кр} \quad -$$

нулевая гипотеза принимается.

Ответ: $U_{набл}$ = -0,62; гипотеза H_0 принимается.

Задача 2.

В серии из 20 независимых испытаний событие A появилось 8 раз, в серии из 15 испытаний – 7 раз. При уровне значимости α = 0,05 проверить нулевую гипотезу H_o: $p_1 = p_2$ (о равенстве двух вероятностей биномиальных распределений) при конкурирующей гипотезе H_1: $p_1 < p_2$. В ответе указать значение $U_{набл}$ и сделать вывод о принятии или отклонении нулевой гипотезы.

Указание

При проверке гипотезы о равенстве двух вероятностей биномиальных распределений наблюдаемое значение критерия вычисляется по формуле

$$U_{набл} = \frac{\frac{m_1}{n_1} - \frac{m_2}{n_2}}{\sqrt{\frac{m_1+m_2}{n_1+n_2}\left(1 - \frac{m_1+m_2}{n_1+n_2}\right) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

а критическое значение и вид критической области такие же, как при проверке гипотезы о вероятности события.

Решение

Критическая область – левосторонняя,

$$\Phi(u_{кр}) = \frac{1-2\cdot 0,05}{2} = 0,45,$$

следовательно, $u_{кр}$ = 1,645, и критическая область имеет вид U < - 1,645. Вычислим

$$U_{набл} = \frac{\frac{8}{20} - \frac{7}{15}}{\sqrt{\frac{15}{35}\cdot\frac{20}{35}\cdot\left(\frac{1}{20}+\frac{1}{15}\right)}} = -0,394.$$

$U_{набл} > -u_{кр}$, следовательно, гипотеза принимается, и можно считать, что вероятность события A в обеих сериях испытаний одинакова.

Ответ: $U_{набл}$ = -0,394; гипотеза H_0 принимается.

Задача 3.

По выборке из 100 значений нормально распределенной случайной величины с дисперсией, равной 4, найдено выборочное среднее $\overline{x}_B = 8,59$.

При уровне значимости α = 0,02 проверить нулевую гипотезу H_o: $M(X)$ = 8 при конкурирующей гипотезе H_1: $M(X)$ > 8. В ответе указать значение $U_{набл}$ и сделать вывод о принятии или отклонении нулевой гипотезы.

Указание

Воспользуйтесь формулами

$$U_{набл} = \frac{(\overline{x}-a_0)\sqrt{n}}{\sigma}, \quad u_{кр} : \Phi(u_{кр}) = \frac{1-2\alpha}{2}.$$

Решение

Критическая область правосторонняя. Найдем $U_{набл}$ и $u_{кр}$:

$$\sigma = \sqrt{D(X)} = \sqrt{4} = 2;$$

$$U_{набл} = \frac{(\overline{x}-a_0)\sqrt{n}}{\sigma} = \frac{(8,59-8)\cdot 10}{2} = 2,95;$$

$$\Phi(u_{кр}) = \frac{1-2\alpha}{2} = \frac{1-0,02}{2} = 0,49 \Rightarrow u_{кр} = 2,32.$$

$$U_{набл} > u_{кр} \quad -$$

гипотеза H_0 отклоняется.

Ответ: $U_{набл}$ = 2,95; гипотеза H_0 отклоняется.

Задача 4.

Имеются независимые выборки значений нормально распределенных случайных величин

X: 2, 2, 3, 3, 4, 4, 4, 5, 5, 6

и

Y: 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5, 5, 7, 8, 9.

Проверить для уровня значимости α = 0,1 при условии равенства генеральных дисперсий нулевую гипотезу H_o: M (X) = M (Y) при конкурирующей гипотезе

$$H_1 : M(X) \neq M(Y).$$

В ответе указать значение $T_{набл}$ и сделать вывод о принятии или отклонении нулевой гипотезы.

Указание

Наблюдаемое значение критерия для проверки гипотезы о равенстве генеральных средних для нормально распределенных случайных величин, дисперсии которых неизвестны, а выборки малы, вычисляется по формуле

$$T_{набл} = \frac{\overline{x}_B - \overline{y}_B}{\sqrt{(m-1)s_x^2 + (n-1)s_y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}},$$

где m и n – объемы выборок. Критическое значение $t_{двуст.кр.}(\alpha,\ k)$ или $t_{прав.кр}$ находится из таблицы критических точек распределения Стьюдента ($k = n + m - 2$ – число степеней свободы).

Критическая область строится в зависимости от вида конкурирующей гипотезы.

а) H_1: M (X) $\neq$ M (Y) – критическая область двусторонняя, задаваемая неравенством $|T| > t_{двуст.кр}$.

б) H_1: M (X) > M (Y) – критическая область правосторонняя, определяемая условием $T > t_{прав.кр.}$.

в) H_1: M (X) < M (Y) – критическая область левосторонняя, $T < -\ t_{прав.кр.}$.

Решение

Объемы выборок m = 10, n = 15. Вычислим выборочные средние и исправленные выборочные дисперсии:

$$\overline{x}_B = 3,8; \quad \overline{y}_B = 4,93; \quad s_x^2 = 1,73; \quad s_y^2 = 3,21.$$

Вычислим наблюдаемое значение критерия:

$$T_{набл} = \frac{3,8 - 4,93}{\sqrt{9 \cdot 1,73 + 14 \cdot 3,21}} \sqrt{\frac{10 \cdot 15 \cdot 23}{25}} = -1,706.$$

Критическая область – двусторонняя, $t_{двуст.кр.}(0,1;\ 23) = 1,71$. Итак, $|T_{набл}| < t_{двуст.кр.}$, следовательно, нет оснований отвергнуть нулевую гипотезу – можно считать, что математические ожидания генеральных совокупностей равны.

Ответ: $T_{набл}$ = -1,706; гипотеза H_0 принимается.

Задача 5.

Даны две независимые выборки объемов n_1 = 10 и n_2 = 15, извлеченные из генеральных совокупностей X и Y, распределенных по нормальному закону. Найдены исправленные выборочные дисперсии

$$s_x^2 = 2,67 \quad и \quad s_y^2 = 1,88.$$

Проверить при уровне значимости α = 0,05 нулевую гипотезу о равенстве генеральных дисперсий при конкурирующей гипотезе H_1: D (X) > D (Y). В ответе указать значение $F_{набл}$ и сделать вывод о принятии или отклонении нулевой гипотезы.

Указание

Критерием служит случайная величина

$$F = \frac{s_б^2}{s_м^2} -$$

отношение большей исправленной дисперсии к меньшей, которая при условии справедливости нулевой гипотезы имеет распределение Фишера-Снедекора со степенями свободы $k_1 = n_1 - 1$ и $k_2 = n_2 - 1$, где n_1 и n_2 – объемы выборок. Критическая область зависит от вида конкурирующей гипотезы: если H_1: D (X) > D (Y), то критическая область правосторонняя:

$$p(F > F_{кр}(\alpha, k_1, k_2)) = \alpha.$$

Критическая точка $F_{кр}(\alpha, k_1, k_2)$ находится по таблице критических точек распределения Фишера-Снедекора. Если

$$F_{набл} = \frac{s_б^2}{s_м^2} < F_{кр} -$$

нулевая гипотеза принимается, в противном случае – отвергается.

Решение

Найдем значение

$$F_{кр}(0,05;\ 9;\ 14) = 2,65.$$

Критическая область – правосторонняя. Вычислим наблюдаемое значение критерия:

$$F_{набл} = \frac{2,67}{1,88} = 1,42 < F_{кр}.$$

Следовательно, нет оснований отвергнуть нулевую гипотезу.

Ответ: $F_{набл}$ = 1,42; гипотеза H_0 принимается.

2.2.2. *Критерии согласия*

В предыдущей лекции рассматривались гипотезы, в которых закон распределения генеральной совокупности предполагался известным. Теперь займемся проверкой гипотез о предполагаемом законе неизвестного распределения, то есть будем проверять нулевую гипотезу о том, что генеральная совокупность распределена по некоторому известному закону. Обычно статистические критерии для проверки таких гипотез называются **критериями согласия**.

Критерий Пирсона

Достоинством критерия Пирсона является его универсальность: с его помощью можно проверять гипотезы о различных законах распределения.

1. Проверка гипотезы о нормальном распределении

Пусть получена выборка достаточно большого объема n с большим количеством различных значений вариант. Доя удобства ее обработки разделим интервал от наименьшего до наибольшего из значений вариант на s равных частей и будем считать, что значения вариант, попавших в каждый интервал, приближенно равны числу, задающему середину интервала. Подсчитав число вариант, попавших в каждый интервал, составим так называемую сгруппированную выборку:

варианты.$x_1 \quad x_2 \quad \ldots \quad x_s$

частоты.$n_1 \quad n_2 \quad \ldots \quad n_s$,

где x_i – значения середин интервалов, а n_i – число вариант, попавших в i-й интервал (эмпирические частоты).

По полученным данным можно вычислить выборочное среднее $\bar{x}_B$ и выборочное среднее квадратическое отклонение σ_B. Проверим предположение, что генеральная совокупность распределена по нормальному закону с параметрами

$$M(X) = \bar{x}_B, \quad D(X) = \sigma_B^2.$$

Тогда можно найти количество чисел из выборки объема n, которое должно оказаться в каждом интервале при этом предположении (то есть теоретические частоты). Для этого по таблице значений функции Лапласа найдем вероятность попадания в i-й интервал:

$$p_i = \Phi\left(\frac{b_i - \bar{x}_B}{\sigma_B}\right) - \Phi\left(\frac{a_i - \bar{x}_B}{\sigma_B}\right),$$

где a_i и b_i - границы i-го интервала. Умножив полученные вероятности на объем выборки n, найдем теоретические частоты: $n_i = n \cdot p_i$. Наша цель – сравнить эмпирические и теоретические частоты, которые, конечно, отличаются друг от друга, и выяснить, являются ли эти различия несущественными, не опровергающими гипотезу о нормальном распределении исследуемой случайной величины, или они настолько велики, что противоречат этой гипотезе. Для этого используется критерий в виде случайной величины

$$\chi^2 = \sum_{i=1}^{s} \frac{(n_i - n_i')^2}{n_i'}.$$

Смысл ее очевиден: суммируются части, которые квадраты отклонений эмпирических частот от теоретических составляют от соответствующих теоретических частот. Можно доказать, что вне зависимости от реального закона распределения генеральной совокупности закон распределения этой случайной величины при $n \to \infty$ стремится к закону распределения χ^2 с числом степеней свободы k = s – 1 – r, где r – число параметров предполагаемого распределения, оцененных по данным выборки. Нормальное распределение характеризуется двумя параметрами, поэтому k = s – 3. Для выбранного критерия строится правосторонняя критическая область, определяемая условием

$$p\left(\chi^2 > \chi^2_{kp}(\alpha, k)\right) = \alpha,$$

где α – уровень значимости. Следовательно, критическая область задается неравенством

$$\chi^2 > \chi^2_{kp}(\alpha, k),$$

а область принятия гипотезы -

$$\chi^2 < \chi^2_{kp}(\alpha, k).$$

Итак, для проверки нулевой гипотезы H_0: генеральная совокупность распределена нормально – нужно вычислить по выборке наблюдаемое значение критерия:

$$\chi^2_{набл} = \sum_{i=1}^{s} \frac{(n_i - n_i')^2}{n_i'}, \qquad (1)$$

а по таблице критических точек распределения χ^2 найти критическую точку $\chi^2_{kp}(\alpha, k)$, используя известные значения α и k = s – 3. Если $\chi^2_{набл} < \chi^2_{kp}$ – нулевую гипотезу принимают, при $\chi^2_{набл} > \chi^2_{kp}$ ее отвергают.

2. Проверка гипотезы о равномерном распределении

При использовании критерия Пирсона для проверки гипотезы о равномерном распределении генеральной совокупности с предполагаемой плотностью вероятности

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & x \in (a,b) \\ 0, & x \notin (a,b) \end{cases}$$

необходимо, вычислив по имеющейся выборке значение $\bar{x}_B$, оценить параметры a и b по формулам:

$$a^* = \bar{x}_B - \sqrt{3}\sigma_B, \quad b^* = \bar{x}_B + \sqrt{3}\sigma_B, \quad (2)$$

где a^* и b^* - оценки a и b. Действительно, для равномерного распределения

$$M(X) = \frac{a+b}{2}, \quad \sigma(x) = \sqrt{D(X)} = \sqrt{\frac{(b-a)^2}{12}} = \frac{b-a}{2\sqrt{3}},$$

откуда можно получить систему для определения a^* и b^*:

$$\begin{cases} \dfrac{b^* + a^*}{2} = \bar{x}_B \\ \dfrac{b^* - a^*}{2\sqrt{3}} = \sigma_B \end{cases},$$

решением которой являются выражения (2).
Затем, предполагая, что

$$f(x)=\frac{1}{b^*-a^*},$$

можно найти теоретические частоты по формулам

$$n_1'=np_1=nf(x)(x_1-a^*)=n\cdot\frac{1}{b^*-a^*}(x_1-a^*);$$

$$n_2'=n_3'=\ldots=n_{s-1}'=n\cdot\frac{1}{b^*-a^*}(x_i-x_{i-1}),\ i=1,2,\ldots,s-1;$$

$$n_s'=n\cdot\frac{1}{b^*-a^*}(b^*-x_{s-1}).$$

Здесь s – число интервалов, на которые разбита выборка.
Наблюдаемое значение критерия Пирсона вычисляется по формуле (1), а критическое – по таблице с учетом того, что число степеней свободы $k = s - 3$. После этого границы критической области определяются так же, как и для проверки гипотезы о нормальном распределении.

3. Проверка гипотезы о показательном распределении

В этом случае, разбив имеющуюся выборку на равные по длине интервалы, рассмотрим последовательность вариант

$$x_i^*=\frac{x_i+x_{i+1}}{2},$$

равноотстоящих друг от друга (считаем, что все варианты, попавшие в i – й интервал, принимают значение, совпадающее с его серединой), и соответствующих им частот n_i (число вариант выборки, попавших в i – й интервал). Вычислим по этим данным $\bar{x}_B$ и примем в качестве оценки параметра λ величину $\lambda^*=\frac{1}{\bar{x}_B}$. Тогда теоретические частоты вычисляются по формуле

$$n_i'=np_i=np(x_i<X<x_{i+1})=n(e^{-\lambda x_i}-e^{-\lambda x_{i+1}}).$$

Затем сравниваются наблюдаемое и критическое значение критерия Пирсона с учетом того, что число степеней свободы $k = s - 2$.

Критерий Колмогорова

Этот критерий применяется для проверки простой гипотезы H_0 о том, что независимые одинаково распределенные случайные величины $X_1, X_2, \ldots, X_n$ имеют заданную непрерывную функцию распределения $F(x)$.
Найдем функцию эмпирического распределения $F_n(x)$ и будем искать границы двусторонней критической области, определяемой условием

$$D_n=\sup_{|x|<\infty}|F_n(x)-F(x)|>\lambda_n.$$

А.Н.Колмогоров доказал, что в случае справедливости гипотезы H_0 распределение статистики D_n не зависит от функции $F(x)$, и при $n\to\infty$

$$p(\sqrt{n}D_n<\lambda)\to K(\lambda),\quad \lambda>0,$$

где

$$K(\lambda)=\sum_{m=-\infty}^{\infty}(-1)^m e^{-2m^2\lambda^2}\quad -$$

- критерий Колмогорова, значения которого можно найти в соответствующих таблицах. Критическое значение критерия $\lambda_n(\alpha)$ вычисляется по заданному уровню значимости α как корень уравнения

$$p(D_n\geq\lambda)=\alpha.$$

Можно показать, что приближенное значение вычисляется по формуле

$$\lambda_n(\alpha)\approx\sqrt{\frac{z}{2n}}-\frac{1}{6n},$$

где z – корень уравнения

$$1-K\left(\sqrt{\frac{\lambda}{2}}\right)=\alpha.$$

На практике для вычисления значения статистики D_n используется то, что

$$D_n=\max(D_n^+,D_n^-),,$$

где

$$D_n^+=\max_{1\leq m\leq n}\left(\frac{m}{n}-F(X_{(m)})\right),\quad D_n^-=\max_{1\leq m\leq n}\left(F(X_{(m)})-\frac{m-1}{n}\right),$$

а

$$X_{(1)}\leq X_{(2)}\leq\ldots\leq X_{(n)}\quad -$$

вариационный ряд, построенный по выборке X_1, X_2, …, X_n.

Можно дать следующее геометрическое истолкование критерия Колмогорова: если изобразить на плоскости Оxy графики функций $F_n(x)$, $F_n(x)\ \underline{+}\lambda_n(\alpha)$ (рис. 1), то гипотеза H_0 верна, если график функции $F(x)$ не выходит за пределы области, лежащей между графиками функций $F_n(x)$ -$\lambda_n(\alpha)$ и $F_n(x)$ +$\lambda_n(\alpha)$.

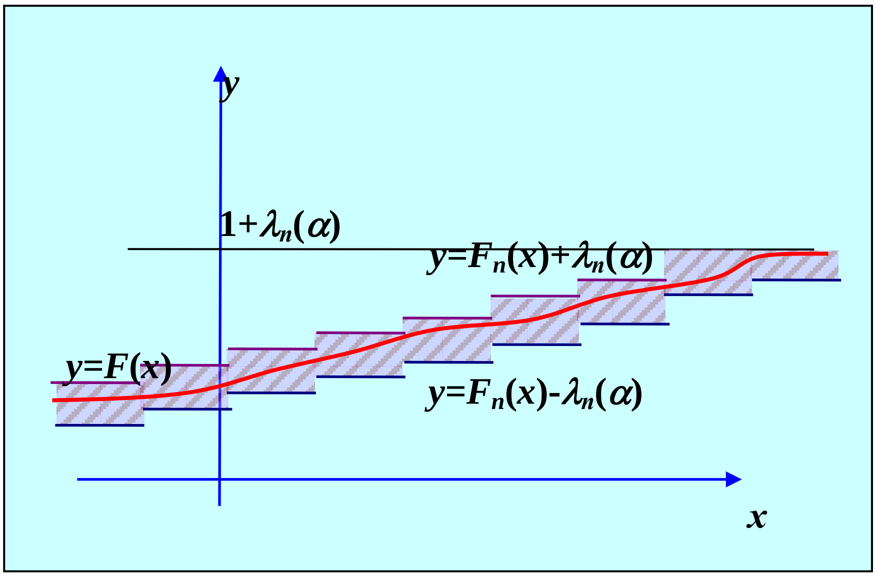


Рис. 1

Приближенный метод проверки нормальности распределения, связанный с оценками коэффициентов асимметрии и эксцесса

Определим по аналогии с соответствующими понятиями для теоретического распределения асимметрию и эксцесс эмпирического распределения.
Асимметрия эмпирического распределения определяется равенством

$$a_s = \frac{m_3}{\sigma_B^3},$$

где m_3 – центральный эмпирический момент третьего порядка.
Эксцесс эмпирического распределения определяется равенством

$$e_k = \frac{m_4}{\sigma_B^4} - 3,$$

где m_4 – центральный эмпирический момент четвертого порядка.
Как известно, для нормально распределенной случайной величины асимметрия и эксцесс равны 0. Поэтому, если соответствующие эмпирические величины достаточно малы, можно предположить, что генеральная совокупность распределена по нормальному закону.

ПРИМЕРЫ РЕШЕНИЯ ЗАДАЧ

Задача 1.
Для выборки, интервальный статистический ряд которой имеет вид

Номер интервала	Границы интервала	Эмпирические частоты
1	2 – 5	6
2	5 – 8	8
3	8 – 11	15
4	11 – 14	22
5	14 – 17	14
6	17 – 20	5

проверить при уровне значимости α = 0,05 гипотезу о показательном законе распределения генеральной совокупности с помощью критерия Пирсона. В ответе указать наблюдаемое значение критерия и сделать вывод о принятии или отклонении гипотезы.

Указание

Для проверки гипотезы о показательном распределении генеральной совокупности в качестве оценки параметра λ принимается

$$\lambda^* = \frac{1}{\overline{x}_B}.$$

Тогда теоретические частоты

$$n_i' = n \cdot P_i, \quad P_i = e^{-\lambda x_i} - e^{-\lambda x_{i+1}}.$$

Показательное распределение определяется одним параметром, поэтому число степеней свободы $k = n - 2$. Наблюдаемое значение критерия Пирсона

$$\chi^2_{набл} = \sum \frac{(n_i - n_i')^2}{n_i'}.$$

Критическая область выбирается правосторонней, и граница ее при заданном уровне значимости α $\chi^2_{кр}(\alpha, k)$ находится по таблице критических точек распределения χ^2. Здесь $k = s - 2$, где s – число частичных интервалов выборки.

Решение

Объем выборки $n = 70$. Будем считать вариантами середины частичных интервалов: $x_1 = 3{,}5$, $x_2 = 6{,}5, \ldots, x_6 = 18{,}5$.

Найдем $\bar{x}_B = 11{,}43$; $\sigma_B = 4{,}03$; $s = 4{,}05$.

Вычислим теоретические частоты в предположении о показательном распределении генеральной совокупности при $\lambda^* = \dfrac{1}{11{,}43} = 0{,}087$:

$$n_1' = 70\left(e^{-0{,}087\cdot 2} - e^{-0{,}087\cdot 5}\right) = 70\left(e^{-0{,}174} - e^{-0{,}435}\right) = 13{,}44;$$

аналогично $n_2' = 10{,}37;$ $n_3' = 8{,}05;$ $n_4' = 6{,}23;$ $n_5' = 4{,}76;$ $n_6' = 3{,}64.$

Наблюдаемое значение критерия

$$\chi^2_{набл} = \frac{(6 - 13{,}44)^2}{13{,}44} + \ldots + \frac{(5 - 3{,}64)^2}{3{,}64} = 69{,}02.$$

Критическая точка

$$\chi^2_{набл}(0{,}05;4) = 9{,}5; \quad \chi^2_{набл} > \chi^2_{кр},$$

и гипотеза о показательном распределении отклоняется.

Ответ: $\chi^2_{набл} = 69{,}02$, гипотеза отклоняется.

Задача 2.

Для выборки, интервальный статистический ряд которой имеет вид

Номер интервала	Границы интервала	Эмпирические частоты
1	2 – 5	6
2	5 – 8	8
3	8 – 11	15
4	11 – 14	22
5	14 – 17	14
6	17 – 20	5

проверить при уровне значимости α = 0,05 гипотезу о равномерном законе распределения генеральной совокупности с помощью критерия Пирсона. В ответе указать наблюдаемое значение критерия и сделать вывод о принятии или отклонении гипотезы.

Указание

Для проверки гипотезы о равномерном распределении генеральной совокупности концы интервала, в котором наблюдались возможные значения X, оцениваются по формулам:

$$a^* = \overline{x}_B - \sqrt{3}\cdot\sigma_B; \quad b^* = \overline{x}_B + \sqrt{3}\cdot\sigma_B.$$

Тогда плотность вероятности

$$f(x) = \frac{1}{b^* - a^*}; \quad n'_1 = \frac{n(x_1 - a^*)}{b^* - a^*};$$

$$n'_2 = n'_3 = \ldots = n'_{s-1} = \frac{n(x_i - x_{i-1})}{b^* - a^*}; \; i = 2,3,\ldots,s-1,$$

$$n'_s = \frac{n(b^* - x_{s-1})}{b^* - a^*}.$$

Число степеней свободы k = n – 3, так как равномерное распределение оценивается двумя параметрами.

Критическая область выбирается правосторонней, и граница ее при заданном уровне значимости α $\chi^2_{кр}(\alpha, k)$ находится по таблице критических точек распределения χ^2. Здесь k = s – 3, где s – число частичных интервалов выборки.

Решение

Объем выборки n = 70. Будем считать вариантами середины частичных интервалов: x_1 = 3,5, x_2 = 6,5,…, x_6 = 18,5.

Найдем $\overline{x}_B$ = 11,43; σ_B = 4,03; s = 4,05.

Для равномерного распределения

$$a^* = 11,43 - \sqrt{3}\cdot 4,03 = 4,45;$$

$$b^* = 11,43 + \sqrt{3}\cdot 4,03 = 18,41.$$

$$f(x) = \frac{1}{18,41 - 4,45} = 0,072;$$

теоретические частоты:

$$n'_1 = 70\cdot(5 - 4,45)\cdot 0,072 = 2,77;$$

$$n'_2 = n'_3 = n'_4 = n'_5 = 70\cdot 3\cdot 0,072 =$$

$$= 15,12; \quad n'_6 = 70\cdot(18,41 - 17)\cdot 0,072 = 7,1.$$

Наблюдаемое значение критерия

$$\chi^2_{набл} = \frac{(6 - 2,77)^2}{2,77} + \ldots + \frac{(5 - 7,1)^2}{7,1} = 10,95.$$

Критическая точка

$$\chi^2(0{,}05;3)=7{,}8; \quad \chi^2_{набл} > \chi^2_{кр},$$

и гипотеза о равномерном распределении отклоняется.

Ответ: $\chi^2_{набл}$ = 10,95, гипотеза отклоняется.

Задача 3.

Для выборки, интервальный статистический ряд которой имеет вид

Номер интервала	Границы интервала	Эмпирические частоты
1	2 – 5	6
2	5 – 8	8
3	8 – 11	15
4	11 – 14	22
5	14 – 17	14
6	17 – 20	5

проверить при уровне значимости α = 0,05 гипотезу о нормальном законе распределения генеральной совокупности с помощью критерия Пирсона. В ответе указать наблюдаемое значение критерия и сделать вывод о принятии или отклонении гипотезы.

Указание

Для проверки гипотезы о нормальном законе распределения

$$n_i' = n \cdot P_i,$$

где n – объем выборки,

$$P_i = \Phi\left(\frac{x_{i+1}-\overline{x}_B}{s}\right) - \Phi\left(\frac{x_i-\overline{x}_B}{s}\right),$$

x_i и x_{i+1} – левая и правая границы i-го интервала, $\overline{x}_B$ – выборочное среднее, s – исправленное среднее квадратическое отклонение. Поскольку нормальное распределение характеризуется двумя параметрами, число степеней свободы $k = n - 3$.

Критическая область выбирается правосторонней, и граница ее при заданном уровне значимости α $\chi^2_{кр}(\alpha, k)$ находится по таблице критических точек распределения χ^2. Здесь $k = s - 3$, где s – число частичных интервалов выборки.

Решение

Объем выборки n = 70. Будем считать вариантами середины частичных интервалов: x_1 = 3,5, x_2 = 6,5,…, x_6 = 18,5.

Найдем $\overline{x}_B$ = 11,43; σ_B = 4,03; s = 4,05.

Теоретические частоты для нормального распределения:

$$n_1' = 70 \cdot \left(\Phi\left(\frac{5-11,43}{4,05} \right) - \Phi\left(\frac{2-11,43}{4,05} \right) \right) =$$
$$= 70 \cdot \left(\Phi(-1,588) - \Phi(-2,328) \right) =$$
$$= 70 \cdot \left(\Phi(2,328) - \Phi(1,588) \right) = 70 \cdot (0,4900 - 0,4441) = 3,2.$$

Так же вычисляются
$$n_2' = 9,9; \quad n_3' = 18,2; \quad n_4' = 19,6; \quad n_5' = 12,5; \quad n_6' = 4,7.$$
Наблюдаемое значение критерия
$$\chi^2_{набл} = \frac{(6-3,2)^2}{3,2} + \ldots + \frac{(5-4,7)^2}{4,7} = 3,87.$$
Критическая точка
$$\chi^2(0,05;3) = 7,8.$$
Поскольку $\chi^2_{набл} < \chi^2_{кр}$, гипотеза о нормальном распределении генеральной совокупности принимается.

Ответ: $\chi^2_{набл}$ = 3,87, гипотеза принимается.

Задача 4.

Выборка из 1000 значений случайной величины X имеет вид:

x_i	98	98,5	99	99,5	100	100,5	101	101,5	102	102,5
p_i	21	47	87	158	181	201	142	97	41	25

Проверить, пользуясь критерием Колмогорова, гипотезу о том, что X распределена по нормальному закону с параметрами a = 100,25 и σ = 1, при уровне значимости α = 0,05. В ответе привести значение λ_q и сделать вывод о принятии или отклонении гипотезы.

Указание

Теоретическая функция распределения $F(x)$ определяется формулой
$$F(x) = \frac{1}{2} + \frac{1}{2}\Phi(x-a).$$
Эмпирическая функция
$$F^*(x_k) = 0,001\left(\sum_{i=1}^{k-1} n_i + 0,5 n_k \right).$$
Найдите максимум модуля разности
$$F^*(x_k) - F(x_k) = D_q$$
и сравните наблюдаемое значение
$$\lambda_q = \sqrt{n} D_q$$
с критическим (для α = 0,05 λ_α = 1,224).

Решение

Теоретическая функция распределения $F(x)$ определяется формулой

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

Эмпирическая функция

$$F^*(x_k)=0{,}001\left(\sum_{i=1}^{k-1}n_i+0{,}5n_k\right).$$

Для вычисления

$$D_q=\max\left|F^*(x_k)-F(x_k)\right|$$

проведем расчеты, результаты которых сведены в таблицу:

$x_k - a$	$F(x_k)$	$F^*(x_k)$	$\mid F^*(x_k) - F(x_k)\mid$
-2,25	0,0123	0,0105	0,0018
-1,75	0,0401	0,0445	0,0044
-1,25	0,1056	0,1115	0,0059
-0,75	0,2266	0,2340	0,0074
-0,25	0,4013	0,4035	0,0022
0,25	0,5987	0,5945	0,0042
0,75	0,7734	0,7660	0,0074
1,25	0,8944	0,8855	0,0089
1,75	0,9599	0,9545	0,0054
2,25	0,9877	0,9875	0,0002

Таким образом,

$$D_q=0{,}0089,\quad \lambda_q=\sqrt{1000}\cdot 0{,}0089=0{,}281,$$

$$\lambda_q<\lambda_\alpha\quad -$$

гипотеза о нормальном распределении принимается.

Ответ: λ_q = 0,281, гипотеза принимается.

2.2.3. *Корреляционный анализ. Регрессионный анализ. Однофакторный дисперсионный анализ*

Проверка гипотезы о значимости выборочного коэффициента корреляции

Рассмотрим выборку объема n, извлеченную из нормально распределенной двумерной генеральной совокупности (X, Y). Вычислим выборочный коэффициент корреляции r_B. Пусть он оказался не равным нулю. Это еще не означает, что и коэффициент корреляции генеральной совокупности не равен нулю. Поэтому при заданном уровне значимости α возникает необходимость проверки нулевой гипотезы H_0: $r_г = 0$ о равенстве нулю генерального коэффициента корреляции при конкурирующей гипотезе H_1: $r_г \neq 0$. Таким образом, при принятии нулевой гипотезы X и Y некоррелированы, то есть не связаны линейной зависимостью, а при отклонении H_0 они коррелированы.

В качестве критерия примем случайную величину

$$\boxed{T = \frac{r_B\sqrt{n-2}}{\sqrt{1-r_B^2}},}$$

которая при справедливости нулевой гипотезы имеет распределение Стьюдента с $k = n - 2$ степенями свободы. Из вида конкурирующей гипотезы следует, что критическая область двусторонняя с границами $\pm t_{кр}$, где значение $t_{кр}(\alpha,\ k)$ находится из таблиц для двусторонней критической области.

Вычислив наблюдаемое значение критерия

$$T_{набл} = \frac{r_B\sqrt{n-2}}{\sqrt{1-r_B^2}}$$

и сравнив его с $t_{кр}$, делаем вывод:

- если $|T_{набл}| < t_{кр}$ – нулевая гипотеза принимается (корреляции нет);
- если $|T_{набл}| > t_{кр}$ – нулевая гипотеза отвергается (корреляция есть).

Ранговая корреляция

Пусть объекты генеральной совокупности обладают двумя качественными признаками (то есть признаками, которые невозможно измерить точно, но которые позволяют сравнивать объекты между собой и располагать их в порядке убывания или возрастания качества). Договоримся для определенности располагать объекты в порядке ухудшения качества.

Пусть выборка объема n содержит независимые объекты, обладающие двумя качественными признаками: A и B. Требуется выяснить степень их связи между собой, то есть установить наличие или отсутствие **ранговой корреляции**.

Расположим объекты выборки в порядке ухудшения качества по признаку A, предполагая, что все они имеют различное качество по обоим признакам.

Назовем место, занимаемое в этом ряду некоторым объектом, его **рангом x_i:** $x_1 = 1, x_2 = 2,\ldots, x_n = n$.

Теперь расположим объекты в порядке ухудшения качества по признаку *B*, присвоив им ранги y_i , где номер i равен порядковому номеру объекта по признаку *A*, а само значение ранга равно порядковому номеру объекта по признаку *B*. Таким образом, получены две последовательности рангов:

по признаку *A* … $x_1, x_2,\ldots, x_n$

по признаку *B* … $y_1, y_2,\ldots, y_n$.

При этом, если, например, $y_3 = 6$, то это означает, что данный объект занимает в ряду по признаку *A* третье место, а в ряду по признаку *B* – шестое. Сравним полученные последовательности рангов.

1. Если $x_i = y_i$ при всех значениях i, то ухудшение качества по признаку *A* влечет за собой ухудшение качества по признаку *B*, то есть имеется «полная ранговая зависимость».

2. Если ранги противоположны, то есть $x_1 = 1$, $y_1 = n$; $x_2 = 2$, $y_2 = n - 1$;…, $x_n = n$, $y_n = 1$, то признаки тоже связаны: ухудшение качества по одному из них приводит к улучшению качества по другому («противоположная зависимость»).

3. На практике чаще всего встречается промежуточный случай, когда ряд y_i не монотонен. Для оценки связи между признаками будем считать ранги $x_1, x_2,\ldots, x_n$ возможными значениями случайной величины *X*, а $y_1, y_2,\ldots, y_n$ – возможными значениями случайной величины *Y*. Теперь можно исследовать связь между *X* и *Y*, вычислив для них выборочный коэффициент корреляции

$$r_B = \frac{\sum n_{uv} uv - n\overline{u}\overline{v}}{n\sigma_u \sigma_v},$$

где

$$u_i = x_i - \overline{x}, \quad v_i = y_i - \overline{y}$$

(условные варианты). Поскольку каждому рангу x_i соответствует только одно значение y_i, то частота любой пары условных вариант с одинаковыми индексами равна 1, а с разными индексами – нулю. Кроме того, из выбора условных вариант следует, что $\overline{u} = \overline{v} = 0$, поэтому формула для выборочного коэффициента корреляции приобретает более простой вид:

$$r_B = \frac{\sum u_i v_i}{n\sigma_u \sigma_v}. \qquad (1)$$

Итак, требуется найти $\sum u_i v_i, \quad \sigma_u \quad и \quad \sigma_v$.

Можно показать, что

$$\sum u_i^2 = \sum v_i^2 = \frac{n^3 - n}{12}.$$

Учитывая, что $\overline{x} = \overline{y}$, можно выразить $\sum u_i v_i$ через разности рангов

$$d_i = x_i - y_i = u_i - v_i.$$

После преобразований получим:

$$\sum u_i v_i = \frac{n^3 - n}{12} - \sum \frac{d_i^2}{2}, \quad \sigma_u = \sigma_v = \sqrt{\frac{n^2 - 1}{12}},$$

откуда

$$n\sigma_u \sigma_v = \frac{n^3 - n}{12}.$$

Подставив эти результаты в (1), получим **выборочный коэффициент ранговой корреляции Спирмена**:

$$\boxed{\rho_B = 1 - \frac{6\sum d_i^2}{n^3 - n}.}$$

Свойства выборочного коэффициента корреляции Спирмена

1. Если между A и B имеется «полная прямая зависимость», то есть ранги совпадают при всех i, то $\rho_B = 1$. Действительно, при этом $d_i = 0$, и из формулы для коэффициента Спирмена следует справедливость свойства 1.
2. Если между A и B имеется «противоположная зависимость», то $\rho_B = -1$. В этом случае, преобразуя $d_i = (2i - 1) - n$, найдем, что

$$\sum d_i^2 = \frac{n^3 - n}{3},$$

тогда из формулы для коэффициента Спирмена получаем, что

$$\rho_B = 1 - \frac{6(n^3 - n)}{3(n^3 - n)} = 1 - 2 = -1.$$

3. В остальных случаях $-1 < \rho_B < 1$, причем зависимость между A и B тем меньше, чем ближе $|\rho_B|$ к нулю.

Итак, требуется при заданном уровне значимости α проверить нулевую гипотезу о равенстве нулю генерального коэффициента ранговой корреляции Спирмена $\rho_Г$ при конкурирующей гипотезе

$$H_1 : \rho_Г \neq 0.$$

Для этого найдем критическую точку:

$$T_{kp} = t_{kp}(\alpha, k)\sqrt{\frac{1 - \rho_B^2}{n - 2}},$$

где n – объем выборки, ρ_B – выборочный коэффициент ранговой корреляции Спирмена, $t_{кр}(\alpha, k)$ – критическая точка двусторонней критической области, найденная по таблице критических точек распределения Стьюдента, число степеней свободы $k = n - 2$.

Тогда, если $|\rho_B| < T_{кр}$, то нулевая гипотеза принимается, то есть ранговая корреляционная связь между признаками незначима.

Если $|\rho_B| > T_{кр}$, то нулевая гипотеза отвергается, и между признаками существует значимая ранговая корреляционная связь.

Можно использовать и другой коэффициент – коэффициент ранговой корреляции Кендалла. Рассмотрим ряд рангов $y_1, y_2, \ldots, y_n$, введенный так же, как и ранее, и зададим величины R_i следующим образом: пусть правее y_1 имеется R_1 рангов, больших y_1; правее y_2 – R_2 рангов, больших y_2 и т.д. Тогда, если обозначить $R = R_1 + R_2 + \ldots + R_{n-1}$, то **выборочный коэффициент ранговой корреляции Кендалла** определяется формулой

$$\boxed{\tau_B = \frac{4R}{n(n-1)} - 1,}$$

где n – объем выборки.

Замечание. Легко убедиться, что коэффициент Кендалла обладает теми же свойствами, что и коэффициент Спирмена.

Для проверки нулевой гипотезы H_0: $\tau_Г = 0$ (генеральный коэффициент ранговой корреляции Кендалла равен нулю) при альтернативной гипотезе

$$H_1 : \tau_Г \neq 0$$

необходимо найти критическую точку:

$$T_{кр} = z_{kp}\sqrt{\frac{2(2n+5)}{9n(n-1)}},$$

где n – объем выборки, а $z_{кр}$ – критическая точка двусторонней критической области, определяемая из условия

$$\Phi(z_{kp}) = \frac{1-\alpha}{2}$$

по таблицам для функции Лапласа.

Если $|\tau_В| < T_{кр}$, то нулевая гипотеза принимается (ранговая корреляционная связь между признаками незначима).

Если $|\tau_В| > T_{кр}$, то нулевая гипотеза отвергается (между признаками существует значимая ранговая корреляционная связь).

Регрессионный анализ

Рассмотрим выборку двумерной случайной величины (X, Y). Примем в качестве оценок условных математических ожиданий компонент их условные средние значения, а именно: **условным средним** $\bar{y}_x$ назовем среднее арифметическое наблюдавшихся значений Y, соответствующих $X = x$. Аналогично **условное среднее** $\bar{x}_y$ – среднее арифметическое наблюдавшихся значений X, соответствующих $Y = y$. Ранее были выведены уравнения регрессии Y на X и X на Y:

$$M(Y|x) = f(x), \quad M(X|y) = \varphi(y).$$

Условные средние $\bar{y}_x$ и $\bar{x}_y$ являются оценками условных математических ожиданий и, следовательно, тоже функциями от x и y, то есть

$$\bar{y}_x = f^*(x) \quad -$$

выборочное уравнение регрессии *Y* на *X*,

$$\overline{x}_y = \varphi^*(y) \quad -$$

выборочное уравнение регрессии *X* на *Y*.

Соответственно функции $f^*(x)$ и $\varphi^*(y)$ называются **выборочной регрессией *Y* на *X* и *X* на *Y***, а их графики – **выборочными линиями регрессии**. Выясним, как определять параметры выборочных уравнений регрессии, если сам вид этих уравнений известен.

Пусть изучается двумерная случайная величина (X, Y), и получена выборка из n пар чисел $(x_1, y_1), (x_2, y_2),\ldots, (x_n, y_n)$. Будем искать параметры прямой линии среднеквадратической регрессии Y на X вида

$$Y = \rho_{xy}x + b, \qquad (2)$$

$$Y = \rho_{yx}x + b\ ,$$

подбирая параметры ρ_{yx} и b так, чтобы точки на плоскости с координатами $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ лежали как можно ближе к прямой (2). Используем для этого метод наименьших квадратов и найдем минимум функции

$$F(\rho, b) = \sum_{i=1}^{n}(Y_i - y_i)^2 = \sum_{i=1}^{n}(\rho x_i + b - y_i)^2.$$

Приравняем нулю соответствующие частные производные:

$$\begin{aligned} \frac{\partial F}{\partial \rho} &= 2\sum_{i=1}^{n}(\rho x_i + b - y_i)x_i = 0 \\ \frac{\partial F}{\partial b} &= 2\sum_{i=1}^{n}(\rho x_i + b - y_i) = 0 \end{aligned}.$$

В результате получим систему двух линейных уравнений относительно ρ и b:

$$\begin{cases} \left(\sum x^2\right)\rho + \left(\sum x\right)b = \sum xy \\ \left(\sum x\right)\rho + nb = \sum y \end{cases}. \qquad (3)$$

Ее решение позволяет найти искомые параметры в виде:

$$\rho_{xy} = \frac{n\sum xy - \sum x \cdot \sum y}{n\sum x^2 - \left(\sum x\right)^2};$$

$$b = \frac{\sum x^2 \cdot \sum y - \sum x \cdot \sum xy}{n\sum x^2 - \left(\sum x\right)^2}.$$

При этом предполагалось, что все значения X и Y наблюдались по одному разу.

Теперь рассмотрим случай, когда имеется достаточно большая выборка (не менее 50 значений), и данные сгруппированы в виде *корреляционной таблицы*:

Y	X				
	x_1	x_2	…	x_k	n_y
y_1	n_{11}	n_{21}	…	n_{k1}	$n_{11}+n_{21}+\ldots+n_{k1}$
y_2	n_{12}	n_{22}	…	n_{k2}	$n_{12}+n_{22}+\ldots+n_{k2}$
…	…	…	…	…	……………..
y_m	n_{1m}	n_{2m}	…	n_{km}	$n_{1m}+n_{2m}+\ldots+n_{km}$
n_x	$n_{11}+\ldots+n_{1m}$	$n_{21}+\ldots+n_{2m}$	…	$n_{k1}+\ldots+n_{km}$	$n=\sum n_x = \sum n_y$

Здесь n_{ij} – число появлений в выборке пары чисел (x_i, y_j).
Поскольку

$$\bar{x}=\frac{\sum x}{n},\quad \bar{y}=\frac{\sum y}{n},\quad \overline{x^2}=\frac{\sum x^2}{n},$$

заменим в системе (3)

$$\sum x = n\bar{x},\quad \sum y = n\bar{y},$$

$$\sum x^2 = n\overline{x^2},\quad \sum xy = \sum n_{xy}xy,$$

где n_{xy} – число появлений пары чисел (x, y). Тогда система (3) примет вид:

$$\begin{cases}(n\overline{x^2})\rho_{yx}+(n\bar{x})b=\sum n_{xy}xy\\ (\bar{x})\rho_{yx}+b=\bar{y}\end{cases}.\quad (4)$$

Можно решить эту систему и найти параметры ρ_{yx} и b, определяющие выборочное уравнение прямой линии регрессии:

$$\bar{y}_x=\rho_{yx}\bar{x}+b.$$

Но чаще уравнение регрессии записывают в ином виде, вводя **выборочный коэффициент корреляции**. Выразим b из второго уравнения системы (4):

$$b=\bar{y}-\rho_{yx}\bar{x}.$$

Подставим это выражение в уравнение регрессии:

$$\bar{y}_x-\bar{y}=\rho_{yx}(x-\bar{x}).$$

Из (4)

$$\rho_{yx}=\frac{\sum n_{xy}xy-n\bar{x}\bar{y}}{n(\overline{x^2}-(\bar{x})^2)}=\frac{\sum n_{xy}xy-n\bar{x}\bar{y}}{n\tilde{\sigma}_x^2},\quad (5)$$

где

$$\tilde{\sigma}_x^2=\overline{x^2}-(\bar{x})^2.$$

Введем понятие **выборочного коэффициента корреляции**

$$r_B=\frac{\sum n_{xy}xy-n\bar{x}\bar{y}}{n\tilde{\sigma}_x\tilde{\sigma}_y}$$

и умножим равенство (5) на $\dfrac{\tilde{\sigma}_x}{\tilde{\sigma}_y}$:

$$\rho_{yx}\frac{\tilde{\sigma}_x}{\tilde{\sigma}_y}=r_B, \quad откуда \quad \rho_{yx}=r_B\frac{\tilde{\sigma}_y}{\tilde{\sigma}_x}.$$

Используя это соотношение, получим выборочное уравнение прямой линии регрессии Y на X вида

$$\overline{y}_x-\overline{y}=r_B\frac{\tilde{\sigma}_y}{\tilde{\sigma}_x}(x-\overline{x}).$$

Однофакторный дисперсионный анализ

Пусть генеральные совокупности $X_1, X_2,\ldots, X_p$ распределены нормально и имеют одинаковую дисперсию, значение которой неизвестно. Найдем выборочные средние по выборкам из этих генеральных совокупностей и проверим при заданном уровне значимости нулевую гипотезу H_0: $M(X_1) = M(X_2) = \ldots = M(X_p)$ о равенстве всех математических ожиданий. Для решения этой задачи применяется метод, основанный на сравнении дисперсий и названный поэтому **дисперсионным анализом**.

Будем считать, что на случайную величину X воздействует некоторый качественный фактор F, имеющий p уровней: F_1, F_2, …, F_p. Требуется сравнить «факторную дисперсию», то есть рассеяние, порождаемое изменением уровня фактора, и «остаточную дисперсию», обусловленную случайными причинами. Если их различие значимо, то фактор существенно влияет на X и при изменении его уровня групповые средние различаются значимо.

Будем считать, что количество наблюдений на каждом уровне фактора одинаково и равно q. Оформим результаты наблюдений в виде таблицы:

Номер испытания	Уровни фактора F_j			
	F_1	F_2	…	F_p
1	x_{11}	x_{12}	…	x_{1p}
2	x_{21}	x_{22}	…	x_{2p}
…	…	…	…	…
q	x_{q1}	x_{q2}	…	x_{qp}
Групповое среднее	$\overline{x}_{гр1}$	$\overline{x}_{гр2}$	…	$\overline{x}_{грp}$

Определим общую, факторную и остаточную суммы квадратов отклонений от среднего:

$$S_{общ}=\sum_{j=1}^{p}\sum_{i=1}^{q}(x_{ij}-\overline{x})^2 \quad -$$

общая сумма квадратов отклонений наблюдаемых значений от общего среднего $\overline{x}$;

$$S_{факт} = q\sum_{j=1}^{p}(\overline{x}_{грj} - \overline{x})^2 \quad -$$

факторная сумма отклонений групповых средних от общей средней, характеризующая рассеяние между группами;

$$S_{ост} = \sum_{i=1}^{q}(x_{i1} - \overline{x}_{гр1})^2 + \sum_{i=1}^{q}(x_{i2} - \overline{x}_{гр2})^2 + \ldots + \sum_{i=1}^{q}(x_{ip} - \overline{x}_{грp})^2 \quad -$$

остаточная сумма квадратов отклонений наблюдаемых значений группы от своего группового среднего, характеризующая рассеяние внутри групп.

Замечание. Остаточную сумму можно найти из равенства

$$S_{ост} = S_{общ} - S_{факт}.$$

Вводя обозначения

$$R_j = \sum_{i=1}^{q} x_{ij}, \quad P_j = \sum_{i=1}^{q} x_{ij}^2,$$

получим формулы, более удобные для расчетов:

$$S_{общ} = \sum_{j=1}^{p} P_j - \frac{\left(\sum_{j=1}^{p} R_j\right)^2}{pq},$$

$$S_{факт} = \frac{\sum_{j=1}^{p} R_j^2}{q} - \frac{\left(\sum_{j=1}^{p} R_j\right)^2}{pq}.$$

Разделив суммы квадратов на соответствующее число степеней свободы, получим общую, факторную и остаточную дисперсии:

$$s_{общ}^2 = \frac{S_{общ}}{pq-1}, \quad s_{факт}^2 = \frac{S_{факт}}{p-1}, \quad s_{ост}^2 = \frac{S_{ост}}{p(q-1)}.$$

Если справедлива гипотеза H_0, то все эти дисперсии являются несмещенными оценками генеральной дисперсии. Покажем, что проверка нулевой гипотезы сводится к сравнению факторной и остаточной дисперсии по критерию Фишера-Снедекора.

1. Пусть гипотеза H_0 правильна. Тогда факторная и остаточная дисперсии являются несмещенными оценками неизвестной генеральной дисперсии и, следовательно, различаются незначимо. Поэтому результат оценки по критерию Фишера-Снедекора F покажет, что нулевая гипотеза принимается. Таким образом, если верна гипотеза о равенстве математических ожиданий генеральных совокупностей, то верна и гипотеза о равенстве факторной и остаточной дисперсий.

2. Если нулевая гипотеза неверна, то с возрастанием расхождения между математическими ожиданиями увеличивается и факторная дисперсия, а вместе с ней и отношение

$$F_{набл} = \frac{s_{факт}^2}{s_{ост}^2}.$$

Поэтому в результате $F_{\text{набл}}$ окажется больше $F_{\text{кр}}$, и гипотеза о равенстве дисперсий будет отвергнута. Следовательно, если гипотеза о равенстве математических ожиданий генеральных совокупностей ложна, то ложна и гипотеза о равенстве факторной и остаточной дисперсий.

Итак, метод дисперсионного анализа состоит в *проверке по критерию F нулевой гипотезы о равенстве факторной и остаточной дисперсий.*

Замечание. Если факторная дисперсия окажется меньше остаточной, то гипотеза о равенстве математических ожиданий генеральных совокупностей верна. При этом нет необходимости использовать критерий F.

Если число испытаний на разных уровнях различно (q_1 испытаний на уровне F_1, q_2 – на уровне F_2, …, q_p - на уровне F_p), то

$$S_{общ} = (P_1 + P_2 + ... + P_p) - (R_1 + R_2 + ... + R_p),$$

где

$$P_j = \sum_{i=1}^{q_j} x_{ij}^2 -$$

сумма квадратов наблюдавшихся значений признака на уровне F_j,

$$R_j = \sum_{i=1}^{q_j} x_{ij} -$$

сумма наблюдавшихся значений признака на уровне F_j . При этом объем выборки, или общее число испытаний, равен

$$n = q_1 + q_2 + ... + q_p.$$

Факторная сумма квадратов отклонений вычисляется по формуле

$$S_{факт} = \left(\frac{R_1^2}{q_1} + \frac{R_2^2}{q_2} + ... + \frac{R_p^2}{q_p} \right) - \frac{(R_1 + R_2 + ... + R_p)^2}{n}.$$

Остальные вычисления проводятся так же, как в случае одинакового числа испытаний:

$$S_{ост} = S_{общ} - S_{факт}, \quad s_{факт}^2 = \frac{S_{факт}}{p-1}, \quad s_{ост}^2 = \frac{S_{ост}}{n-p}.$$

ПРИМЕРЫ РЕШЕНИЯ ЗАДАЧ

Задача 1.

Для выборки двумерной случайной величины

i	1	2	3	4	5	6	7	8	9	10
x_i	1,2	1,5	1,8	2,1	2,3	3,0	3,6	4,2	5,7	6,3
y_i	5,6	6,8	7,8	9,4	10,3	11,4	12,9	14,8	15,2	18,5

вычислить выборочный коэффициент корреляции.

Указание

Найдите для X и Y выборочные средние и выборочные средние квадратические отклонения и воспользуйтесь формулой

$$r_B = \frac{\sum x_i y_i - n\overline{x}_B \overline{y}_B}{n\sigma_x \sigma_y}.$$

Решение

$$\overline{x}_B = \frac{1,2 + 1,5 + ... + 6,3}{10} = 3,17;$$

$$\overline{y}_B = \frac{5,6 + 6,8 + .. + 18,5}{10} = 11,27.$$

$$D_B(X) = 0,1(1,2^2 + 1,5^2 + ... + 6,3^2) - 3,17^2 = 2,7921;$$

$$\sigma_X = \sqrt{2,7921} = 1,671.$$

$$D_B(Y) = 0,1(5,6^2 + 6,8^2 + ... + 18,5^2) - 11,27^2 = 15,146;$$

$$\sigma_Y = \sqrt{15,146} = 3,892.$$

Для определения выборочного коэффициента корреляции вычислим предварительно

$$\sum_{i=1}^{10} x_i y_i = 1,2 \cdot 5,6 + 1,5 \cdot 6,8 + ... + 6,3 \cdot 18,5 = 420,38.$$

Тогда

$$r_B = \frac{\sum x_i y_i - n\overline{x}_B \overline{y}_B}{n\sigma_x \sigma_y} = \frac{420,38 - 10 \cdot 3,17 \cdot 11,27}{10 \cdot 1,671 \cdot 3,892} = 0,97.$$

Ответ: 0,97.

Задача 2.

По выборке объема n = 150, извлеченной из нормально распределенной двумерной генеральной совокупности, вычислен выборочный коэффициент корреляции r_B = - 0,37. Проверить при уровне значимости α = 0,01 нулевую гипотезу H_o: $r_Г$ = 0 о равенстве нулю генерального коэффициента корреляции при конкурирующей гипотезе H_1: $r_Г \neq 0$. В ответе указать наблюдаемое значение критерия $T_{набл}$ и сделать вывод о принятии или отклонении гипотезы.

Указание

Критерием является случайная величина

$$T = \frac{r_B \sqrt{n-2}}{\sqrt{1 - r_B^2}},$$

имеющая при справедливости нулевой гипотезы распределение Стьюдента с $k = n - 2$ степенями свободы. Критическая область при заданном виде конкурирующей гипотезы является двусторонней и задается неравенством $|T| > t_{кр}$, где $t_{кр}(\alpha,\ k)$ находится по таблице критических точек распределения Стьюдента.

Решение

Критическая точка

$$t_{кр}(0{,}01;\ 150) = 2{,}58.$$

Вычислим наблюдаемое значение критерия:

$$T_{набл} = -\frac{0{,}37\sqrt{148}}{\sqrt{1-0{,}37^2}} = -4{,}85.$$

Поскольку $|T_{набл}| > t_{кр}$, нулевая гипотеза отвергается, то есть X и Y коррелированы.

Ответ: $T_{набл}$ = -4,85, гипотеза отклоняется.

Задача 3.

Десять школьников сдавали выпускной экзамен ЕГЭ по математике и вступительный экзамен по системе централизованного тестирования. Результаты обоих экзаменов оценивались по 100-балльной шкале и оказались следующими (1-я строка – оценки ЕГЭ, вторая – централизованного тестирования):

87	82	80	79	63	55	40	34	33	29
57	92	80	69	71	43	49	51	20	19

Найти выборочный коэффициент корреляции Спирмена.

Указание

Воспользуйтесь формулой

$$\rho_B = 1 - \frac{6\sum d_i^2}{n^3 - n},$$

где $d_i = x_i - y_i$, n – объем выборки.

Решение

Составим последовательности рангов по убыванию баллов на каждом экзамене:

x_i	1	2	3	4	5	6	7	8	9	10
y_i	5	1	2	4	3	8	7	6	9	10 .

Вычислим d_i: $d_1 = 1 - 5 = -4$; $d_2 = 2 - 1 = 1$; $d_3 = 3 - 2 = 1$; $d_4 = 4 - 4 = 0$; $d_5 = 5 - 3 = 2$; $d_6 = 6 - 8 = -2$; $d_7 = 7 - 7 = 0$; $d_8 = 8 - 6 = 2$; $d_9 = d_{10} = 0$.

Найдем

$$\sum d_i^2 = 16 + 1 + 1 + 4 + 4 + 4 = 30.$$

Тогда выборочный коэффициент ранговой корреляции Спирмена

$$\rho_B = 1 - \frac{6 \cdot 30}{1000 - 10} = 0{,}818.$$

Ответ: 0,818.

Задача 4.
Десять школьников сдавали выпускной экзамен ЕГЭ по математике и вступительный экзамен по системе централизованного тестирования. Результаты обоих экзаменов оценивались по 100-балльной шкале и оказались следующими (1-я строка – оценки ЕГЭ, вторая – централизованного тестирования):

87	82	80	79	63	55	40	34	33	29
57	92	80	69	71	43	49	51	20	19

Найти выборочный коэффициент корреляции Кендалла.

Указание

Воспользуйтесь формулой

$$\tau_B = \frac{4R}{n(n-1)} - 1,$$

где $R = R_1 + R_2 + \ldots + R_n$ (величины R_1, R_2, ..., R_n, где R_i – количество чисел, больших y_i, стоящих справа от y_i в последовательности рангов по признаку B).

Решение

Определим, сколько рангов, больших данного, располагается справа от каждого y_i:
$R_1 = 5$; $R_2 = 8$; $R_3 = 7$; $R_4 = 5$; $R_5 = 5$; $R_6 = 2$; $R_7 = 2$; $R_8 = 2$; $R_9 = 1$; $R_{10} = 0$;
$R = 5 + 8 + 7 + 5 + 5 + 2 + 2 + 2 + 1 = 37$;

$$\tau_B = \frac{4 \cdot 37}{10 \cdot 9} - 1 = 0{,}644.$$

Ответ: 0,644.

Задача 5.
Для выборки двумерной случайной величины

i	1	2	3	4	5	6	7	8	9	10
x_i	1,2	1,5	1,8	2,1	2,3	3,0	3,6	4,2	5,7	6,3
y_i	5,6	6,8	7,8	9,4	10,3	11,4	12,9	14,8	15,2	18,5

составить выборочное уравнение прямой линии регрессии Y на X.

Указание

Выборочное уравнение прямой линии регрессии Y на X имеет вид:

$$y - \overline{y}_B = r_B \frac{\sigma_y}{\sigma_x}(x - \overline{x}_B).$$

Можно воспользоваться тем, что параметры этой выборки уже вычислены в задаче 1.

Решение

$$\overline{x}_B = 3,17; \quad \overline{y}_B = 11,27; \quad D_B(X) = 2,7921;$$

$$\sigma_X = \sqrt{2,7921} = 1,671;$$

$$D_B(Y) = 15,146; \quad \sigma_Y = \sqrt{15,146} = 3,892;$$

$$r_B = \frac{420,38 - 10 \cdot 3,17 \cdot 11,27}{10 \cdot 1,671 \cdot 3,892} = 0,97.$$

Выборочное уравнение прямой линии регрессии Y на X имеет вид:

$$y - 11,27 = 0,97 \cdot \frac{3,892}{1,671}(x - 3,17),$$

или

$$y = 2,26x - 4,104.$$

Ответ: $y = 2,26x - 4,104.$

Задача 6.

Произведено по 4 испытания на каждом из трех уровней фактора F. Результаты испытаний приведены в таблице:

Номер испытания	Уровни фактора		
i	F_1	F_2	F_3
1	38	20	21
2	36	24	22
3	35	26	31
4	31	30	34
$\overline{x}_{гр\ j}$	35	25	27

Методом дисперсионного анализа при уровне значимости 0,05 проверить нулевую гипотезу о равенстве генеральных средних (математических ожиданий генеральных совокупностей), предполагая, что генеральные дисперсии равны. В ответе указать значение $F_{набл}$ и сделать вывод о принятии или отклонении гипотезы.

Указание

Найдите выборочное среднее для всей выборки и перейдите к уменьшенным вариантам

$$y_{ij} = x_{ij} - \overline{x}.$$

Наблюдаемое значение критерия вычисляется по формуле

$$F_{набл} = \frac{s^2_{факт}}{s^2_{ост}}.$$

Решение

Выборочное среднее $\bar{x} = 29$. Перейдем к уменьшенным вариантам: $y_{ij} = x_{ij} - 29$ и внесем результаты вспомогательных расчетов в таблицу:

Номер испытания	Уровни фактора					
i	F_1		F_2		F_3	
	y_{i1}	y_{i1}^2	y_{i2}	y_{i2}^2	y_{i3}	y_{i3}^2
1	9	81	-9	81	-8	64
2	7	49	-5	25	-7	49
3	6	36	-3	9	2	4
4	2	4	1	1	5	25

При этом

$$T_1 = \sum_{i=1}^{3} y_{i1} = 24, \quad T_2 = \sum_{i=1}^{3} y_{i2} = -16, \quad T_3 = \sum_{i=1}^{3} y_{i3} = -8, \quad \sum T_j = 0;$$

$$S_1 = \sum_{i=1}^{3} y^2{}_{i1} = 170, \quad S_2 = \sum_{i=1}^{3} y^2{}_{i2} = 116, \quad S_3 = \sum_{i=1}^{3} y^2{}_{i3} = 142;$$

$$\sum S_j = 428, \quad \sum T_j^2 = 896.$$

$$S_{общ} = \sum S_j - \frac{\left(\sum T_j\right)^2}{3 \cdot 4} = 428 - 0 = 428,$$

$$S_{факт} = \frac{\sum T_j^2}{4} - \frac{\left(\sum T_j\right)^2}{3 \cdot 4} = 224, \quad S_{ост} = 428 - 224 = 204.$$

Найдем факторную и остаточную дисперсии, учитывая, что число степеней свободы для факторной дисперсии равно $p - 1 = 3 - 1 = 2$, а для остаточной $p(q - 1) = 3(4 - 1) = 9$:

$$s^2_{факт} = \frac{224}{2} = 112, \quad s^2_{ост} = \frac{204}{9} = 22{,}67;$$

$$F_{набл} = \frac{s^2_{факт}}{s^2_{ост}} = \frac{112}{22{,}67} = 4{,}94.$$

По таблице критических точек распределения Фишера-Снедекора найдем

$$F_{кр}(\alpha, k_1, k_2) = F_{кр}(0{,}05; 2; 9) = 4{,}26.$$

$$F_{набл} > F_{кр} \quad -$$

нулевая гипотеза отклоняется (групповые средние различаются значимо).

Ответ: $F_{набл} = 4{,}94$, гипотеза отклоняется.

2.2.4. *Моделирование случайных величин методом Монте-Карло (статистических испытаний)*

Задачу, для решения которой применяется метод Монте-Карло, можно сформулировать так: требуется найти значение *a* изучаемой случайной величины. Для его определения выбирается случайная величина *X*, математическое ожидание которой равно *a*, и для выборки из *n* значений *X*, полученных в *n* испытаниях, вычисляется выборочное среднее:

$$\bar{x} = \frac{\sum x_i}{n},$$

которое принимается в качестве оценки искомого числа *a*:

$$a \approx a^* = \bar{x}.$$

Этот метод требует проведения большого числа испытаний, поэтому его иначе называют **методом статистических испытаний**. Теория метода Монте-Карло исследует, как наиболее целесообразно выбрать случайную величину *X*, как найти ее возможные значения, как уменьшить дисперсию используемых случайных величин, чтобы погрешность при замене *a* на *a** была возможно меньшей.

Поиск возможных значений *X* называют **разыгрыванием случайной величины**. Рассмотрим некоторые способы разыгрывания случайных величин и выясним, как оценить допускаемую при этом ошибку.

Оценка погрешности метода Монте-Карло

Если поставить задачу определения верхней границы допускаемой ошибки с заданной доверительной вероятностью γ, то есть поиска числа δ, для которого

$$p(|\bar{X} - a| \le \delta) = \gamma,$$

то получим известную задачу определения доверительного интервала для математического ожидания генеральной совокупности. Воспользуемся результатами решения этой задачи для следующих случаев:

1) случайная величина *X* распределена нормально и известно ее среднее квадратическое отклонение. Тогда

$$\delta = \frac{t\sigma}{\sqrt{n}}$$

(см. формулу для доверительного интервала при известном σ), где *n* – число испытаний, σ - известное среднее квадратическое отклонение, а *t* – аргумент функции Лапласа, при котором Φ(*t*) = $\gamma/2$.

2) Случайная величина *X* распределена нормально с неизвестным σ. Тогда из формулы для доверительного интервала при неизвестном σ следует, что

$$\delta = \frac{t_\gamma s}{\sqrt{n}},$$

где *s* – исправленное выборочное среднее квадратическое отклонение, а t_γ определяется по соответствующей таблице.

3) Если случайная величина распределена по иному закону, то при достаточно большом количестве испытаний ($n > 30$) можно использовать для оценки δ предыдущие формулы, так как при $n \to \infty$ распределение Стьюдента стремится к нормальному, и границы интервалов, полученные по формулам пунктов 1 и 2, различаются незначительно.

Разыгрывание случайных величин

Случайными числами называют возможные значения r непрерывной случайной величины R, распределенной равномерно в интервале (0; 1).

1. Разыгрывание дискретной случайной величины

Пусть требуется разыграть дискретную случайную величину X, то есть получить последовательность ее возможных значений, зная закон распределения X:

X	x_1	x_2	…	x_n
p	p_1	p_2	…	p_n

Рассмотрим равномерно распределенную в (0, 1) случайную величину R и разобьем интервал (0, 1) точками с координатами p_1, $p_1 + p_2$, …, $p_1 + p_2 + \ldots + p_{n-1}$ на n частичных интервалов $\Delta_1, \Delta_2, \ldots, \Delta_n$, длины которых равны вероятностям с теми же индексами.

Теорема 1. Если каждому случайному числу $r_j (0 \le r_j < 1)$, которое попало в интервал Δ_i, ставить в соответствие возможное значение x_i, то разыгрываемая величина будет иметь заданный закон распределения:

X	x_1	x_2	…	x_n
p	p_1	p_2	…	p_n

Доказательство.

Возможные значения полученной случайной величины совпадают с множеством x_1, x_2,… x_n, так как число интервалов равно n, а при попадании r_j в интервал Δ_i случайная величина может принимать только одно из значений x_1, x_2,… x_n.

Так как R распределена равномерно, то вероятность ее попадания в каждый интервал равна его длине, откуда следует, что каждому значению x_i соответствует вероятность p_i. Таким образом, разыгрываемая случайная величина имеет заданный закон распределения.

Пример 1.

Разыграть 10 значений дискретной случайной величины X, закон распределения которой имеет вид:

X	2	3	6	8
p	0,1	0,3	0,5	0,1

Решение. Разобьем интервал (0, 1) на частичные интервалы: Δ_1- (0; 0,1), Δ_2 – (0,1; 0,4), Δ_3 - (0,4; 0,9), Δ_4 – (0,9; 1). Выпишем из таблицы случайных чисел 10 чисел: 0,09; 0,73; 0,25; 0,33; 0,76; 0,52; 0,01; 0,35; 0,86; 0,34. Первое и седьмое числа лежат на интервале Δ_1, следовательно, в этих случаях разыгрываемая случайная величина приняла значение $x_1 = 2$; третье, четвертое, восьмое и десятое числа попали в интервал Δ_2, что соответствует $x_2 = 3$; второе, пятое, шестое и девятое числа оказались в интервале Δ_3 – при этом $X = x_3 = 6$; на последний интервал не попало ни одного числа. Итак, разыгранные возможные значения X таковы: 2, 6, 3, 3, 6, 6, 2, 3, 6, 3.

2. Разыгрывание противоположных событий

Пусть требуется разыграть испытания, в каждом из которых событие A появляется с известной вероятностью p. Рассмотрим дискретную случайную величину X, принимающую значения 1 (в случае, если событие A произошло) с вероятностью p и 0 (если A не произошло) с вероятностью $q = 1 - p$. Затем разыграем эту случайную величину так, как было предложено в предыдущем пункте.

Пример 2.

Разыграть 10 испытаний, в каждом из которых событие A появляется с вероятностью 0,3.

Решение. Для случайной величины X с законом распределения

X	1	0
p	0,3	0,7

получим интервалы Δ_1 – (0; 0,3) и Δ_2 – (0,3; 1). Используем ту же выборку случайных чисел, что и в предыдущем примере, для которой в интервал Δ_1 попадают числа №1,3 и 7, а остальные – в интервал Δ_2. Следовательно, можно считать, что событие A произошло в первом, третьем и седьмом испытаниях, а в остальных – не произошло.

3. Разыгрывание полной группы событий.

Если события A_1, A_2, …, A_n, вероятности которых равны p_1 , p_2 ,… p_n, образуют полную группу, то для из разыгрывания (то есть моделирования последовательности их появлений в серии испытаний) можно разыграть дискретную случайную величину X с законом распределения

X	1	2	…	n
p	p_1	p_2	…	p_n

сделав это так же, как в пункте 1. При этом считаем, что если X принимает значение $x_i = i$, то в данном испытании произошло событие A_i.

4. Разыгрывание непрерывной случайной величины.

а) Метод обратных функций.

Пусть требуется разыграть непрерывную случайную величину X, то есть получить последовательность ее возможных значений x_i ($i = 1, 2, \ldots, n$), зная функцию распределения $F(x)$.

Теорема 2. Если r_i – случайное число, то возможное значение x_i разыгрываемой непрерывной случайной величины X с заданной функцией распределения $F(x)$, соответствующее r_i , является корнем уравнения

$$F(x_i) = r_i. \qquad (1)$$

Доказательство.

Так как $F(x)$ монотонно возрастает в интервале от 0 до 1, то найдется (причем единственное) значение аргумента x_i , при котором функция распределения примет значение r_i . Значит, уравнение (1) имеет единственное решение: $x_i = F^{-1}(r_i)$, где F^{-1}- функция, обратная к F. Докажем, что корень уравнения (1) является возможным значением рассматриваемой случайной величины X. Предположим вначале, что x_i – возможное значение некоторой случайной величины ξ, и докажем, что вероятность попадания ξ в интервал (c, d) равна $F(d) – F(c)$. Действительно,

$$c < x_i < d \Leftrightarrow F(c) < r_i < F(d)$$

в силу монотонности $F(x)$ и того, что $F(x_i) = r_i$. Тогда

$$c < \xi < d \Leftrightarrow F(c) < R < F(d),$$

следовательно,

$$p(c < \xi < d) = p(F(c) < R < F(d)) = F(d) - F(c).$$

Значит, вероятность попадания ξ в интервал (c, d) равна приращению функции распределения $F(x)$ на этом интервале, следовательно, $\xi = X$.

Пример 3.

Разыграть 3 возможных значения непрерывной случайной величины X, распределенной равномерно в интервале (5; 8).

Решение.

$$F(x) = \frac{x-5}{3},$$

то есть требуется решить уравнение

$$\frac{x_i - 5}{3} = r_i, \quad x_i = 3r_i + 5.$$

Выберем 3 случайных числа: 0,23; 0,09 и 0,56 и подставим их в это уравнение. Получим соответствующие возможные значения X:

$$x_1 = 5{,}69; \quad x_2 = 5{,}27; \quad x_3 = 6{,}68.$$

б) Метод суперпозиции.

Если функция распределения разыгрываемой случайной величины может быть представлена в виде линейной комбинации двух функций распределения:

$$F(x) = C_1F_1(x) + C_2F_2(x) \quad (C_{1,2} > 0),,$$

то $C_1 + C_2 = 1$, так как при $x \to \infty \quad F(x) \to 1$.

Введем вспомогательную дискретную случайную величину Z с законом распределения

Z	1	2
p	C_1	C_2

Выберем 2 независимых случайных числа r_1 и r_2 и разыграем возможное значение Z по числу r_1 (см. пункт 1). Если $Z = 1$, то ищем искомое возможное значение X из уравнения $F_1(x) = r_2$, а если $Z = 2$, то решаем уравнение $F_1(x) = r_2$.

Можно доказать, что при этом функция распределения разыгрываемой случайной величины равна заданной функции распределения.

в) Приближенное разыгрывание нормальной случайной величины.

Так как для R, равномерно распределенной в (0, 1),

$$M(R) = \frac{1}{2}, \quad D(R) = \frac{1}{12},$$

то для суммы n независимых, равномерно распределенных в интервале (0,1) случайных величин $\sum_{j=1}^{n} R_j$

$$M\left(\sum_{j=1}^{n} R_j\right) = \frac{n}{2}, \quad D\left(\sum_{j=1}^{n} R_j\right) = \frac{n}{12}, \quad \sigma = \sqrt{\frac{n}{12}}.$$

Тогда в силу центральной предельной теоремы нормированная случайная величина

$$\frac{\sum_{j=1}^{n} R_j - \frac{n}{2}}{\sqrt{\frac{n}{12}}}$$

при $n \to \infty$ будет иметь распределение, близкое к нормальному, с параметрами $a = 0$ и $\sigma = 1$. В частности, достаточно хорошее приближение получается при $n = 12$: $\sum_{j=1}^{12} R_j - 6$.

Итак, чтобы разыграть возможное значение нормированной нормальной случайной величины x, надо сложить 12 независимых случайных чисел и из суммы вычесть 6.

ПРИМЕРЫ РЕШЕНИЯ ЗАДАЧ

Задача 1.
Разыграть 10 значений дискретной случайной величины X, закон распределения которой имеет вид:

X	-3	0	5	7
p	0,2	0,2	0,4	0,2

используя следующие значения случайных чисел:

0,66 0,95 0,90 0,91 0,17 0,39 0,29 0,27 0,49 0,45.

Указание

Разбейте интервал (0, 1) на частичные интервалы: Δ_1- (0; 0,2), Δ_2 – (0,2; 0,4), Δ_3 - (0,4; 0,8), Δ_4 – (0,8; 1), и каждому случайному числу $r_j (0 \le r_j < 1)$, которое попало в интервал Δ_i, поставьте в соответствие возможное значение x_i.

Решение

Разобьем интервал (0, 1) на частичные интервалы: Δ_1- (0; 0,2), Δ_2 – (0,2; 0,4), Δ_3 - (0,4; 0,8), Δ_4 – (0,8; 1). На первый из них попало одно число из выборки случайных чисел, на второй – три числа, на третий – три и на четвертый – три. С учетом последовательности случайных чисел результат разыгрывания случайной величины имеет вид: 5,7,7,7,-3,0,0,0,5,5.
Ответ: 5,7,7,7,-3,0,0,0,5,5.

Задача 2.
Разыграть 10 испытаний, в каждом из которых событие A появляется с вероятностью 0,6, используя следующие значения случайных чисел:

0,15 0,95 0,33 0,47 0,64 0,35 0,08 0,03 0,36 0,06.

Указание

Разбейте интервал (0, 1) на частичные интервалы: Δ_1- (0; 0,6) и Δ_2 – (0,6; 1). При попадании случайного числа на первый интервал считаем, что в данном испытании событие A произошло, а при попадании на второй интервал – не произошло.

Решение

Разобьем интервал (0, 1) на частичные интервалы: Δ_1- (0; 0,6) и Δ_2 – (0,6; 1).

На первый интервал попали случайные числа с номерами 1,3,4,6,7,8,9,10, на второй – числа с номерами 2 и 5. Следовательно, событие A произошло во всех испытаниях, кроме второго и пятого.
Ответ: A произошло во всех испытаниях, кроме второго и пятого.

Задача 3.
Разыграть 10 испытаний, в которых могут появиться события A_1, A_2 и A_3, образующие полную группу, если вероятности этих событий равны соответственно 0,4, 0,5 и 0,1. Воспользоваться следующими значениями случайных чисел:

0,37 0,08 0,92 0,00 0,48 0,61 0,19 0,69 0,04 0,46.

Указание

Разыграйте дискретную случайную величину X с законом распределения

X	1	2	3
p	0,4	0,5	0,1

Решение

Разыграем дискретную случайную величину X с законом распределения

X	1	2	3
p	0,4	0,5	0,1

Разобьем интервал (0, 1) на частичные интервалы: Δ_1- (0; 0,4), Δ_2 – (0,4; 0,9), Δ_3 - (0,9; 1). На первый попали числа с номерами 1,2,4,7 и 9, на второй – с номерами 5,6,8 и 10, на третий – число с номером 3. Следовательно, можно считать, что последовательность появления событий в серии испытаний такова: $A_1, A_1, A_3, A_1, A_2, A_2, A_1, A_2, A_1, A_2$.
Ответ: $A_1, A_1, A_3, A_1, A_2, A_2, A_1, A_2, A_1, A_2$.

Задача 4.
Разыграть 5 возможных значений непрерывной случайной величины, функция распределения которой имеет вид:

$$F(x)=\begin{cases} 0, & x\le 0, \\ \dfrac{x^2}{16}, & 0<x\le 4, \\ 1, & x>4. \end{cases}$$

Воспользоваться следующими значениями случайных чисел:

0,40 0,21 0,95 0,25 0,63.

Указание

Если r_i – случайное число, то возможное значение x_i разыгрываемой непрерывной случайной величины X с заданной функцией распределения $F(x)$, соответствующее r_i , является корнем уравнения

$$F(x_i)=r_i.$$

Решение

Решим для каждого случайного числа уравнение вида

$$F(x_i)=r_i.$$

$$1)\quad \frac{x^2}{16}=0,40,\quad x^2=6,4,\quad x_1=2,53.$$

$$2)\quad \frac{x^2}{16}=0,21,\quad x^2=3,36,\quad x_2=1,83.$$

$$3)\quad \frac{x^2}{16}=0,95,\quad x^2=15,2,\quad x_3=3,90.$$

$$4)\quad \frac{x^2}{16}=0,25,\quad x^2=4,\quad x_4=2,00.$$

$$5)\quad \frac{x^2}{16}=0,63,\quad x^2=10,08,\quad x_5=3,17.$$

Ответ: 2,53; 1,83; 3,90; 2,00; 3,17.

Задача 5.

Разыграть значение нормально распределенной случайной величины с математическим ожиданием 3 и дисперсией 4. Воспользоваться следующими значениями случайных чисел:

0,07 0,20 0,73 0,17 0,90 0,23 0,52 0,37 0,83 0,17 0,73 0,20.

Указание

Чтобы разыграть возможное значение **нормированной** нормальной случайной величины y, надо сложить 12 независимых случайных чисел и из суммы вычесть 6. При этом значение данной случайной величины связано с y соотношением

$$y=\frac{x-a}{\sigma}=\frac{x-M(X)}{\sqrt{D(X)}}.$$

Решение

Найдем соответствующее данной последовательности случайных чисел значение **нормированной** нормальной случайной величины y:

$$y=0,07+\ldots+0,20-6=5,12-6=-0,88.$$

Тогда

$$x=\sigma y+a=-2\cdot 0,88+3=1,24.$$

Ответ: 1,24.