Тема 1

Теория вероятностей – математическая наука, изучающая закономерности в явлениях и опытах, результаты которых не могут быть заранее предсказаны.

Историческая справка

Возникновение теории вероятностей как науки относят к средним векам, к романтическому времени королей и мушкетеров, прекрасных дам и благородных рыцарей. Первоначальным толчком к развитию теории вероятностей послужили задачи, относящиеся к азартным играм, таким, как орлянка, кости, карты, рулетка, когда в них начали применять количественные подсчеты и прогнозирование шансов на успех. В переводе с французского «азарт» (le hazard) означает случай. Такого рода задачи неоднократно ставились в средневековой литературе, в том числе, и художественной, и решались иногда верно, а иногда неверно. Мощным стимулом развития теории явились запросы страхового дела, которое зародилось еще в XIV веке, а также, начиная с XVII века, демографии или, как тогда говорили, политической арифметики.

Зарождение теории вероятностей началось с того, что придворный французского короля, шевалье (кавалер) де Мере (1607-1648), сам азартный игрок, обратился к французскому физику, математику и философу Блезу Паскалю (1607-1648) с вопросами к задаче об очках. До нас дошли два знаменитых вопроса де Мере к Паскалю:

1) сколько раз надо бросить две игральные кости, чтобы случаев выпадения сразу двух шестерок было больше половины от общего числа бросаний;

2) как справедливо разделить поставленные на кон деньги, если игроки прекратили игру преждевременно? В 1654 г. Паскаль обратился к математику Пьеру Ферма (1601-1665) и переписывался с ним по поводу этих задач. Они вдвоем установили некоторые исходные положения ТВ, в частности пришли к понятию математического ожидания и теоремам сложения и умножения вероятностей. Далее голландский ученый Х.Гюйгенс (1629-1695) в книге «О расчетах при азартных играх» (1657 г.) попытался дать собственное решение вопросов, затронутых в этой переписке.

Другим толчком для развития теории вероятностей послужило страховое дело, а именно с конца XVII века на научной основе стало производиться страхование от несчастных случаев и стихийных бедствий. В XVI-XVII веках во всех странах Западной Европы получило распространение страхование судов и страхование от пожара. В XVIII веке были созданы многочисленные страховые компании и лотереи в Италии, Фландрии, Нидерландах. Затем методы ТВ стали широко применять в демографии, например при ведении статистики рождения и смерти. Важную роль для развития математической статистики сыграли работы Э. Галлея по демографии. Заметим, что «по основной специальности» этот ученый был астрономом, и его именем названа знаменитая комета.

Стала зарождаться новая наука, вырисовываться ее специфика и методология: определения, теоремы, методы.

Становление ТВ связано с именем известного швейцарского математика Якоба Бернулли (1654-1705). В его трактате «Искусство предположений» (1713), над которым он работал 20 лет и который был издан уже после смерти автора, впервые введено и широко использовалось классическое определение вероятности, а также применялась статистическая концепция вероятности.

Следующий важный этап в развитии ТВ связан с именами Муавра (1667-1754), Лапласа (1749-1827), Гаусса (1777-1855), Пуассона (1781-1840). Далее, в XIX веке, большую роль сыграли представители Петербургской математической школы В.Я. Буняковский (1804-1889), П.Л. Чебышев (1821-1894), А.А. Марков (1856-1922), А.А. Ляпунов (1857-1918).

Большой вклад в последующее развитие ТВ и математической статистики внесли российские математики С.Н. Бернштейн, В.И. Романовский (1879-1954), А.Н. Колмогоров, А.Я. Хинчин (1894-1959), Ю.В. Ленник, Б.В. Гнеденко, Н.В. Смирнов и др., а также ученые англо-американской школы Стьюдент (псевдоним В. Госсета), Р. Фишер, Э. Пирсон, Е. Нейман, А. Вальд и др. Особо следует отметить неоценимый вклад академика А.Н. Колмогорова в становлении теории вероятностей как математической науки. Фундаментом современного здания теории вероятностей является аксиоматический подход, предложенный А.Н. Колмогоровым в книге «Основные понятия теории вероятностей». В настоящее время аксиоматический подход является общепринятым. Следует отметить, что в других разделах математики аксиоматический подход был принят значительно раньше, чем в теории вероятностей.

ТВ и математическая статистика и в настоящее время развиваются и применяются на практике: при организации производства, анализе экономических процессов, контроле качества продукции, маркетинговых и социологических исследованиях, страховом деле и т.д.

Основные понятия

В ТВ вводятся специальные понятия и строятся специфические математические модели. Исходными понятиями в ТВ являются понятие случайного события и вероятности. Под ***испытанием*** (опытом, экспериментом) понимается выполнение определенного комплекса условий, в которых наблюдается то или иное явление, фиксируется тот или иной результат. В теории вероятностей рассматриваются ***испытания***, результаты которых нельзя предсказать заранее, а сами испытания можно повторять, хотя бы теоретически, произвольное число раз при неизменном комплексе условий. Испытаниями, например, являются: подбрасывание монеты, выстрел из винтовки, проведение денежно-вещевой лотереи.

Случайным событием (*возможным событием* или просто *событием*) называется любой факт, который в результате испытания может произойти или не произойти. Для приведенных выше испытаний приведем примеры случайных событий: появление герба (реверса), попадание (промах) в цель, выигрыш автомобиля по билету лотереи. Случайное событие – это не какое-

нибудь происшествие, а лишь возможный ***исход***, результат испытания (опыта, эксперимента). События обозначаются прописными (заглавными) буквами латинского алфавита: *A,B,C.*

В реальности примерами случайных событий являются: соотношение курсов валют, доходность акций, цена реализованной продукции, стоимость выполнения больших проектов, продолжительность жизни человека, броуновское движение частиц, как результат их взаимных соударений и многое другое. Случайность и потребность в консолидации усилий по борьбе со стихией (природы, рынка и т.д.), точнее создание структур для возмещения неожиданного ущерба за счет взносов всех участников, породило теорию и институты страхования.

Если при каждом испытании, при котором происходит событие A, происходит и событие B, то говорят, что A *влечет за собой* событие B (*входит* в B) или B *включает* событие A и обозначают $A \subset B$. Если одновременно $A \subset B$ и $B \subset A$, то в этом случае события A и B называются *равносильными*. События A и B называются ***несовместными***, если наступление одного из них исключает появление другого в одном и том же испытании. События A и B называются ***совместными*** если они могут произойти вместе в одном и том же испытании.

Пример 1. Испытание состоит в однократном подбрасывании игральной кости с шестью гранями. Событие A – появление трех очков, событие B – появление четного числа очков, C – появление нечетного числа очков. События A и C совместны, поскольку число 3 – нечетное, а значит, если выпало 3 очка, то произошло и событие A и событие C. Кроме того, событие A влечет за собой событие C. События A и B несовместны, т.к. если произошло и событие A, то не произойдет событие B, а если произошло событие B, то не произойдет событие A. События B и C также являются несовместными.

События $A_1, A_2, ..., A_n$ называются ***попарно несовместными*** *(или* ***взаимоисключающими****)*, если любые два из них несовместны.

Пример 2. Испытание – сдача студентом экзамена по определенной дисциплине. События $A_1, A_2, ..., A_{10}$ – соответственно студент получит на экзамене один балл, два, три и т.д. Эти события являются попарно несовместными.

События $A_1, A_2, ..., A_n$ образуют ***полную группу*** для данного испытания, если они попарно несовместны и в результате испытания обязательно появится одно из них.

В примере 2 события $A_1, A_2, ..., A_{10}$ образуют полную группу, а события $A_1, A_2, ..., A_6$ – нет.

Для одного и того же испытания можно рассматривать различные полные группы событий. Так, в примере 2 полную группу также образуют события B={студент получил отметку не выше 5}, C={студент получил отметку выше 5}. Другой пример полной группы событий в этом же испытании – события $B, A_6, A_7, A_8, A_9, A_{10}$.

Два несовместных события, из которых одно должно обязательно произойти, называются ***противоположными*** (в литературе такие события называют также ***взаимно-дополнительными***). Событие, противоположное событию A, будем обозначать $\bar{A}$. Противоположные события являются простейшим примером полной группы.

Пример 3. «Выигрыш» и «проигрыш» по одному билету денежно-вещевой лотереи – события противоположные.

Событие называется ***достоверным***, если в результате испытания оно обязательно должно произойти. Событие называется ***невозможным***, если в данном испытании оно заведомо не может произойти. Обозначим достоверное событие Ω, а невозможное $\varnothing$.

Пример 4. На склад поступила партия, все изделия которой стандартны. Извлечение из нее стандартного изделия – событие достоверное, извлечение же бракованного изделия есть событие невозможное.

События называются ***равновозможными,*** если нет оснований считать, что одно из них является более возможным, чем другое. Или другими словами: под равновозможными понимают события, которые в силу тех или других причин (например, симметрии) не имеют объективного преимущества одного перед другим.

Примеры равновозможных событий: 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

Для практической деятельности важно уметь сравнивать события по степени возможности их наступления. Например, интуитивно ясно, что 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 для сравнения событий нужна определенная мера.

Численная мера степени объективной возможности наступления события называется ***вероятностью события*** и является, наряду с понятием случайного события, вторым основным понятием теории вероятностей. Это определение, *качественно* отражающее понятие вероятности события, не является математическим. Чтобы оно таковым стало, необходимо определить его *количественно*.

Отметим, что строгое математическое определение вероятности, как и случайного события, является аксиоматическим (то, что принимается как данность) и поэтому не поддается строгому определению. То, что в дальнейшем будет называться различными определениями вероятности, представляет собой *способы* или *методы вычисления* этой величины.

Пусть производится испытание с конечным числом равновозможных исходов $\omega_1, \omega_2 ..., \omega_n$, образующих полную группу событий. Элементарный

исход ω_i называется ***благоприятствующим*** появлению события A, если наступление исхода ω_i влечет за собой наступление события A.

Пусть число возможных исходов опыта равно n (общее число элементарных исходов), а при m из них происходит некоторое событие А (число благоприятных исходов), тогда при сделанных ранее предположениях на испытание, ***вероятность P(A) случайного события A***, наступившего в данном испытании вычисляется по формуле:

$$P(A) = \frac{m}{n}. \qquad (1.1)$$

Это, так называемое, *классическое определение вероятности*.

Свойства вероятности:

1. Вероятность достоверного события Ω 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 исходя из (1.1), $P(\Omega) = 1$.

2. Вероятность невозможного события $\varnothing$ равна нулю.

Доказательство. Для невозможного события ни один исход опыта не является благоприятным, поэтому $m = 0$ и на основании формулы (1.1) имеем $P(\varnothing) = 0$.

3. Вероятность *любого* события удовлетворяет двойному неравенству $0 \le P(A) \le 1$.

Доказательство. Случайное событие происходит при некоторых благоприятствующих исходах опыта m, удовлетворяющих неравенству $0 \le m \le n$ (0–для невозможного события и n–для достоверного), и из (1.1) следует, что $0 \le P(A) \le 1$.

События, вероятности которых очень малы (близки к нулю) или очень велики (близки к единице), называются ***практически невозможными*** или ***практически достоверными*** событиями.

Пример 5. 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 применить классическую схему. Следовательно, число возможных исходов равно 10, а число исходов, благоприятных событию A (появлению белого шара) – 6 (таково количество белых шаров в урне). Значит,

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

Следует обратить особое внимание на то, что формула (1.1) справедлива только в случае всех равновозможных исходов. Пренебрежение этим требованием приводило к ошибкам при решении простых вероятностных задач. Приведем хрестоматийный пример – ошибку Ж.Даламбера, попавшую даже во французскую энциклопедию. Ответ Ж.Даламбера на вопрос о вероятности

выпадения «герба» (Г) хотя бы один раз при двух бросаниях монеты гласил– $\frac{2}{3}$. Вероятно, он считал, что при двух бросаниях монеты возможны три следующих исхода: Г–Г, Г–Р, Р–Р, и среди них только последний является неблагоприятным. На самом же деле, для того чтобы все исходы были равновозможными, необходимо учитывать, что помимо исхода Г–Р, возможен и исход Р–Г. С учетом этого искомая вероятность равна $\frac{3}{4}$.

Относительная частота. Статистическое определение вероятности

Классическое определение вероятности применимо только для очень узкого класса задач, где все исходы опыта удовлетворяют жестким условиям, и не является пригодным для изучения произвольных случайных событий. В большинстве реальных задач эта схема неприменима. Так, она неприемлема, если результаты испытания не равновозможны. Например, при бросании неправильной монеты выпадение ее различных граней не равновозможно.

В таких ситуациях требуется определять вероятность события иным образом. Для этого введем вначале понятие ***относительной частоты (частости) W(A)*** события A как отношения числа опытов, в которых наблюдалось событие A, к общему количеству проведенных испытаний:

$$W(A) = \frac{M}{N}, \qquad (1.2)$$

где N – общее число опытов, M – число опытов, в которых появилось событие A.

Большое количество экспериментов показало, что если опыты проводятся в одинаковых условиях, то для большого количества испытаний относительная частота $W(A)$ изменяется мало, колеблясь около некоторого постоянного числа $P^*(A)$. Это число $P^*(A) = W(A)$ можно считать вероятностью рассматриваемого события. В отличие от «математической» вероятности $P(A)$, рассматриваемой в классическом определении, статистическая вероятность $P^*(A)$ является характеристикой *опытной, экспериментальной*.

Таким образом, ***статистической вероятностью события*** считают его относительную частоту или число, близкое к ней.

Замечание 1. Из формулы (1.2) следует, что свойства вероятности, доказанные для ее классического определения, справедливы и для статистического определения вероятности.

Статистическое определение вероятности применимо не к любым событиям с неопределенными исходами, которые в житейской практике считаются случайными, а только к тем из них, которые обладают определенными свойствами:

1) Рассматриваемые события должны быть исходами только тех испытаний, которые могут быть воспроизведены неограниченное число раз при одном и том же комплексе условий;
2) События должны обладать, так называемой, статистической устойчивостью или устойчивостью относительных частот. Это означает,

что в различных сериях испытаний относительная частота события изменяется незначительно (тем меньше, чем больше число испытаний), колеблясь около постоянного числа.

Замечание 2. Недостатком статистического определения является неоднозначность статистической вероятности.

Пример 6. 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, например, около семидесяти раз из каждой сотни выстрелов.

Основные формулы комбинаторики

При вычислении вероятностей часто приходится использовать некоторые формулы *комбинаторики* – раздела математики, изучающего комбинации, которые можно составить по определенным правилам из элементов заданного, обычно конечного, множества. Определим основные такие комбинации.

1. **Правило суммы.**

Если объект A можно выбрать m способами, а другой объект B можно выбрать n способами, то выбор «A *или* B» можно осуществить $m+n$ способами.

2. **Правило умножения.**

Если объект A можно выбрать m способами и если после каждого такого выбора объект B можно выбрать n способами, то выбор «A *и* B» в указанном порядке можно осуществить mn способами.

Эти правила дают удобные универсальные методы решения многих комбинаторных задач.

Пример 7. Три человека независимо друг от друга решили поместить свои вклады в банк. Банков всего 5.

Решение. По правилу произведения общее число способов выбора равно $N = 8 \cdot 8 \cdot 8 = 8^3$

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

$$P_n = n! \qquad (1.3)$$

Пример 8. 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

$$A_n^m = n(n-1)(n-2)...(n-m+1). \qquad (1.4)$$

Пример 9. 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

$$C_n^m = \frac{n!}{m!(n-m)!}. \qquad (1.5)$$

Пример 10. 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:

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

Геометрическая вероятность

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

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

$$P(A) = \frac{l_{CD}}{L_{MN}}, \qquad (1.6)$$

где l – длина отрезка CD, а L – длина отрезка MN.

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

$$P(A) = \frac{s_g}{S_G}, \qquad (1.6`)$$

где s – площадь части g области G, а S – площадь всей области G.

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

$$P(A) = \frac{v}{V}, \qquad (1.6``)$$

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

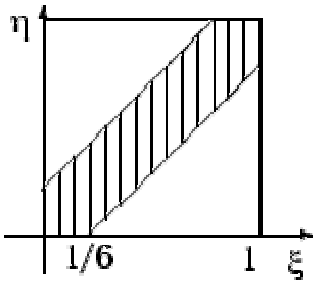
Обобщим приведенные выше формулы. Пусть событие Ω означает, что точка случайным образом попадает во множество Ω и, аналогично, A– точка попадет в подмножество $A \subset \Omega$, причем точка наверняка попадает во множество Ω, т.е. событие A– достоверно. Пусть далее вероятность $P(A)$

события A пропорциональна *геометрической мере* mes(A) (от французского mesure–мера) и мера mes(Ω) множества Ω конечна. Тогда естественно определить $P(A)$ соотношением

$$P(A) = \frac{mes(A)}{mes(\Omega)}$$

Пример 11. Два лица X и Y условились встретиться в определённом месте между двумя и тремя часами дня. Пришедший первым ждет другого в течение 10 минут, после чего уходит. Чему равна вероятность встречи этих лиц, если каждый из них может прийти в любое время в течение указанного часа независимо от другого?

Решение. Будем считать интервал с 14 до 15 часов отрезком [0, 1] длиной в 1 час. Пусть ξ(«кси») и η(«эта») — моменты прихода X и Y— точки отрезка [0, 1]. Все возможные результаты эксперимента — точки квадрата со стороной 1: $\Omega = \{(\xi,\eta) : 0 \le \xi \le 1,\ 0 \le \eta \le 1\} = [0,1] \times [0,1]$.



Можно считать, что эксперимент сводится к бросанию точки наудачу в квадрат. При этом благоприятными исходами являются точки множества A:

$$A = \{(\xi,\eta) : |\xi - \eta| \le 1/6\}$$

(10 минут = 1/6 часа). Попадание в множество A наудачу брошенной в квадрат точки означает, что X и Y встретятся. Тогда вероятность встречи равна

$$P(A) = \frac{mes(A)}{mes(\Omega)} = \frac{S_A}{S_\Omega} = \frac{1 - (5/6)^2}{1} = \frac{11}{36}$$

Пример 12. 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 $x + y = z$, $x + z = y$, $y + z = x$

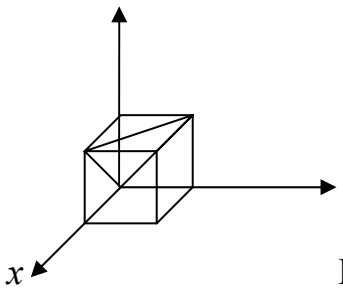


Рис.1.

(одна из них, плоскость $x + y = z$, проведена на рис.1). 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 $p = \frac{v}{V} = \frac{1}{2} : 1 = \frac{1}{2}$.

Тема 2

Алгебра событий

Приведем *теоретико-множественную трактовку* основных понятий теории вероятностей, рассмотренных выше.

Множество всех взаимоисключающих исходов эксперимента называется ***пространством элементарных событий***. Пространство элементарных событий будем обозначать буквой Ω, а его исходы – буквой ω, т.е. $\omega \in \Omega$.

Пример 1. Выпадение на игральной кости: одного очка ω_1,….., выпадение шести очков ω_6. Это элементарные события и их уже *нельзя разбить* на более мелкие.

На практике интересуют события *неэлементарные.*

Событие может быть определено как произвольное подмножество из пространства элементарных событий Ω, если Ω конечно ($\Omega = \{\omega_1, \omega_2 ..., \omega_N\}$) или счетно* ($\Omega = \{\omega_1, \omega_2, ..., \omega_n, ...\}$).

В примере 1 пространство элементарных событий имеет вид $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$. Событие A, состоящее в выпадении четного числа очков, есть $A = \{\omega_2, \omega_4, \omega_6\}$, $A \in \Omega$.

Введем операции над событиями, которые эквивалентны операциям над соответствующими множествами.

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

Пример 2. Два стрелка делают по одному выстрелу по мишени. Если событие А – попадание первого стрелка, а событие В – второго, то сумма А+В – это хотя бы одно попадание при двух выстрелах.

В примере 1 событие B, состоящее в выпадении нечетного числа очков, есть $B = \omega_1 + \omega_3 + \omega_5$.

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

В примере 2 событием AB будет попадание обоих стрелков.

Пример 3. 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.

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

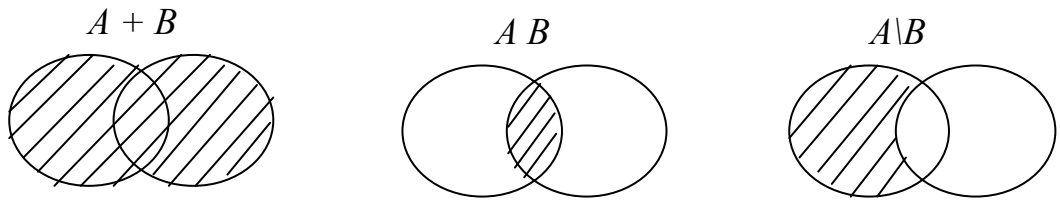
---

* множество называется счетным, если его элементы можно перенумеровать натуральными числами.

Смысл события $A-B$ состоит в том, что событие A наступает, но при этом не наступает событие B. Определенное ранее противоположное событие A можно представить в виде $\overline{A}=\Omega\setminus A$.

В примере 3 $A\setminus B$ – извлечение из колоды любой карты пиковой масти, кроме дамы. Наоборот, $B\setminus A$ – извлечение дамы любой масти, кроме пик.

Дадим геометрическую интерпретацию основных действий над событиями с помощью *диаграмм Венна*.



Замечание. Для определенных ранее несовместных событий A и B справедливо $AB=\varnothing$. Если изобразить графически области исходов опыта, благоприятных несовместным событиям, то они не будут иметь общих точек.

Как правило, для определения вероятностей событий применяются не непосредственные прямые методы, а косвенные, позволяющие по известным вероятностям одних событий определять вероятности других событий, с ними связанных. Применяя эти косвенные методы, мы всегда в той или иной форме пользуемся *основными теоремами* теории вероятностей. Этих теорем две: теорема сложения вероятностей и теорема умножения вероятностей.

Теорема сложения вероятностей

Теорема 1. *Вероятность P(A + B) суммы событий A и B равна*

$$P(A+B)=P(A)+P(B)-P(AB). \qquad (2.1)$$

Доказательство. Докажем теорему сложения для схемы случаев. Пусть n – число возможных исходов опыта, m_A – число исходов, благоприятных событию A, m_B – число исходов, благопри-ятных 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 $(m_A + m_B)$ число m_{AB} учтено дважды: как исходы, благоприятные A, и исходы, благоприятные B). Следовательно, вероятность суммы можно определить по формуле (1.1):

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

что и требовалось доказать.

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

$$P(A+B+C)=P(A)+P(B)+P(C)-P(AB)-P(AC)-P(BC)+P(ABC) \qquad (2.2)$$

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

$$P(A + B) = P(A) + P(B). \qquad (2.3)$$

Следствие 3. Сумма вероятностей противоположных событий равна 1:

$$P(A) + P(\overline{A}) = 1. \qquad (2.4)$$

Доказательство. Так как A и $\overline{A}$ образуют полную группу, то одно из них обязательно произойдет в результате опыта, то есть событие $A + \overline{A}$ является достоверным. Следовательно, $P(A + \overline{A}) = 1$. Но, так как A и $\overline{A}$ несовместны, из (2.3) следует, что $P(A + \overline{A}) = P(A) + P(\overline{A})$. Значит, $P(A) + P(\overline{A}) = 1$, 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 (2.4).

Пример 4. Из урны, содержащей 2 белых и 6 черных шаров, случайным образом извлекаются 5 шаров. Найти вероятность того, что вынуты шары разных цветов.

Решение. Событие $\overline{A}$, 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 (1.5):

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

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

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

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

Теорема умножения вероятностей

При изучении реальных случайных явлений иногда возникает или искусственно создается ситуация, когда мы получаем дополнительную информацию о возможных исходах опыта Ω.

Остановимся более подробно на следующем примере иллюстративного характера. Допустим, что студент из 30 билетов успел выучить билеты с 1-го по 3-й и с 28-го по 30-й. На экзамен он пришел одиннадцатым, и оказалось, что к его приходу остались только билеты с 1-го по 20-й (событие А). Вероятность события В={студент получил выученный билет} без дополнительной информации о том, что событие А произошло, может быть вычислена по классическому определению с Ω={1,2,…,30}. Согласно формуле (1.1) имеем: $P(B) = \frac{6}{30} = \frac{1}{5}$. При дополнительной информации (событие А произошло) множество возможных исходов А состоит из 20 элементарных исходов, а событие В вместе с А наступает в 3 случаях. Следовательно, в рассматриваемом примере естественно определить ***условную вероятность***

$P(B \mid A) = P_A(B)$ ***события B при условии, что событие A произошло***, как $P_A(B) = \frac{3}{20}$.

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

Пример 5. 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: $P(B) = P(A) = \frac{4}{32} = \frac{1}{8} = 0{,}125$. 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 6. 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 2 (теорема умножения). *Вероятность произведения двух событий равна произведению вероятности одного из них на условную вероятность другого при условии, что первое событие произошло:*

$$P(AB) = P(A) \cdot P_A(B). \qquad (2.5)$$

Доказательство. Воспользуемся обозначениями теоремы 1. 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,

$P_A(B) = \frac{m_{AB}}{m_A} = \frac{m_{AB}}{n} \cdot \frac{n}{m_A} = \frac{P(AB)}{P(A)}$, откуда следует утверждение теоремы.

Пример 7. 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,

$$P(A) \cdot P_A(B) = P(B) \cdot P_B(A). \qquad (2.6)$$

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 (2.6) 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:

$$P(AB)=P(A)\cdot P(B)\,, \qquad (2.7)$$

то есть вероятность произведения независимых событий равна произведению их вероятностей.

Теорема умножения вероятностей легко обобщается на случай произвольного числа событий:

$$P(A_1A_2A_3\ldots A_n)=P(A_1)P_{A_1}(A_2)P_{A_1A_2}(A_3)\ldots P_{A_1A_2\ldots A_{n-1}}(A_n) \qquad (2.8)$$

При решении задач теоремы сложения и умножения обычно применяются вместе.

Пример 8. 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

$A = H_1 + H_2$, $B = H_1\cdot\overline{H}_2 + \overline{H}_1\cdot H_2$, $C = H_1\cdot H_2$, $D = \overline{H}_1\cdot\overline{H}_2$. События H_1 и H_2 совместны и независимы, поэтому теорема сложения применяется в общем виде, а теорема умножения – в виде (2.7). 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 $H_1\cdot\overline{H}_2$ и $\overline{H}_1\cdot H_2$ 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 3. *Вероятность появления хотя бы одного из попарно независимых событий $A_1, A_2, \ldots, A_n$ равна*

$$P(A) = 1 - q_1q_2\ldots q_n\,, \qquad (2.8)$$

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 $\overline{A}_1\overline{A}_2\ldots\overline{A}_n$ противоположны, поэтому по следствию 3 из теоремы 1 сумма их вероятностей равна 1. Кроме того, поскольку A_1, A_2,…, A_n независимы, то независимы и $\overline{A}_1, \overline{A}_2, \ldots, \overline{A}_n$, следовательно, $P(\overline{A}_1\overline{A}_2\ldots\overline{A}_n)=P(\overline{A}_1)P(\overline{A}_2)\ldots P(\overline{A}_n) = q_1q_2\ldots q_n$. Отсюда следует справедливость формулы (2.8).

Пример 9. 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 $n > \log_2 10 \geq 4$.

Формула полной вероятности. Формула Байеса

Следствием двух основных теорем теории вероятностей – теоремы сложения и умножения – являются формулы полной вероятности и формулы Байеса.

На языке алгебры событий набор $H_1, H_2, \ldots, H_n$ называется ***полной группой событий***, если:

1. $H_i H_j = \varnothing,\ \forall i \neq j,\ i = 1,2,...,n;\ j = 1,2,...,n$
2. $H_1 + H_2 + \ldots + H_n = \Omega$.

Теорема 4 (Формула полной вероятности). *Если событие A может произойти только при условии появления одного из событий (гипотез) H_i, $i = \overline{1,n}$, образующих полную группу, то вероятность события A равна*

$$P(A) = \sum_{i=1}^{n} P(H_i) \cdot P_{H_i}(A). \qquad (2.9)$$

Доказательство. Так как гипотезы $H_1, H_2, \ldots, H_n$ – единственно возможные, а событие A по условию теоремы может произойти только вместе с одной из гипотез, то

$$A = \Omega A = (H_1 + H_2 + \ldots + H_n)A = H_1 A + H_2 A + \ldots + H_n A$$

Из несовместности гипотез H_i, $i = \overline{1,n}$ следует несовместность $H_i A$, $i = \overline{1,n}$. Применяем теорему сложения вероятностей в виде (2.3):

$$P(A) = P(H_1 A) + P(H_2 A) + \ldots + P(H_n A) = \sum_{i=1}^{n} P(H_i A) \qquad (2.10)$$

По теореме умножения $P(H_i A) = P(H_i) P_{H_i}(A)$. Подставляя данное представление в формулу (2.10), окончательно имеем:

$P(A) = P(H_1)P_{H_1}(A) + P(H_2)P_{H_2}(A) + \ldots + P(H_n)P_{H_n}(A) = \sum_{i=1}^{n} P(H_i)P_{H_i}(A)$, что и требовалось доказать.

Пример 10. Экспортно-импортная фирма собирается заключить контракт на поставку сельскохозяйственного оборудования в одну из развивающихся стран. Если основной конкурент фирмы не станет одновременно претендовать на заключение контракта, то вероятность получения контракта оценивается в 0,45; в противном случае – в 0,25. По оценкам экспертов компании вероятность того, что конкурент выдвинет свои предложения по заключению контракта, равна 0,40. Чему равна вероятность заключения контракта?

Решение. A = «фирма заключит контракт». H_1 – «конкурент выдвинет свои предложения». H_2 – « конкурент не выдвинет свои предложения». По условию задачи $P(H_1) = 0{,}4$, $P(H_2) = 1 - 0{,}4 = 0{,}6$. Условные вероятности по заключению контракта для фирмы $P_{H_1}(A) = 0{,}25$, $P_{H_2}(A) = 0{,}45$. По формуле полной вероятности

$$P(A) = P(H_1) \cdot P_{H_1}(A) + P(H_2) \cdot P_{H_2}(A).$$

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

Следствием теоремы умножения и формулы полной вероятности является формула Байеса.

Она применяется, когда событие А, которое может появиться только с одной из гипотез $H_1, H_2,..., H_n$, образующих полную группу событий, произошло и необходимо провести количественную переоценку априорных вероятностей этих гипотез $P(H_1), P(H_2),..., P(H_n)$ известных до испытания, т.е. надо найти апостериорные (получаемые после проведения испытания) условные вероятности гипотез $P_A(H_1), P_A(H_2),..., P_A(H_n)$.

Теорема 5 (Формула Байеса). *Если событие A произошло, то апостериорные условные вероятности гипотез H_i, $\left(i=\overline{1,n}\right)$ вычисляются по формуле, которая носит название формулы Байеса:*

$$P_A(H_i)=\frac{P(H_i)\cdot P_{H_i}(A)}{P(A)}=\frac{P(H_i)\cdot P_{H_i}(A)}{\sum_{i=1}^{n}P(H_i)\cdot P_{H_i}(A)} \qquad (2.11)$$

Доказательство. Для получения искомой формулы запишем теорему умножения вероятностей событий A и H_i в двух формах:

$$P(AH_i)=P(A)P_A(H_i)=P(H_i)P_{H_i}(A),$$

откуда $P_A(H_i)=\frac{P(H_i)P_{H_i}(A)}{P(A)}$ или с учетом (2.10): $P_A(H_i)=\frac{P(H_i)P_{H_i}(A)}{\sum_{i=1}^{n}P(H_i)P_{H_i}(A)}$,

что и требовалось доказать.

Значение формулы Байеса состоит в том, что при наступлении события *A,* т.е. по мере получения новой информации, мы можем проверять и корректировать выдвинутые до испытания гипотезы. Такой подход, называемый байесовским, дает возможность корректировать управленческие решения в экономике, оценки неизвестных параметров распределения изучаемых признаков в статистическом анализе и т.п.

Пример 11. Экономист-аналитик условно подразделяет экономическую ситуацию в стране на «хорошую», «посредственную» и «плохую» и оценивает их вероятности для данного момента времени в 0,15; 0,70 и 0,15 соответственно. Некоторый индекс экономического состояния возрастает с вероятностью 0,60, когда ситуация «хорошая»; с вероятностью 0,30, когда ситуация посредственная, и с вероятностью 0,10, когда ситуация «плохая». Пусть в настоящий момент индекс экономического состояния возрос. Чему равна вероятность того, что экономика страны на подъеме?

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

$$P_A(H_1)=\frac{P(H_1)\cdot P_{H_1}(A)}{P(H_1)\cdot P_{H_1}(A)+P(H_2)\cdot P_{H_2}(A)+P(H_3)\cdot P_{H_3}(A)};$$

$$P_A(H_1)=\frac{0{,}15\cdot 0{,}6}{0{,}15\cdot 0{,}6+0{,}7\cdot 0{,}3+0{,}15\cdot 0{,}1}=\frac{0{,}09}{0{,}09+0{,}21+0{,}015}=\frac{0{,}09}{0{,}315}\approx 0{,}286.$$

Пример 12. В торговую фирму поступили телевизоры от трех поставщиков в соотношении 1:4:5. Практика показала, что телевизоры, поступающие от 1-го, 2-го и 3-го поставщиков, не потребуют ремонта в течение гарантийного срока соответственно в 98%, 88% и 92% случаев.

1. Найти вероятность того, что поступивший в торговую фирму телевизор не потребует ремонта в течение гарантийного срока.
2. Проданный телевизор потребовал ремонта в течение гарантийного срока. От какого поставщика вероятнее всего поступил этот телевизор.

Решение. Обозначим события:

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

A – телевизор не потребует ремонта в течение гарантийного срока.

По условию:

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

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

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

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

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

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

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

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

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

По формуле Байеса (2.11)

$$P_{\overline{A}}(H_1)=\frac{0{,}1\cdot 0{,}02}{0{,}09}=0{,}022 \quad P_{\overline{A}}(H_2)=\frac{0{,}4\cdot 0{,}12}{0{,}09}=0{,}533 \quad P_{\overline{A}}(H_3)=\frac{0{,}5\cdot 0{,}08}{0{,}09}=0{,}444$$

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

Тема 3

Повторные независимые испытания (схема Бернулли)

На практике часто приходится сталкиваться с задачами, которые можно представить в виде многократно повторяющихся испытаний при данном комплексе условий, в которых представляет интерес вероятность числа m наступлений некоторого события A в n испытаниях.

Последовательные испытания называются ***независимыми относительно события*** A, если вероятность осуществления любого исхода в n-м по счету испытании не зависит от реализации исходов предыдущих испытаний. Если независимые повторные испытания проводятся при одном и том же комплексе условий, то вероятность наступления события A в каждом испытании одна и та же. Простейшим классом повторных независимых испытаний является ***последовательность независимых испытаний с двумя исходами*** («успех» и «неуспех») *и с* ***неизменными вероятностями «успеха»*** p ***и «неуспеха»*** $q = 1 - p$ ***в каждом испытании.*** Описанная последовательность независимых испытаний получила название ***схемы Бернулли.***

Теорема 1. *Если вероятность* p *наступления события* A *в каждом испытании постоянна, то вероятность* $P_n(m)$ *того, что событие* A *наступит ровно* m *раз в* n *независимых испытаниях вычисляется по формуле, называемой* ***формулой Бернулли****:*

$$P_n(m) = C_n^m p^m (1-p)^{n-m} = C_n^m p^m q^{n-m}. \qquad (3.1)$$

Доказательство. Пусть A_i и $\overline{A}_i$ – соответственно появление и непоявление события A в i-м испытании $(i = 1,2,...,n)$, а B_m – событие, состоящее в том, что в n независимых испытаниях событие A появилось m раз. Представим событие B_m через элементарные события A_i. Например, при $n = 3,\ m = 2$ событие $B_2 = A_1 A_2 \overline{A}_3 + A_1 \overline{A}_2 A_3 + \overline{A}_1 A_2 A_3$.

В общем виде

$$B_m = A_1 A_2 ... A_m \overline{A}_{m+1} ... \overline{A}_n + A_1 \overline{A}_2 A_3 ... A_m \overline{A}_{m+1} ... \overline{A}_{n-1} A_n + ... + \overline{A}_1 \overline{A}_2 ... \overline{A}_{n-m} A_{n-m+1} ... A_n, \quad (3.2)$$

Т.е. каждый вариант появления события B_m (каждый член суммы (3.2)) состоит из m событий A и $n-m$ событий $\overline{A}$ с различными индексами. Число всех комбинаций (слагаемых суммы (3.2)) равно числу способов выбора из n испытаний m, в которых событие A произошло, т.е. числу сочетаний C_n^m. Вероятность каждой такой комбинации по теореме умножения для независимых событий равна $p^m q^{n-m}$. В связи с тем, что комбинации между собой несовместны, то по теореме сложения вероятностей получим

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

Пример 1. Изделия некоторого производства содержат 5% брака. Найти вероятность того, что среди пяти взятых наугад изделий: а) нет ни одного испорченного; б) два испорченных.

Решение.

а) По условию задачи $n=5,\ p=0{,}05$. Так как вероятность наступления события A (появление бракованной детали) постоянна для каждого испытания, то задача подходит под схему Бернулли. Находим вероятность того, что среди пяти взятых наудачу изделий нет ни одного испорченного $n=5,\ m=0,\ p=0{,}05$. По формуле Бернулли:

$P_5(0)=C_5^0\cdot 0{,}05^0\cdot 0{,}95^5=1\cdot 1\cdot 0{,}774=0{,}774$.

б) $n=5,\ m=2,\ p=0{,}05$:

$$P_5(2)=C_5^2\cdot 0{,}05^2\cdot 0{,}95^3=\frac{5\cdot 4}{2}\cdot 0{,}0025\cdot 0{,}857=0{,}021.$$

Число m_0 наступления события A в n независимых испытаниях называется ***наивероятнейшим***, если вероятность осуществления этого события $P_n(m_0)$ по крайней мере не меньше вероятностей других событий $P_n(m)$ при любом m.

Можно доказать, что наивероятнейшее число наступлений события A в n испытаниях заключено между числами $np-q$ и $np+p$:

$$np-q\le m_0\le np+p. \qquad (3.3)$$

Отметим, что, так как разность $np+p-(np-q)=p+q=1$, то всегда существует целое число m_0, удовлетворяющее неравенству (3.3). При этом если $np-q$ – целое число, то наивероятнейших числа два $np-q$ и $np+p$.

Пример 2. По данным примера 1 найти наивероятнейшее число появления бракованных деталей из 5 отобранных и вероятность этого числа.

Решение. По формуле (3.3) $5\cdot 0{,}05-0{,}95\le m_0\le 5\cdot 0{,}05+0{,}05$ или $-0{,}7\le m_0\le 0{,}3$. Единственное целое число, удовлетворяющее полученному неравенству, $m_0=0$, а его вероятность $P_5(0)=0{,}774$ была получена в примере 1.

Пример 3. В помещении четыре лампы. Вероятность работы в течение года для каждой лампы 0,8. Чему равно наивероятнейшее число ламп, которые будут работать в течение года?

Решение. По формуле $np-q\le m_0\le np+p$ найдем m_0. По условию $n=4,\ p=0{,}8,\ q=1-0{,}8=0{,}2$:

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

Следовательно, имеется два наивероятнейших числа $m_0=3$ или $m_0=4$.

Асимптотические формулы для подсчета вероятностей по схеме Бернулли

Предположим, что мы хотим вычислить вероятность $P_n(m)$ появления события A при большом числе испытаний n, например, $P_{500}(300)$. По формуле Бернулли (3.1) имеем: $P_{500}(300) = C_{500}^{300} p^{300} q^{200}$. Ясно, что в этом случае непосредственное вычисление по формуле Бернулли технически сложно, тем более, если учесть, что сами p и q – числа дробные. Поэтому возникает естественное желание иметь более простые, пусть даже и приближенные, формулы для вычисления $P_n(m)$ при больших n. Такие формулы, называемые ***асимптотическими***, существуют, среди которых наиболее известны теорема Пуассона, локальная и интегральная теоремы Лапласа.

Теорема 2. *Если вероятность p наступления события A в каждом испытании стремится к 0 ($p \to 0$) при неограниченном увеличении числа n испытаний ($n \to \infty$), причем произведение np стремится к постоянному числу λ ($np \to \lambda$), то вероятность $P_n(m)$ того, что событие A появится m раз в n независимых испытаниях, удовлетворяет предельному равенству*

$$\lim_{n\to\infty} P_n(m) = \frac{\lambda^m e^{-\lambda}}{m!} \qquad (3.4)$$

Строго говоря, условие теоремы Пуассона $p \to 0$ при $n \to \infty$, так что $np \to \lambda$, противоречит исходной предпосылке схемы испытаний Бернулли, согласно которой вероятность наступления события в каждом испытании $p = const$. Однако, если вероятность p –постоянна и мала ($p < 0{,}1$), число испытаний n –велико ($n > 100$) и число $\lambda = np \le 10$, то из предельного равенства (3.4) вытекает приближенная формула Пуассона:

$$P_n(m) \approx \frac{\lambda^m e^{-\lambda}}{m!} \qquad (3.5)$$

Пример 4. На факультете насчитывается 1825 студентов. Какова вероятность того, что 1 сентября является днем рождения одновременно четырех студентов факультета?

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

$$P_{1825}(4) \approx \frac{5^4 e^{-5}}{4!} \approx 0{,}1755.$$

Таким образом, искомая вероятность составляет 17,5%.

Если в схеме Бернулли вероятность p появления события A близка к 1, а число испытаний n велико, для вычисления вероятности $P_n(m)$ также можно использовать формулу Пуассона. При этом находят вероятность того, что событие $\overline{A}$ произойдет $n - m$ раз.

Теорема 3. *Если в схеме Бернулли вероятность p появления события A в каждом из n испытаний существенно отличается от 0 ($p \geq 0{,}1$) и 1 ($p \leq 0{,}9$), то вероятность $P_n(m)$ того, что событие A произойдет m раз в n независимых испытаниях при достаточно большом числе n ($n > 100$) приближенно равна:*

$$P_n(m) \approx \frac{\varphi(x)}{\sqrt{npq}}, \qquad (3.6)$$

где $\varphi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2}$ – функция Гаусса, $x = \frac{m - np}{\sqrt{npq}}$.

Таблица значений функции $\varphi(x)$ приведена в приложении любого учебника по теории вероятности. Пользуясь этой таблицей, необходимо иметь в виду очевидные свойства функции $\varphi(x)$:

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

Приближенную формулу (3.6) называют *локальной формулой Муавра-Лапласа.*

Пример 5. Вероятность найти белый гриб среди прочих равна $\frac{1}{4}$. Какова вероятность того, что среди 300 грибов белых будет 75?

Решение. По условию задачи $p = \frac{1}{4}, m = 75, n = 300$, $q = 1 - p = 1 - \frac{1}{4} = \frac{3}{4}$.

Находим $x = \frac{m - np}{\sqrt{npq}} = \frac{75 - 300 \cdot \frac{1}{4}}{\sqrt{300 \cdot \frac{1}{4} \cdot \frac{3}{4}}} = 0$. По таблице находим $\varphi(0) = 0{,}3989$.

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

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

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

В принципе вычислить каждое слагаемое можно по локальной формуле Муавра-Лапласа (3.6), но большое количество слагаемых делает расчет весьма громоздким. В таких случаях используется следующая теорема.

Теорема 4. *Если вероятность p наступления события A в каждом из n независимых испытаниях постоянна и отлична от нуля и единицы, а число*

испытаний достаточно велико, то вероятность того, что число m наступления события A в n испытаниях заключено между m_1 и m_2 включительно при достаточно большом числе n приближенно равна

$$P_n(m_1 \le m \le m_2) \approx \frac{1}{2}\left(\Phi\left(\frac{m_2 - np}{\sqrt{npq}}\right) - \Phi\left(\frac{m_1 - np}{\sqrt{npq}}\right)\right), \qquad (3.7)$$

где p – вероятность появления успеха в каждом испытании, $q = 1 - p$, $\Phi(x) = \frac{2}{\sqrt{2\pi}} \cdot \int_0^x e^{\frac{-t^2}{2}}\, dt$ – функция (или интеграл вероятностей) Лапласа, значения $\Phi(x)$ приведены в приложениях любого учебника по теории вероятностей.

Приближенную формулу (3.7) называют *интегральной формулой Муавра-Лапласа.* Чем больше n, тем точнее эта формула. При выполнении условия $npq \ge 20$ интегральная формула (3.7), также как и локальная дает незначительную погрешность вычисления вероятности.

Отметим свойства функции $\Phi(x)$:

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

Пример 6. В партии из 768 арбузов каждый арбуз оказывается неспелым с вероятностью $\frac{1}{4}$. Найти вероятность того, что количество спелых арбузов будет в пределах от 564 до 600.

Решение. По условию $n = 768,\ p = 0{,}75,\ m_1 = 564,\ m_2 = 600$. По интегральной теореме Лапласа

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

$$= \frac{1}{2}\left(\Phi\left(\frac{600 - 576}{12}\right) - \Phi\left(\frac{564 - 576}{12}\right)\right) = \frac{1}{2}\left(\Phi(2) + \Phi(1)\right) \approx \frac{1}{2}(0{,}9545 + 0{,}6827) = 0{,}8186.$$

Пример 7. Город ежедневно посещает 1000 туристов, которые днем идут обедать. Каждый из них выбирает для обеда один из двух городских ресторанов с равными вероятностями и независимо друг от друга. Владелец одного из ресторанов желает, чтобы с вероятностью приблизительно 0,99 все пришедшие в его ресторан туристы могли там одновременно пообедать. Сколько мест должно быть для этого в его ресторане?

Решение. Пусть A = «турист пообедал у заинтересованного владельца». Наступление события A будем считать «успехом», $p = P(A) = 0{,}5$, $n = 1000$. Нас интересует такое наименьшее число k, что вероятность наступления не менее чем k «успехов» в последовательности из $n = 1000$ независимых испытаний с вероятностью успеха p = 0,5 приблизительно равна 1 – 0,99 = 0,01. Это как раз вероятность переполнения ресторана. Таким образом, нас интересует такое

наименьшее число k, что $P_{1000} = (k, 1000) \approx 0{,}01$. Применим интегральную теорему Муавра-Лапласа:

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

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

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

Используя таблицу для *Φ(x)*, находим $\frac{k-500}{5\sqrt{10}} \approx 2{,}33$, значит $k = 2{,}33 \cdot 5\sqrt{10} + 500 \approx 536{,}8$. Следовательно, в ресторане должно быть 537 мест.

Тема 4

Случайные величины (дискретные и непрерывные)

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

Случайной величиной называется переменная величина, которая в результате испытания в зависимости от случая принимает одно из возможного множества своих значений, причем заранее неизвестно, какое именно.

В данной теме будем рассматривать скалярные случайные величины, которые принимают значения из множества $\mathbb{R}^1$.

Будем обозначать случайные величины заглавными буквами латинского алфавита ($X,Y,Z,\ldots$), а их возможные значения – соответствующими строчными буквами ($x, y,\ldots$).

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, но счетно 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 (ДСВ)***, если множество $\{x_1, x_2, \ldots, x_n, \ldots\}$ ее возможных значений конечно или счетно (т.е. если все ее значения можно занумеровать).

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

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

Законом распределения случайной величины называется всякое соотношение, устанавливающее связь между возможными значениями случайной величины и вероятностями, с которыми она принимает эти значения.

В этом случае про случайную величину говорят, что она распределена по данному закону распределения или подчинена этому закону распределения.

Закон распределения дискретной случайной величины

Для задания дискретной случайной величины нужно знать все ее возможные значения и вероятности, с которыми принимаются эти значения, т.е. задать ее ***закон распределения,*** который может иметь вид таблицы, формулы или графика.

Таблица, в которой перечислены все ее возможные значения дискретной случайной величины и соответствующие им вероятности, называется ***рядом распределения***. Для удобства возможные значения дискретной случайной величины располагают в таблицу в порядке их возрастания:

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

где $p_i = P(X = x_i)$, $i = 1, 2, ..., n, ...$

Заметим, что события $X = x_1, X = x_2, \ldots, X_n = x_n$ образуют полную группу, следовательно, сумма вероятностей этих событий равна единице

$$\sum_{i=1}^{n(\infty)} p_i = 1. \qquad (4.1)$$

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

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

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

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

$$P(X = 1) = P(\overline{A}_1 A_2 + A_1 \overline{A}_2) = 0,4 \cdot 0,7 + 0,6 \cdot 0,3 = 0,46$$

$$P(X = 2) = P(A_1 A_2) = 0,6 \cdot 0,7 = 0,42$$

Следовательно, ряд распределения имеет вид:

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

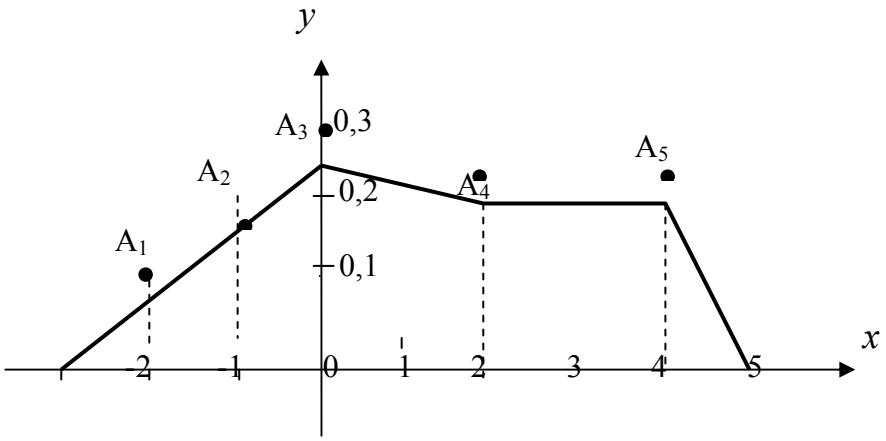
Ряд распределения дискретной случайной величины можно изобразить графически в виде ***полигона*** или ***многоугольника распределения*** вероятностей. Для этого по горизонтальной оси в выбранном масштабе нужно отложить значения случайной величины, а по вертикальной вероятности этих значений. Тогда точки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

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

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

Построить полигон и многоугольник распределения вероятностей.

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



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

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

Решение. Возможные значения случайной величины X–чистого выигрыша на один билет равны $0-7=-7$ ден.ед. (если билет не выиграл), $200-7=193$, 250-7=243, 5000-7=4993 ден.ед.(если на билет выпал выигрыш соответственно видеомагнитофона, телевизора или автомобиля). Учитывая, что из 1000 билетов число невыигравших составляет 990, а указанных выигрышей соответственно 5, 4 и 1, используя классическое определение вероятности, получим:

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

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

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

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

Ряд распределения имеет вид:

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

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

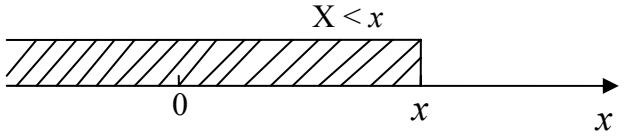
Функция распределения случайной величины

Для описания закона распределения случайной величины X возможен и другой подход: рассматривать не вероятности событий $X = x$ для разных x, как это имеет место в ряде распределения для ДСВ, а вероятности события $X < x$, где x –текущая переменная. Вероятность $P(X < x)$, очевидно, зависит от x, т.е. является некоторой функцией от x.

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

$$F(x) = P(X < x) \qquad (4.2)$$

Если значения случайной величины – точки на числовой оси, то геометрически функция распределения интерпретируется как вероятность того, что случайная величина X попадает левее заданной точки x:



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 $(0 \le p \le 1)$.

2) Функция распределения является неубывающей функцией на всей числовой оси, то есть $F(x_2) \ge F(x_1)$ при $x_2 > x_1$. Это следует из того, что

$$F(x_2)=P(X<x_2)=P((X<x_1)+(x_1 \le X<x_2))=P(X<x_1)+P(x_1 \le X<x_2)=$$
$$=F(x_1)+P(x_1 \le X<x_2). \qquad (4.3)$$

Т.к. вероятность $P(x_1 \le X < x_2) \ge 0$, то из (4.3) вытекает $F(x_2) \ge F(x_1)$.

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

$$\lim_{x \to x_0 - 0} F(x) = F(x_0) \text{ или } F(x_0 - 0) = F(x_0) \qquad (4.4)$$

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

5) $$P(X = x_0) = F(x_0 + 0) - F(x_0 - 0) = F(x_0 + 0) - F(x_0) \qquad (4.5)$$

6) Вероятность того, что случайная величина примет значение из интервала $[a, b)$, равна приращению ее функции распределения на этом интервале:

$$P(a \le \mathrm{X} < \mathrm{b}) = F(b) - F(a). \qquad (4.6)$$

Справедливость этого утверждения следует непосредственно из формулы (4.3).

Таким образом, каждая функция распределения является неубывающей, непрерывной слева и удовлетворяющей условиям $F(-\infty)=0,\ F(+\infty)=1$ функцией. Верно и обратное: каждая функция, удовлетворяющая перечисленным условиям 1)-6), может рассматриваться как функция распределения некоторой случайной величины.

Законом распределения СВ называется любое правило, позволяющее определить ее функцию распределения.

Для дискретной случайной величины значение $F(x)$ в каждой точке представляет собой сумму вероятностей тех ее возможных значений, которые меньше аргумента функции.

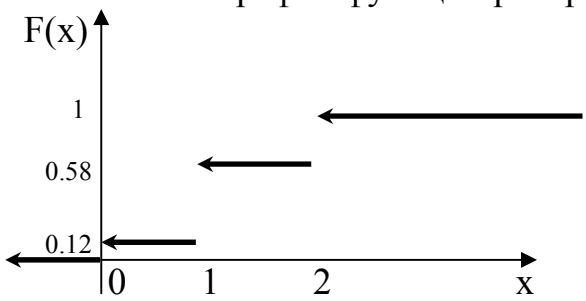
$$F(x)=\sum_{x_k<x} P(X=x_k) \qquad (4.7)$$

Функция распределения любой ДСВ разрывна, возрастает скачками при тех значениях x, которые являются возможными значениями X, а величина скачка находится по формуле (4.5). Если два возможных значения величины X разделены интервалом, в котором других возможных значений X нет, то в этом интервале функция распределения $F(x)$ постоянна. Если возможных значений X конечное число, например n, то функция распределения $F(x)$ представляет собой ступенчатую кусочно-постоянную кривую с $n+1$ интервалом постоянства.

Пример 4. 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:



Непрерывные случайные величины (НСВ).

Случайная величина называется ***непрерывной***, если ее функция распределения непрерывна на всей числовой оси и дифференцируема кроме, быть может, конечного числа точек. Из этого определения и формулы (4.5) следует $P(X=x_0)=0$, т.е. *для НСВ X вероятность того, что она примет одно, заданное определенное значение равна нулю*. Поэтому для НСВ справедливы равенства:

$$P(x_1\le X\le x_2)=P(x_1<X\le x_2)=P(x_1\le X<x_2)=P(x_1<X<x_2). \qquad (4.8)$$

Очевидно, что с учетом вышеизложенного, описание НСВ с помощью ряда распределения теряет смысл.

Однако, для НСВ существует неотрицательная функция $p(x)$, удовлетворяющая при любых x равенству:

$$p(x)=\lim_{\Delta x\to 0}\frac{P(x<X<x+\Delta x)}{\Delta x} \qquad (4.9)$$

Функция $p(x)$ называется ***плотностью распределения вероятностей.*** График функции $p(x)$ называется кривой распределения.

Свойства плотности распределения вероятностей.

1) $$F\left(x\right)=P\left(X<x\right)=\int_{-\infty}^{x}p(\tau)d\tau \qquad (4.10)$$

Формула (4.10) определяет, так называемую, интегральную связь между функциями $p(x)$ и $F(x)$. Функцию $F\left(x\right)$ иногда называют ***интегральной функцией распределения*** или интегральным законом распределения.

2) Следовательно, если функция $p(x)$ непрерывна в точке x, то функция распределения $F(x)$ дифференцируема в этой точке, причем

$$p(x)=F'\left(x\right) \qquad (4.11)$$

Действительно, из формулы (4.9) следует

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

Формула (4.11) определяет, так называемую, дифференциальную связь между функциями $F(x)$ и $p(x)$. Следует отметить, что плотность распределения $p(x)$ называют также ***дифференциальной функцией распределения***.

3) Непосредственно из определения (4.9) или из (4.11) с учетом того, что $F(x)$ – неубывающая функция вытекает $p(x)\geq 0$ при всех x.

4) $$\int_{-\infty}^{+\infty}p(x)dx=1 \qquad (4.12)$$

Заметим, что (4.12) для НСВ является аналогом формулы (4.1) для ДСВ.

Геометрически свойства 3) и 4) означают, что график плотности распределения лежит не ниже оси Ox и площадь под графиком плотности равна 1.

5) Вероятности попадания НСВ X в интервал, отрезок или полуинтервал с одними и теми же концами одинаковы и равны определенному интегралу от плотности вероятности на этом промежутке:

$$P\left(a\leq X\leq b\right)=P\left(a<X\leq b\right)=P\left(a\leq X<b\right)=P\left(a<X<b\right)=\int_{a}^{b}p(x)dx \qquad (4.13)$$

Из (4.13) следует, что геометрически вероятность $P\left(a\leq X\leq b\right)$ представляет собой площадь криволинейной трапеции, ограниченной графиком плотности вероятности и отрезками прямых $y=0,\ x=a,\ x=b$.

Тема 5

Математические операции над дискретными случайными величинами

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

Назовем ***произведением*** kX ***случайной величины X на постоянную величину*** k случайную величину, которая принимает значения kx_i с теми же вероятностями p_i $(i=1,2,...,n)$, что и случайная величина X.

m***-ой степенью случайной величины X,*** т.е. X^m**,** называется случайная величина, которая принимает значения x_i^m с теми же вероятностями p_i $(i=1,2,...,n)$.

Пример 1. Дана случайная величина X

x_i	-3	1	3
p_i	0,5	0,4	0,1

Найти закон распределения случайных величин $Y=2X$ и $Z=X^2$

Решение.

Y:

y_i	-6	2	6
p_i	0,5	0,4	0,1

Z:

z_i	1	9
p_i	0,4	0,6

Назовем ***произведением (суммой или разностью) случайных величин X и Y*** случайную величину, возможные значения которой имеют вид $x_i \cdot y_j$ ($x_i + y_j$ или $x_i - y_j$), $i=1,2,...,n$; $j=1,2,...,m$ с вероятностями p_{ij} того, что случайная величина X примет значение x_i, а Y – значение y_j : $p_{ij} = P\big((X=x_i)(Y=y_j)\big)$.

Если случайные величины X *и* Y независимы, т.е. независимы любые события $X=x_i$ и $Y=y_j$, то по теореме умножения вероятностей для независимых событий

$$p_{ij} = P\big(X=x_i\big)\cdot P\big(Y=y_j\big) = p_i \cdot p_j \qquad (5.1)$$

Пример 2. Даны законы распределения двух независимых случайных величин

X:

x_i	1	3
p_i	0,4	0,6

Y:

y_j	-1	0	1
p_j	0,5	0,4	0,1

Найти закон распределения случайных величин $Z = X - Y$ и $U = XY$.

Решение.

Z:

x_i \ y_j	p_i \ p_j	-1	0	1
		0,5	0,4	0,1
1	0,4	2 / 0,2	1 / 0,16	0 / 0,04
3	0,6	4 / 0,3	3 / 0,24	2 / 0,06

Таким образом, закон распределения $Z = X - Y$ имеет вид:

z_k	0	1	2	3	4
p_k	0,04	0,16	0,26	0,24	0,3

U:

x_i \ y_j	p_i \ p_j	-1	0	1
		0,5	0,4	0,1
1	0,4	-1 / 0,2	0 / 0,16	1 / 0,04
3	0,6	-3 / 0,3	0 / 0,24	3 / 0,06

Таким образом, закон распределения $U = XY$ имеет вид:

u_k	-3	-1	0	1	3
p_k	0,3	0,2	0,4	0,04	0,06

Основные числовые характеристики дискретных и непрерывных случайных величин: математическое ожидание, дисперсия и среднее квадратическое отклонение. Их свойства и примеры

Закон распределения (функция распределения и ряд распределения или плотность вероятности) полностью описывают поведение случайной величины. Однако такой закон распределения бывает трудно обозримым, не всегда удобным и даже необходимым для анализа. В ряде задач достаточно знать некоторые *числовые характеристики* исследуемой величины (например, ее среднее значение и возможное отклонение от него), чтобы ответить на интересующий вопрос. Рассмотрим основные числовые характеристики дискретных и непрерывных случайных величин.

Начнем с примера.

Пример 3. Известны законы распределения случайных величин X и Y – числа очков, выбиваемых первым и вторым стрелками.

x_i	0	1	2	3	4	5	6	7	8	9	10
p_i	0,1	0,1	0,05	0,03	0,05	0,12	0,11	0,04	0,05	0,13	0,22

y_j	0	1	2	3	4	5	6	7	8	9	10
p_j	0,01	0,02	0,06	0,09	0,1	0,25	0,2	0,1	0,1	0,05	0,02

Необходимо выяснить, какой из двух стрелков стреляет лучше.

Изучая ряды распределения случайных величин X и Y, ответить на этот вопрос далеко не просто из-за обилия числовых значений. К тому же у первого стрелка достаточно большие вероятности имеют крайние значения числа выбиваемых очков, а у второго стрелка – промежуточные значения. Очевидно, что из двух стрелков лучше стреляет тот, кто в *среднем* выбивает большее количество очков. Таким средним значением СВ является ее математическое ожидание.

Математическое ожидание. Математическое ожидание ДСВ

Математическим ожиданием ДСВ называется сумма произведений ее возможных значений на соответствующие им вероятности:

$$M(X) = x_1 p_1 + x_2 p_2 + \ldots + x_n p_n = \sum_{i=1}^{n} x_i p_i \qquad (5.2)$$

Если число возможных значений случайной величины бесконечно, но счетно, то $M(X) = \sum_{i=1}^{\infty} x_i p_i$, если полученный ряд сходится абсолютно. Так как данный ряд может и расходиться, то соответствующая СВ может и не иметь математического ожидания.

Пример 4. Ряд распределения СВ X имеет вид:

x_i	2	2^2	2^3	…	2^i	…
p_i	$1/2$	$1/2^2$	$1/2^3$	…	$1/2^i$	…

$$M(X)=\sum_{i=1}^{\infty}x_i p_i=\sum_{i=1}^{\infty}2^i\ 1/2^i=\sum_{i=1}^{\infty}1=\infty .$$

На практике, как правило, множество возможных значений случайной величины принадлежат лишь ограниченному участку оси абсцисс и, значит, математическое ожидание существует.

Пример 5. Найдем математическое ожидание случайных величин X и Y из примера 3.

Решение.

$$M(X)=\sum_{i=1}^{11}x_i p_i=0\cdot 0{,}1+1\cdot 0{,}1+\ldots 9\cdot 0{,}13+10\cdot 0{,}22=5{,}8$$

$$M(Y)=\sum_{j=1}^{11}y_j p_j=0\cdot 0{,}01+1\cdot 0{,}02+\ldots 9\cdot 0{,}05+10\cdot 0{,}02=5{,}41$$

Таким образом, сравнивая $M(X)$ и $M(Y)$, можно утверждать, что первый стрелок в среднем стреляет лучше второго.

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

Пример 6. Найдем математическое ожидание случайной величины X – числа стандартных деталей среди трех, отобранных из партии в 10 деталей, среди которых 2 бракованных.

Решение. Составим ряд распределения для X. Из условия задачи следует, что X может принимать значения 1, 2, 3.

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

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

Свойства математического ожидания.

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

Действительно, обозначая a и b наименьшее и наибольшее значение среди $x_1, x_2,\ldots,x_n$, из (5.2) имеем:

$$a(p_1+p_2+\ldots+p_n)\le M(X)=x_1p_1+x_2p_2+\ldots+x_np_n\le b(p_1+p_2+\ldots+p_n)$$

Учитывая, что $p_1+p_2+\ldots+p_n=1$, получаем $a\le M(X)\le b$.

2) Математическое ожидание постоянной равно самой постоянной:

$$M(C)=C. \qquad (5.3)$$

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

3) Постоянный множитель можно выносить за знак математического ожидания:

$$M(CX) = C\,M(X). \qquad (5.4)$$

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

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

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

$$M(XY) = M(X)M(Y). \qquad (5.5)$$

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

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

Замечание 2. Свойство 4 справедливо для произведения любого числа независимых случайных величин, что доказывается методом математической индукции.

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

$$M(X + Y) = M(X) + M(Y). \qquad (5.6)$$

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 сумма любого числа случайных величин равна сумме математических ожиданий слагаемых.

Пример 7. Найти математическое ожидание суммы числа очков, выпавших при броске пяти игральных костей.

Решение. 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 5) $M(X_1 + X_2 + ... + X_5) = 5\cdot\frac{7}{2} = \frac{35}{2}$.

Математическое ожидание НСВ

Введем понятие математического ожидания для непрерывной случайной величины. Пусть НСВ X, определяемая плотностью распределения $p(x)$, принимает значения, принадлежащие отрезку $[a,b]$. Отрезок $[a,b]$ разобьем на n элементарных отрезков $[a,x_1],[x_1,x_2],...,[x_{n-1},b]$, длины которых выражаются формулой $\Delta x_i = x_i - x_{i-1},\ i = 1,2,...,n,$

$x_0 = a,\ x_n = b$. В каждом из элементарных отрезков $[x_{i-1},x_i]$ выберем произвольную точку ξ_i и составим произведение $p(\xi_i)\Delta x_i$, которое по формуле (4.9) приближенно выражается вероятностью p_i попадания значений X в интервал $[x_{i-1},x_i]$.

Предположив дать определение понятия математического ожидания для НСВ по аналогии с указанным понятием для ДСВ по форме (5.2), имеем

$$M(X) = \sum_{i=1}^{n}\xi_i p_i \approx \sum_{i=1}^{n}\xi_i p(\xi_i)\Delta x_i$$

Переходя к пределу в последней сумме (интегральной сумме Римана) при $\max\Delta x_i \to 0$, получаем

$$\lim_{\max\Delta x_i\to 0}\sum_{i=1}^{n}\xi_i p(\xi_i)\Delta x_i = \int_a^b xp(x)dx$$

Этот определенный интеграл называют математическим ожиданием рассматриваемой НСВ X, т.е. по определению полагают

$$M(X) = \int_a^b xp(x)dx \qquad (5.7)$$

Если значения НСВ X принадлежат бесконечному интервалу $(-\infty,+\infty)$, то ее математическое ожидание определяется формулой

$$M(X) = \int_{-\infty}^{+\infty} xp(x)dx \qquad (5.8)$$

при условии, что несобственный интеграл первого рода сходится абсолютно.
Замечание. Математическое ожидание НСВ обладает теми же свойствами, что и математическое ожидание ДСВ.

Разность $X - M(X)$ называется *отклонением* случайной величины X от ее математического ожидания $M(X)$. Отклонение является случайной величиной. Покажем, что математическое ожидание отклонения равно нулю. Действительно,

$$M(X - M(X)) = M(X) - M(M(X)) = M(X) - M(X) = 0 \qquad (5.9)$$

Это равенство объясняется тем, что отклонения могут быть как положительными, так и отрицательными; в результате их взаимного погашения среднее значение отклонения равно нулю.

Дисперсия

Для того чтобы иметь представление о поведении случайной величины, недостаточно знать только ее математическое ожидание: зная математическое ожидание, нельзя сказать, какие значения принимает случайная величина и как они отклоняются от среднего значения. Чтобы знать, как рассеяны значения случайной величины вокруг ее математического ожидания, вводят другую числовую характеристику, называемую дисперсией. Из равенства (5.9) видно, что за меру рассеяния нельзя принять отклонение СВ от ее математического ожидания. Вследствие этого рассматривают их квадраты.

Пример 8. Рассмотрим две случайные величины: X и Y, заданные рядами распределения вида

x_i	49	50	51
p_i	0,1	0,8	0,1

y_j	0	100
p_j	0,5	0,5

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 $M(X)$ 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 математического ожидания:

$$D(X) = M(X - M(X))^2. \qquad (5.10)$$

Пример 9. Найдем дисперсию случайной величины X (числа стандартных деталей среди отобранных) в примере 6 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. Из определения дисперсии следует, что эта величина принимает только неотрицательные значения.

Для вычисления дисперсии существует также формула:

$$D(X) = M(X^2) - M^2(X). \qquad (5.11)$$

Доказательство. Используя то, что $M(X)$ – постоянная величина, и свойства математического ожидания, преобразуем формулу (5.10) к виду:

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

Пример 10. Вычислим дисперсии случайных величин X и Y, рассмотренных в примере 9.

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

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

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:

$$D(C) = 0. \qquad (5.12)$$

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:

$$D(CX) = C^2 D(X). \qquad (5.13)$$

Доказательство.

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

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

$$D(X + Y) = D(X) + D(Y). \qquad (5.14)$$

Доказательство. По формуле (5.11)

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

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

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:

$$D(X - Y) = D(X) + D(Y). \qquad (5.15)$$

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 отклонением.

Средним квадратическим отклонением σ случайной величины X называется квадратный корень из дисперсии:

$$\sigma=\sqrt{D(X)}\,. \qquad (5.16)$$

Очевидно, размерность дисперсии равна квадрату размерности случайной величины, поэтому среднее квадратическое отклонение имеет ту же размерность, что и СВ X. Среднее квадратическое отклонение применяется тогда, когда желательно получить оценку рассеяния СВ в тех же единицах, в которых выражены значения самой величины.

Пример 11. В примере 11 средние квадратические отклонения X и Y равны соответственно $\sigma_x=\sqrt{0{,}2}\approx 0{,}447;\quad \sigma_y=\sqrt{2500}=50.$

Замечание. Общее определение дисперсии дано как для ДСВ так и для НСВ, однако с учетом специфики вычисления математического ожидания для дискретного и непрерывного случаев приведем формулы для расчета дисперсии для ДСВ:

$$D(X)=\sum_{i=1}^{n}\big(x_i-M(X)\big)^2\,p_i \qquad (5.17)$$

или

$$D(X)=\sum_{i=1}^{n}x_i^{\,2}p_i-(M(X))^2 \qquad (5.18)$$

и НСВ:

$$D(X)=\int_{-\infty}^{+\infty}(x-M(X))^2\,p(x)dx \qquad (5.19)$$

или

$$D(X)=\int_{-\infty}^{+\infty}x^2p(x)dx-M^2(X) \qquad (5.20)$$

Замечание. Если все возможные значения непрерывной случайной величины не выходят за пределы интервала [*a, b*], то интегралы в формулах (5.19) и (5.20) вычисляются в этих пределах.

Пример 12. Плотность распределения случайной величины X имеет вид:

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

Найти $M(X)$, $D(X)$, σ.

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

$$D(X) = -\frac{3}{4}\int_2^4 x^2(x^2 - 6x + 8)dx - 9 = -\frac{3}{4}\left(\frac{x^5}{5} - \frac{3x^4}{2} + \frac{8x^3}{3}\right)\Bigg|_2^4 - 9 = 9,2 - 9 = 0,2; \quad \sigma = \sqrt{0,2} \approx 0,447.$$

Интерпретация математического ожидания и дисперсии в финансовом анализе

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

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

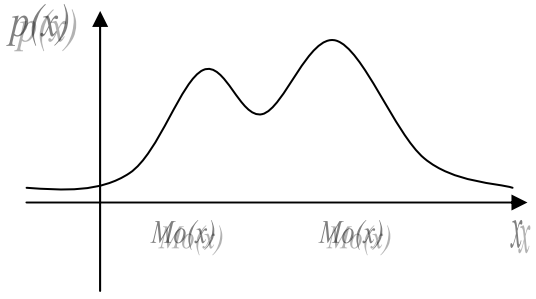
В теории вероятностей числовые характеристики играют большую роль. Часто удается решать вероятностные задачи, оперируя лишь числовыми характеристиками СВ. Применение вероятностных методов для решения практических задач в значительной мере определяется умением пользоваться числовыми характеристиками СВ, оставляя в стороне законы распределения.

Мода и медиана. Квантили. Моменты случайных величин. Асимметрия и эксцесс

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

Модой $Mo(X)$ случайной величины X называется ее наиболее вероятное значение (для которого вероятность p_i или плотность $p(x)$ достигает максимума).

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



Медианой $Me(X)$ непрерывной случайной величины X называется такое ее значение, для которого справедливо равенство

$$P(x < Me(X)) = P(x > Me(X)) = \frac{1}{2} \qquad (5.21)$$

Очевидно равенство $P(x < Me(X)) = F(x = Me(X)) = \frac{1}{2}$.

Квантилем уровня q (или q***-квантилем***) называется такое значение x_q случайной величины, при котором функция ее распределения принимает значение, равное q, т.е.

$$F(x_q) = P(x < x_q) = q \qquad (5.22)$$

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

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

Начальным ***моментом*** k***-го порядка*** СВ X называется математическое ожидание k-ой степени этой величины:

$$\nu_k = M(X^k) \qquad (5.23)$$

Центральным моментом k***-го порядка*** СВ X называется математическое ожидание k-ой степени отклонения СВ X от ее математического ожидания.

$$\mu_k = M(X - M(X))^k \qquad (5.24)$$

При $k = 1$ первый начальный момент СВ X есть ее математическое ожидание, т.е. $\nu_1 = M(X)$; при $k = 2$ второй центральный момент–дисперсия, т.е. $\mu_2 = D(X)$.

Центральные моменты μ_k могут быть выражены через начальные моменты ν_k по формулам:

$\mu_1 = 0$

$\mu_2 = \nu_2 - \nu_1^2$

$\mu_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3$

$\mu_4 = \nu_4 - 4\nu_1\nu_3 + 6\nu_1^2\nu_2 - 3\nu_1^4$ и т.д.

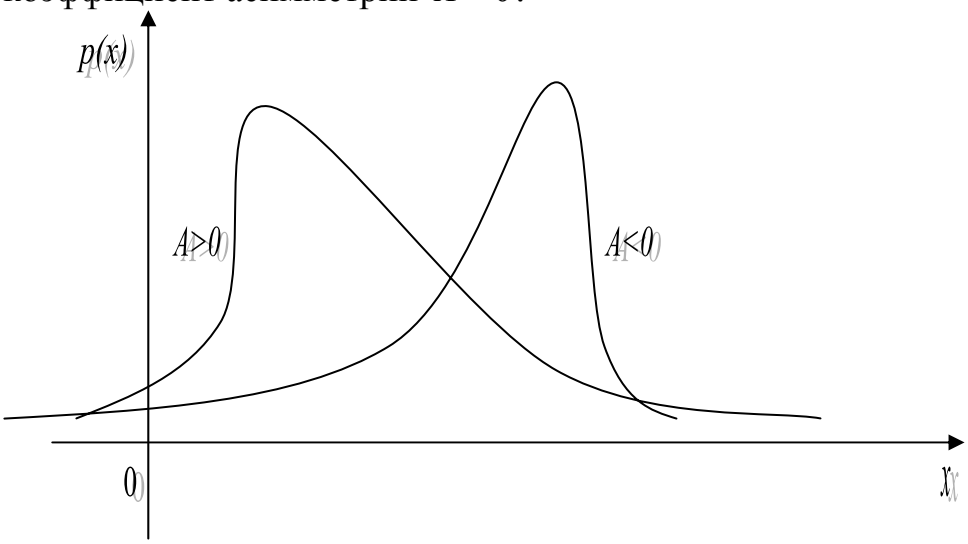
Например,

$$\mu_3 = M(X - M(X))^3 = M(X^3 - 3X^2M(X) + 3XM(X)^2 - M(X)^3) =$$
$$= M(X^3 - 3X^2\nu_1 + 3X\nu_1^2 - \nu_1^3) = M(X^3) - 3M(X^2)\nu_1 + 3M(X)\nu_1^2 - \nu_1^3 =$$
$$= \nu_3 - 3\nu_1\nu_2 + 3\nu_1^3 - \nu_1^3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3$$

Третий центральный момент μ_3 служит для характеристики ***асимметрии*** (***скошенности***) распределения. Он имеет размерность куба СВ. Чтобы получить безразмерную величину, ее делят на σ^3. Полученная величина A называется *коэффициентом асимметрии* СВ:

$$A = \frac{\mu_3}{\sigma^3} \qquad (5.25)$$

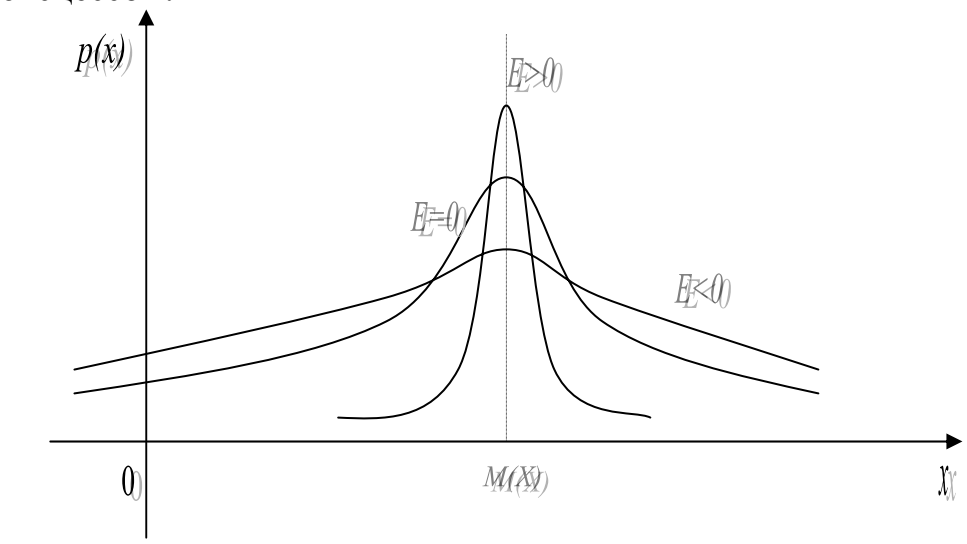
Если распределение симметрично относительно математического ожидания, то коэффициент асимметрии $A = 0$.



Четвертый центральный момент μ_4 служит для характеристики *крутости* (*островершинности* или *плосковершинности*) распределения.
Эксцессом СВ называется число

$$E = \frac{\mu_4}{\sigma^4} - 3 \qquad (5.26)$$

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



Тема 6

Часто на практике мы имеем дело со случайными величинами, распределенными определенным типовым образом, то есть такими, закон распределения которых имеет некоторую стандартную форму. В данной лекции рассматриваются основные законы распределения дискретных и непрерывных случайных величин, используемых для построения теоретико-вероятностных моделей реальных социально-экономических явлений.

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,…,m,…,n. Соответствующие им вероятности можно вычислить по формуле Бернулли:

$$p(X=m)=C_n^m p^m q^{n-m} \qquad (6.1)$$

(p – вероятность появления A в каждом испытании).

Ряд распределения биномиального закона имеет вид:

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

Очевидно, что определение биномиального закона распределения корректно, так как основное свойство ряда распределения (4.1) выполнено, ибо $\sum_{i=0}^{n} p_i$ есть не что иное, как сумма всех членов разложения бинома Ньютона (отсюда и название закона – ***биномиальный***):

$$C_n^0 q^n + C_n^1 p^1 q^{n-1} \ldots + C_n^m p^m q^{n-m} + \ldots + C_n^{n-1} p^{n-1} q^1 + C_n^n p^n = (q+p)^n = 1^n = 1. \qquad (6.2)$$

Теорема 1. Математическое ожидание и дисперсия случайной величины X, распределенной по биномиальному закону, вычисляются соответственно по формулам:

$$M(X)=np \qquad (6.3)$$

$$D(X)=npq \qquad (6.4)$$

Доказательство. СВ X –число m наступлений события A в n независимых испытаниях–можно представить в виде суммы n независимых величин $X=X_1+X_2+\ldots+X_n=\sum_{k=1}^{n} X_k$, каждая из которых имеет один и тот же закон распределения:

x_i	0	1
p_i	q	p

Случайная величина X_k, которую называют индикатором события A, выражает число наступлений события A в k-м испытании $k=1,2,\ldots,n$, т.е. при наступлении события A $X_k=1$ с вероятностью p, при ненаступлении A $X_k=0$ с вероятностью q. Найдем числовые характеристики индикатора события A по формулам (5.2), (5.18):

$$M(X_k) = x_1 p_1 + x_2 p_2 = 1 \cdot p + 0 \cdot q = p \qquad (6.3^*)$$

$$D(X_k) = x_1^2 p_1 + x_2^2 p_2 - (M(X_k))^2 = 1^2 \cdot p + 0^2 \cdot q - p^2 = p(1-p) = pq \qquad (6.4^*)$$

Таким образом, математическое ожидание и дисперсия рассматриваемой СВ X:

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

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

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

Из определения моды и формулы (3.3) вытекает, что мода случайной величины, распределенной по биномиальному закону, является целым числом, которое находится по формуле:

$$np - q \le Mo(X) \le np + p \qquad (6.5)$$

Пример 1. Составить ряд распределения, найти математическое ожидание и дисперсию случайной величины X – числа попаданий при 5 выстрелах, если вероятность попадания при одном выстреле равна 0,8.

Решение.

P(X=0)=1·(0,2)^5 = 0,00032; P(X=1) = 5·0,8·(0,2)^4 = 0,0064;
P(X=2) = 10·(0,8)^2·(0,2)^3 = 0,0512; P(X=3) = 10·(0,8)^3·(0,2)^2 = 0,2048;
P(X=4) = 5·(0,8)^4·0,2 = 0,4096; P(X=5) = 1·(0,8)^5 = 0,32768.

Таким образом, ряд распределения имеет вид:

x_i	0	1	2	3	4	5
p_i	0.00032	0.0064	0.0512	0.2048	0.4096	0.32768

По формуле (6.3) математическое ожидание случайной величины X равняется $M(X) = np = 5 \cdot 0,8 = 4$, по формуле (6.4) вычисляется дисперсия $D(X) = npq = 5 \cdot 0,8 \cdot 0,2 = 0,8$.

Распределение Пуассона

Рассмотрим дискретную случайную величину X, принимающую только целые неотрицательные значения (0, 1, 2,…, m,…), последовательность которых бесконечна, но счетна. Такая случайная величина называется распределенной ***по закону Пуассона***, если вероятность того, что она примет значение m, выражается формулой Пуассона (3.5):

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

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

Ряд распределения закона Пуассона имеет вид:

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

Очевидно, что определение закона Пуассона корректно, так как основное свойство ряда распределения (4.1) выполнено, ибо сумма всех вероятностей равна 1:

$$\sum_{m=0}^{\infty} p(X=m) = e^{-\lambda} + \lambda e^{-\lambda} + \frac{\lambda^2 e^{-\lambda}}{2!} + \frac{\lambda^m e^{-\lambda}}{m!} + \ldots = e^{-\lambda} \sum_{m=0}^{\infty} \frac{\lambda^m}{m!} = e^{-\lambda} \cdot e^{\lambda} = 1$$

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

Замечание. В лекции 3 говорилось о том, что формула Пуассона выражает биномиальное распределение при большом числе опытов и малой вероятности события. Поэтому закон Пуассона часто называют *законом редких явлений*.

Теорема 2. Математическое ожидание и дисперсия случайной величины *X,* распределенной по закону Пуассона, вычисляются соответственно по формулам:

$$M(X) = \lambda \qquad (6.6)$$

$$D(X) = \lambda \qquad (6.7)$$

Доказательство.

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

Дисперсию найдем по формуле (5.18). Для этого найдем

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

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

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

Теорема 3. Сумма двух независимых случайных величин, подчиняющихся распределению Пуассона с параметрами λ_1 и λ_2, также имеет распределение Пуассона с параметром $\lambda_1 + \lambda_2$.

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

Геометрическое распределение

ДСВ X, принимающую только целые положительные значения (1, 2,…, m,…), последовательность которых бесконечна, но счетна, имеет геометрическое распределение, если вероятность того, что она примет значение m, выражается формулой:

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

Ряд распределения геометрического закона имеет вид:

x_i	1	2	…	m	…
p_i	p	pq	…	pq^{m-1}	…

Покажем, что определение геометрического закона корректно:

$\sum_{i=1}^{\infty} p_i = p + pq + pq^2 + ... + pq^{m-1} + ... = \frac{p}{1-q} = \frac{p}{p} = 1$, здесь использована формула $S = \frac{b_1}{1-q}$ – суммы бесконечной убывающей геометрической прогрессии.

Теорема 4. Математическое ожидание и дисперсия случайной величины X, распределенной по геометрическому закону, вычисляются соответственно по формулам:

$$M(X) = \frac{1}{p} \qquad (6.8)$$

$$D(X) = \frac{q}{p^2} \qquad (6.9)$$

Пример 2. Определить математическое ожидание и дисперсию случайной величины X – числа бросков монеты до первого появления герба. Эта величина может принимать бесконечное число значений (множество возможных значений есть множество натуральных чисел).

Решение.

Ряд ее распределения имеет вид:

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

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

Гипергеометрическое распределение

Пусть имеется N элементов, из которых M элементов обладают некоторым признаком A. Извлекаются случайным образом без возвращения n элементов. X – дискретная случайная величина, число элементов обладающих признаком A, среди отобранных n элементов. Вероятность, что $X=m$, где $m = 0,1,2,...,\min\{n,M\}$, определяется по формуле

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

Теорема 5. Математическое ожидание и дисперсия случайной величины, распределенной по гипергеометрическому закону, определяются формулами:

$$M(X)=n\frac{M}{N}, \qquad (6.10)$$

$$D(X)=n\frac{M}{N-1}\left(1-\frac{M}{N}\right)\left(1-\frac{n}{N}\right). \qquad (6.11)$$

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

Решение.

Число угаданных видов спорта в лотерее «6 из 36» есть случайная величина, имеющая гипергеометрическое распределение с параметрами $n=6,\ M=6,\ N=36$. Ряд распределения СВ X имеет вид

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

Вероятность получения денежного приза

$$P(3\le X\le 6)=\sum_{i=3}^{6}P(X=i)=\frac{1450}{34782}+\frac{2175}{649264}+\frac{15}{162316}+\frac{1}{1947792}\approx 0{,}043$$

По формулам (6.10) и (6.11) найдем:

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

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

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

Равномерный закон распределения

Выше были рассмотрены примеры таких законов распределения для дискретных случайных величин. Для непрерывных случайных величин тоже существуют часто встречающиеся виды законов распределения, и в качестве первого из них рассмотрим равномерный закон.

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

$$p(x)=\begin{cases} c \ \text{ при } a\le x\le b, \\ 0 \text{ при } \ x<a, \quad x>b. \end{cases} \tag{6.12}$$

Из условия нормировки следует, что $\int_a^b f(x)dx=\int_a^b cdx=c(b-a)=1,$ откуда $f(x)=c=\frac{1}{b-a}$, таким образом, (6.12) имеет вид

$$p(x)=\begin{cases} \dfrac{1}{b-a} & \text{при } a\le x\le b, \\ 0 & \text{при } \ x<a, \quad x>b. \end{cases} \tag{6.13}$$

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

$$P(\alpha<X<\beta)=\int_\alpha^\beta \frac{1}{b-a}dx=\frac{\beta-\alpha}{b-a}. \tag{6.14}$$

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

$$F(x)=\begin{cases} 0 & \text{при } \ x\le a, \\ \dfrac{x-a}{b-a} & \text{при } a<x\le b, \\ 1 & \text{при } \ x>b. \end{cases} \tag{6.15}$$

Действительно,

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

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

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

Математическое ожидание $M(X)=\frac{a+b}{2}$, (6.16)

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

дисперсия $D(X)=\frac{(b-a)^2}{12}$, (6.17)

По формуле (5.20) имеем

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

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

а среднее квадратическое отклонение $\sigma(X) = \frac{b-a}{2\sqrt{3}}$. (6.18)

Пример 4. 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 по формуле (6.14) $p(0 \le x \le 2) = \frac{2}{5} = 0,4.$

Показательный (экспоненциальный) закон распределения

Непрерывная случайная величина X имеет ***показательный (экспоненциальный)*** закон распределения с параметром λ, если ее плотность вероятности имеет вид:

$$p(x) = \begin{cases} 0 & \text{при } x < 0, \\ \lambda e^{-\lambda x} & \text{при } x \ge 0. \end{cases} \quad (6.19)$$

Функция распределения случайной величины, распределенной по показательному закону, равна

$$F(x) = \begin{cases} 0 & \text{при } x < 0 \\ 1 - e^{-\lambda x} & \text{при } x \ge 0. \end{cases} \quad (6.20)$$

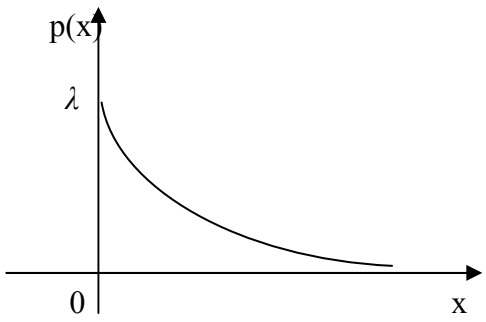
Действительно,

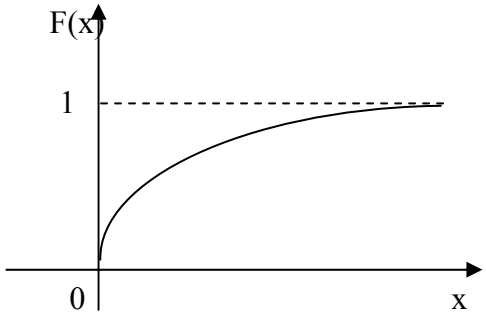
при $x \le 0$ $F(x) = \int_{-\infty}^{x} p(\tau)d\tau = \int_{-\infty}^{x} 0 d\tau = 0$

при $x > 0$

$$F(x) = \int_{-\infty}^{x} p(\tau)d\tau = \int_{-\infty}^{0} 0 d\tau + \int_{0}^{x} \lambda e^{-\lambda\tau} d\tau = -\int_{0}^{x} e^{-\lambda\tau} d(-\lambda\tau) = -\int_{0}^{x} d(e^{-\lambda\tau}) = -e^{-\lambda\tau}\Big|_0^x = -e^{-\lambda x} + 1$$

Кривая распределения *p(x)* и график функции распределения $F(x)$ приведены ниже:





Для случайной величины, распределенной по показательному закону

$$M(X)=\frac{1}{\lambda}; \qquad (6.21)$$

По формуле (5.8) имеем:

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

$$=-\lim_{b\to+\infty}\int_{0}^{b} xd(e^{-\lambda x})=-\lim_{b\to+\infty}\left(xe^{-\lambda x}\Big|_{0}^{b}-\int_{0}^{b} e^{-\lambda x}dx\right)=-\lim_{b\to+\infty}\left(be^{-b\lambda}-0e^{-0\lambda}+\frac{1}{\lambda}\int_{0}^{b} d(e^{-\lambda x})\right)=$$

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

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

$$D(X)=\frac{1}{\lambda^2}, \qquad \sigma(X)=\frac{1}{\lambda}. \qquad (6.22)$$

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

$$P(a<X<b)=e^{-\lambda a}-e^{-\lambda b}. \qquad (6.22)$$

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

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

$$F(t)=P(T<t)=1-e^{-\lambda t} \quad (\lambda>0) \qquad (6.23)$$

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

Функция надежности

$$R(t)=e^{-\lambda t} \qquad (6.24)$$

определяет вероятность безотказной работы элемента за время длительностью t.

Пример 5. Установлено, что время ремонта магнитофонов есть случайная величина X, распределенная по показательному закону. Определить вероятность того, что на ремонт магнитофона потребуется не менее 15 дней, если среднее время ремонта магнитофонов составляет 12 дней. Найти плотность вероятности, функцию распределения и среднее квадратическое отклонение случайной величины X.

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

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

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

Пример 6. Испытывают три элемента, которые работают независимо один от другого. Длительность времени безотказной работы элементов распределена по показательному закону: для первого элемента $F_1(t)=1-e^{-0,1t}$; для второго $F_2(t)=1-e^{-0,2t}$; для третьего элемента $F_3(t)=1-e^{-0,3t}$. Найти вероятности того, что в интервале времени $(0;5)$ч. откажут: а) только один элемент; б) только два элемента; в) все три элемента.

Решение. Вероятность отказа первого элемента

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

Вероятность отказа второго элемента

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

Вероятность отказа третьего элемента

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

Искомая вероятность

а) $P=p_1q_2q_3+q_1p_2q_3+q_1q_2p_3=0{,}034+0{,}084+0{,}1749=0{,}2929$.

б) $P=p_1p_2q_3+p_1q_2p_3+q_1p_2p_3=0{,}057+0{,}1187+0{,}2925=0{,}4682$.

в) $P=p_1p_2p_3=0{,}1985$.

Нормальный закон распределения

В теории вероятностей и математической статистике важнейшую роль играет так называемое нормальное или гауссовское распределение. Оно также широко применяется и при решении прикладных задач. Значимость нормального распределения определяется тем, что оно служит хорошим приближением для большого числа наборов случайных величин, получаемых при наблюдениях и экспериментах. Нормальное распределение почти всегда имеет место, когда наблюдаемые случайные величины формируются под влиянием большого числа случайных факторов, ни один из которых существенно не превосходит остальные.

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

$$p(x)=\frac{1}{\sigma\sqrt{2\pi}}\cdot e^{-\frac{(x-a)^2}{2\sigma^2}},\ -\infty<x<+\infty. \qquad (6.25)$$

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

Для изучения вида этой кривой методами дифференциального исчисления найдем точки экстремума и точки перегиба.

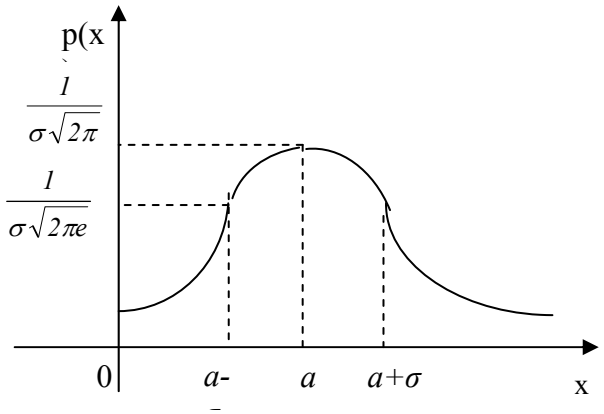
$$p'(x)=-\frac{2(x-a)}{2\sigma^2}\frac{1}{\sigma\sqrt{2\pi}}\cdot e^{-\frac{(x-a)^2}{2\sigma^2}},\ -\infty<x<+\infty$$

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

Т.к. первая производная обращается в 0 при $x=a$ и меняет знак при переходе через эту точку с «+» на «-», то в точке $x=a$ функция (6.25) принимает максимальное значение, равное $p_{\max}(a)=\frac{1}{\sigma\sqrt{2\pi}}$.

Т.к. вторая производная обращается в 0 при $x=a\pm\sigma$ и меняет знак при переходе через эти точки, то в точках $\left(a+\sigma,\frac{1}{\sigma\sqrt{2\pi e}}\right)$ и $\left(a-\sigma,\frac{1}{\sigma\sqrt{2\pi e}}\right)$ функция (6.25) меняет направление выпуклости.

Ниже приведена нормальная кривая *p(x)* с параметрами a и σ^2, т.е. $N(a;\sigma^2)$:



Для случайной величины, распределенной по нормальному закону,

$$M(X)=a, \qquad (6.26)$$

$$D(X)=\sigma^2. \qquad (6.27)$$

Выясним как будет меняться нормальная кривая при изменении параметров a и σ. Если $\sigma=const$ и меняется параметр a –центр симметрии

распределения, то нормальная кривая будет смещаться вдоль оси абсцисс, не меняя формы.

Если $a = const$ и меняется параметр σ–разброс значений случайной величины от центра симметрии распределения, то при увеличении σ $p_{\max}(a) = \frac{1}{\sigma\sqrt{2\pi}}$ уменьшается, но т.к. площадь под любой кривой распределения должна оставаться равной 1, то кривая становится более плоской, растягиваясь вдоль оси Ox, при уменьшении σ $p_{\max}(a) = \frac{1}{\sigma\sqrt{2\pi}}$ увеличивается и нормальная кривая вытягивается вверх, одновременно сжимаясь с боков.

Сложность непосредственного нахождения функции распределения случайной величины, распределенной по нормальному закону, по формуле (4.9) и вероятности ее попадания на некоторый промежуток по формуле (4.12) связана с тем, что интеграл от функции (6.25) не берется в элементарных функциях. Поэтому ее выражают через функцию Лапласа (интеграл вероятностей), для которой составлены таблицы.

$$\Phi(x) = \frac{2}{\sqrt{2\pi}} \cdot \int_0^x e^{\frac{-t^2}{2}} dt \qquad (6.28)$$

Функция распределения случайной величины *X*, распределенной по нормальному закону, выражается через функцию Лапласа *Φ(x)* по формуле

$$F(x) = \frac{1}{2} + \frac{1}{2}\Phi\left(\frac{x-a}{\sigma}\right) \qquad (6.29)$$

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

$$P(\alpha \le x \le \beta) = F(\beta) - F(\alpha) = \frac{1}{2}\left[\Phi\left(\frac{\beta-a}{\sigma}\right) - \Phi\left(\frac{\alpha-a}{\sigma}\right)\right]. \qquad (6.30)$$

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

$$P(|x-a| \le \delta) = P(a-\delta \le x \le a+\delta) = \frac{1}{2}\left[\Phi\left(\frac{a+\delta-a}{\sigma}\right) - \Phi\left(\frac{a-\delta-a}{\sigma}\right)\right] = \Phi\left(\frac{\delta}{\sigma}\right). \qquad (6.31)$$

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

$$P(|x-a| \le 3\sigma) = \Phi\left(\frac{3\sigma}{\sigma}\right) = \Phi(3) = 0{,}9973. \qquad (6.32)$$

Отклонение по абсолютной величине нормально распределенной СВ X больше, чем на 3σ, является событием практически невозможным, т.к. его вероятность весьма мала:

$$P\left(|x-a|>3\sigma\right)=1-\Phi\left(\frac{3\sigma}{\sigma}\right)=1-\Phi(3)=1-0{,}9973=0{,}0027$$

Т.к. кривая Гаусса симметрична относительно математического ожидания, то коэффициент асимметрии нормального распределения $A=0$.

Эксцесс нормального распределения E=0 и крутость других распределений определяется по отношению к нормальному.

Пример 7. Определить закон распределения случайной величины X, если ее плотность распределения вероятностей задана функцией:

$$p(x)=\frac{1}{6\sqrt{2\pi}}\cdot e^{-\frac{(x-1)^2}{72}}.$$

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

Решение. Сравнивая данную функцию $p(x)$ с функцией плотности вероятности для случайной величины, распределенной по нормальному закону, заключаем, что случайная величина X распределена по нормальному закону с параметрами $a = 1$ и $\sigma=6$.

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

Функция распределения случайной величины X имеет вид:

$$F(x)=\frac{1}{2}+\frac{1}{2}\Phi\left(\frac{x-1}{6}\right). \qquad (6.33)$$

Пример 8. Текущая цена акции может быть смоделирована с помощью нормального закона распределения с математическим ожиданием 15 ден.ед. и средним квадратическим отклонением 0,2 ден. ед.

Найти вероятность того, что цена акции: а) не выше 15,3 ден. ед.; б) не ниже 15,4 ден. ед.; в) от 14,9 до 15,3 ден. ед. Найти границы, в которых будет находиться текущая цена акции.

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

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

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

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

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

$$P(14{,}9\le x\le 15{,}3)=\frac{1}{2}\left[\Phi\left(\frac{15{,}3-15}{0{,}2}\right)-\Phi\left(\frac{14{,}9-15}{0{,}2}\right)\right]=\frac{1}{2}\left[\Phi(1{,}5)+\Phi(0{,}5)\right]=$$

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

По правилу трех сигм $P\left(\left|X-15\right|\le 0{,}6\right)=0{,}9973$ и, следовательно, $15-0{,}6\le X\le 15+0{,}6$. Окончательно $14{,}4\le X\le 15{,}6$.

Пример 9. Автомат изготавливает детали, контролируя их диаметры X. Считая, что случайная величина X распределена нормально с параметрами $\sigma=0{,}1$ мм и $a=10$ мм, найти интервал, в котором с вероятностью 0,9973 будут заключены диаметры изготовленных деталей.

Решение. Найдем отклонение δ по известным вероятности отклонения и $\sigma=0{,}1$ (по формуле (6.31)):

$$\Phi\left(\frac{\delta}{\sigma}\right)=0{,}9973.$$

По таблице значений функции Лапласа находим, что $\frac{\delta}{\sigma}=3$. Следовательно, $\delta=3\sigma=0{,}3$. Из неравенства $\left|X-10\right|<0{,}3$ получаем $-0{,}3<X-10<0{,}3$ или $9{,}7<X<10{,}3$

Пример 10. Рост взрослых мужчин является случайной величиной, распределенной по нормальному закону. Пусть математическое ожидание ее равно 175 см, а среднее квадратическое отклонение – 6 см. Определить вероятность того, что хотя бы один из наудачу выбранных пяти мужчин будет иметь рост от 170 до 180 см.

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

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

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

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

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

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

Тема 6

Закон больших чисел

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

Неравенства Маркова и Чебышева

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

Теорема 1 (неравенства Маркова). *Если случайная величина X принимает только неотрицательные значения и имеет математическое ожидание, то для любого положительного числа A верны неравенства*

$$P(x > A) \le \frac{M(X)}{A} \qquad (7.1)$$

$$P(x \le A) \ge 1 - \frac{M(X)}{A} \qquad (7.2)$$

Доказательство. Пусть X задается рядом распределения, в котором ее значения x_i располагаются в порядке возрастания:

X	x_1	x_2	…	x_n
p	p_1	p_2	…	p_n

Рассмотрим три возможных случая расположения числа A и значений x_i:

1. Пусть $A < x_1$, тогда событие $x > A$ является достоверным и $P(x > A) = 1$. По свойству 1 математического ожидания следует $x_1 \le M(X) \le x_n$, откуда очевидно $A < M(X)$ и $\frac{M(X)}{A} > 1$. Поэтому неравенство (7.1) в этом случае справедливо.
2. Пусть $A > x_n$, тогда событие $x > A$ является невозможным и $P(x > A) = 0$. По свойству 1 математического ожидания следует $x_1 \le M(X) \le x_n$, откуда очевидно $A < M(X)$ и $\frac{M(X)}{A} > 1$. Поэтому неравенство (7.1) и в этом случае справедливо.
3. Пусть часть значений $x_1, x_2, ..., x_k$ будут не более числа A, а другая часть $x_{k+1}, x_{k+2}, ..., x_n$ будут больше числа A. Математическое ожидание ДСВ X вычисляется по формуле (5.2):

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

Отбрасывая первые k неотрицательных слагаемых, получим

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

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

$M(X) \geq A\left(p_{k+1} + \ldots + p_n\right)$ или $p_{k+1} + \ldots + p_n \leq \dfrac{M(X)}{A}$.

Сумма вероятностей в левой части полученного неравенства представляет собой сумму вероятностей событий $X = x_{k+1}, \ldots, X = x_n$, т.е. вероятность события $X > A$. Поэтому $P(x > A) \leq \dfrac{M(X)}{A}$.

Отметим, что события $X > A$ и $X \leq A$ противоположные, поэтому заменяя $P(x > A)$ в уже доказанном неравенстве (7.1) на $1 - P(x \leq A)$, придем к другой форме неравенства Маркова (7.2).

Теорема доказана.

Пример 1. Оценить вероятность того, что в течение ближайшего дня потребность в воде в населенном пункте превысит 150 000 л, если среднесуточная потребность в ней составляет 50 000 л.

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

Ответ: $P\left(X > 150\,000\right) \leq \dfrac{1}{3}$.

Пример 2. Среднее число солнечных дней в году для данной местности равно 90. Оценить вероятность того, что в течение года в этой местности будет не более 240 солнечных дней.

Решение. Согласно неравенству $P\left(X \leq A\right) \geq 1 - \dfrac{M\left(X\right)}{A}$, имеем

$P\left(X \leq 240\right) \geq 1 - \dfrac{90}{240} = 1 - 0{,}375 = 0{,}625$.

Ответ: $P\left(X \leq 240\right) \geq 0{,}625$.

Теорема 2 (неравенства Чебышева). *Для любой случайной величины X, имеющей математическое ожидание и дисперсию, справедливы неравенства*

$$P\left(\left|X - M(X)\right| > \varepsilon\right) \leq \frac{D(X)}{\varepsilon^2}, \qquad (7.3)$$

$$P\left(\left|X - M(X)\right| \leq \varepsilon\right) \geq 1 - \frac{D(X)}{\varepsilon^2}, \qquad (7.4)$$

где $\varepsilon > 0$.

Доказательство. Применим неравенство Маркова (7.1) к случайной величине $\tilde{X} = \left(X - M(X)\right)^2$, выбрав в качестве положительного числа $A = \varepsilon^2$:

$$P(\left(X - M(X)\right)^2 > \varepsilon^2) \le \frac{M(X - M(X))^2}{\varepsilon^2} \qquad (7.5)$$

Т.к. неравенство $\left(X - M(X)\right)^2 > \varepsilon^2$ равносильно неравенству $\left|X - M(X)\right| > \varepsilon$, $M(X - M(X))^2 = D(X)$, то из неравенства (7.5) получаем неравенство (7.3). Учитывая, что события $\left|X - M(X)\right| > \varepsilon$ и $\left|X - M(X)\right| \le \varepsilon$ противоположные, из (7.3) получаем другое представление (7.4) неравенства Чебышева, что и требовалось доказать.

Пример 3. Оценить вероятность того, что отклонение любой случайной величины от ее математического ожидания по абсолютной величине будет не более трех средних квадратических отклонений.

Решение. Воспользуемся неравенством Чебышева (7.3), учитывая, что $\varepsilon = 3\sigma$, $D(X) = \sigma^2$:

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

Пример 4. Среднесуточное потребление электроэнергии в населенном пункте равно 20 000 квт-ч, а среднеквадратичное отклонение – 200 квт-ч. Какого потребления электроэнергии в этом населенном пункте можно ожидать в ближайшие сутки с вероятностью, не меньшей 0,96?

Решение. Воспользуемся неравенством Чебышева (7.3):

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

$$1 - \frac{40\,000}{\varepsilon^2} \ge 0{,}96 \quad \Leftrightarrow \quad \frac{40\,000}{\varepsilon^2} \le 0{,}04 \quad \Leftrightarrow \quad \varepsilon^2 \ge \frac{40\,000}{0{,}04}, \quad \varepsilon \ge 1000\,.$$

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

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

Ответ: от 19000 до 21000.

Теоремы Чебышева и Бернулли

Теорема 3 (теорема Чебышева). *Если X_1, X_2,..., X_n – попарно независимые случайные величины, дисперсии которых ограничены одной и той же постоянной, т.е. $D(X_i) \le C$, то при неограниченном увеличении числа n и для сколь угодно малого числа ε имеет место равенство:*

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - \frac{M(X_1) + M(X_2) + \ldots + M(X_n)}{n}\right| < \varepsilon\right) = 1. \qquad (7.6)$$

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

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

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

Так как рассматриваемые случайные величины независимы, то, учитывая условие теоремы, имеем: $D\left(\frac{X_1 + X_2 + ... + X_n}{n}\right) = \frac{D(X_1) + D(X_2) + ... + D(X_n)}{n^2} \leq \frac{Cn}{n^2} = \frac{C}{n}$. Используя этот результат, представим предыдущее неравенство в виде:

$$P\left(\left|\frac{X_1 + X_2 + ... + X_n}{n} - \frac{M(X_1) + M(X_2) + ... + M(X_n)}{n}\right| < \varepsilon\right) \geq 1 - \frac{C}{n\varepsilon^2}. \qquad (7.7)$$

Перейдем к пределу при $n \to \infty$:

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

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

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

Теорема доказана.

Замечание. Формулу (7.6) можно записать в виде:

$$\frac{\sum_{i=1}^{n} X_i}{n} \underset{n\to\infty}{\xrightarrow{P}} \frac{\sum_{i=1}^{n} M(X_i)}{n}. \qquad (7.8)$$

Формула (7.8) отражает тот факт, что при выполнении условий теоремы Чебышева, средняя арифметическая случайных величин *сходится по вероятности* к средней арифметической их математических ожиданий.

В отличие от записи $\frac{\sum_{i=1}^{n} X_i}{n} \underset{n\to\infty}{\to} \frac{\sum_{i=1}^{n} M(X_i)}{n}$, которая обозначает, что начиная с некоторого n для сколь угодно малого числа ε неравенство

$$\left|\frac{X_1 + X_2 + ... + X_n}{n} - \frac{M(X_1) + M(X_2) + ... + M(X_n)}{n}\right| < \varepsilon \qquad (7.9)$$

будет верно всегда, из (7.8) не следует такого же категоричного утверждения. Возможно, что в отдельных случаях неравенство (7.9) выполняться не будет, однако, с увеличением числа n вероятность неравенства (7.9) стремится к 1,

что означает практическую достоверность выполнения этого неравенства при $n \to \infty$.

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 $\left|\frac{X_1 + X_2 + ... + X_n}{n} - a\right| < \varepsilon$ будет как угодно близка к 1, если число случайных величин достаточно велико. Иначе говоря,

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + ... + X_n}{n} - a\right| < \varepsilon\right) = 1. \qquad (7.10)$$

Вывод: среднее арифметическое достаточно большого числа *случайных величин* принимает значения, близкие к сумме их математических ожиданий, то есть как угодно мало отличается от *неслучайной величины*. Например, если проводится серия измерений какой-либо физической величины, причем:
а) результат каждого измерения не зависит от результатов остальных, то есть все результаты представляют собой попарно независимые случайные величины;
б) измерения производятся без систематических ошибок (их математические ожидания равны между собой и равны истинному значению a измеряемой величины);
в) обеспечена определенная точность измерений, следовательно, дисперсии рассматриваемых случайных величин равномерно ограничены;
то при достаточно большом числе измерений их среднее арифметическое окажется сколь угодно близким к истинному значению измеряемой величины.

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

Пример 5. За значение некоторой величины принимают среднеарифметическое достаточно большого числа ее измерений. Предполагая, что среднеквадратичное отклонение возможных результатов каждого измерения не превосходит 5 мм, оценить вероятность того, что при 1000

измерений неизвестной величины отклонение принятого значения от истинного по абсолютной величине не превзойдет 0,5 мм.

Решение. Воспользуемся неравенством

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

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

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

Ответ: $P\geq 0{,}9$.

Теорема 4 (теорема Бернулли). *Частость события в n повторных независимых испытаниях, в каждом из которых оно может произойти с одной и той же вероятностью p, при неограниченном увеличении числа n сходится по вероятности к вероятности p этого события в отдельном испытании:*

$$\lim_{n\to\infty}P\left(\left|\frac{m}{n}-p\right|<\varepsilon\right)=1. \qquad (7.11)$$

или

$$\frac{m}{n}\xrightarrow[n\to\infty]{P}p \qquad (7.12)$$

Доказательство.

Введем случайные величины $X_1, X_2, \ldots, X_n$, где X_i *(индикатор события A* см. в лекции 6 биномиальный закон распределения*)*– число появлений A в i-м опыте. При этом X_i могут принимать только два значения: 1(с вероятностью p) и 0 (с вероятностью $q = 1 - p$). Кроме того, рассматриваемые случайные величины попарно независимы и их дисперсии равномерно ограничены (так как по формуле (6.4*) $D(X_i) = pq$, $p + q = 1$, откуда $pq \leq ¼$). Следовательно, к ним можно применить теорему Чебышева при $M_i = p$ *см.* (6.3*):

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

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

Замечание. *Для индикаторов события A справедлива оценка (7.7),* *которая с учетом формул (6.3*) и* (6.4*) *приводит к часто применяемой на практике оценке*

$$P\left(\left|\frac{m}{n}-p\right|<\varepsilon\right)\geq 1-\frac{pq}{n\varepsilon^2}. \qquad (7.13)$$

Пример 6. При контрольной проверке изготавливаемых приборов было установлено, что в среднем 15 шт. из 100 оказывается с теми или иными дефектами. Оценить вероятность того, что доля приборов с дефектами среди 400 изготовленных будет по абсолютной величине отличаться от математического ожидания этой доли не более чем на 0,05.

Решение. Воспользуемся неравенством (7.13).По условию $n = 400$, $e = 0{,}05$. В качестве p возьмем величину, полученную при проверке для доли брака $p = \dfrac{15}{100} = 0{,}15$.

Итак, $P\left(\left|\dfrac{m}{n} - p\right| < \varepsilon\right) \geq 1 - \dfrac{0{,}15 \cdot 0{,}85}{400 \cdot 0{,}05^2} = 0{,}8725$.

Ответ: $P \geq 0{,}8725$.

Пример 7. Вероятность того, что изделие является качественным, равна 0,9. Сколько следует проверить изделий, чтобы с вероятностью не меньшей 0,95 можно было утверждать, что абсолютная величина отклонения доли качественных изделий от 0,9 не превысит 0,01?

Решение. Воспользуемся неравенством (7.13).По условию $p = 0{,}9$, $q = 1 - 0{,}9 = 0{,}1$, $\varepsilon = 0{,}01$. Подставим в правую часть вышеприведенного неравенства эти значения:

$$1 - \frac{0{,}9 \cdot 0{,}1}{n \cdot 0{,}0001} \geq 0{,}95 \quad \Leftrightarrow \quad \frac{900}{n} \leq 0{,}05 \quad \Leftrightarrow \quad n \geq 18\,000.$$

Ответ: $n \geq 18\,000$.

Тема 8

Системы нескольких случайных величин

Наряду с одномерными случайными величинами, возможные значения которых определяются одним числом, в теории вероятностей рассматриваются и многомерные случайные величины, которые возникают когда испытание характеризуется не одной случайной величиной, а некоторой системой случайных величин: $X_1, X_2, \ldots X_n$. Геометрической иллюстрацией этого понятия служат точки *n*-мерного пространства, каждая координата которых является случайной величиной (дискретной или непрерывной), или *n*-мерные векторы. Поэтому многомерные случайные величины называют еще случайными векторами $\vec{X} = (X_1, X_2, \ldots X_n)$. Каждое возможное значение такой величины представляет собой упорядоченный набор нескольких чисел $\vec{x} = (x_1, x_2, \ldots x_n)$. Вектор $\vec{x} = (x_1, x_2, \ldots x_n)$ называется реализацией случайного вектора $\vec{X} = (X_1, X_2, \ldots X_n)$.

Двумерные случайные величины

Дискретные двумерные случайные величины

Закон распределения дискретной двумерной случайной величины (*X, Y*) имеет вид таблицы с двойным входом, задающей перечень возможных значений каждой компоненты и вероятности произведения событий $X = x_i$ и $Y = y_j$ $p_{ij}=p(x_i, y_j)=P((X = x_i)(Y = y_j))$, с которыми двумерная случайная величина (*X, Y*) принимает значение (x_i, y_j):

Y	*X*					
	x_1	x_2	…	x_i	…	x_n
y_1	$p(x_1, y_1)$	$p(x_2, y_1)$	…	$p(x_i, y_1)$	…	$p(x_n, y_1)$
…	…	…	…	…	…	…
y_j	$p(x_1, y_j)$	$p(x_2, y_j)$	…	$p(x_i, y_j)$	…	$p(x_n, y_j)$
…	…	…	…	…	…	…
y_m	$p(x_1, y_m)$	$p(x_2, y_m)$	…	$p(x_i, y_m)$	…	$p(x_n, y_m)$

Так как события $[(X = x_i)(Y = y_j)]$ ($i=1,2,\ldots,n$; $j=1,2,\ldots,m$), состоящие в том, что случайная величина *X* примет значение x_i, а случайная величина *Y* примет значение y_j, несовместны и единственно возможны, т.е. образуют полную группу, то сумма их вероятностей равна 1, т.е.

$$\sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} = 1$$

Зная закон распределения двумерной случайной величины, можно найти законы распределения ее составляющих. Действительно, событие $X=x_1$ представляет собой сумму несовместных событий ($X=x_1$, $Y=y_1$), ($X=x_1$, $Y=y_2$),…, ($X=x_1$, $Y=y_m$), поэтому $p(X=x_1)=p(x_1, y_1) + p(x_1, y_2) +\ldots+ p(x_1, y_m)$ (в правой части находится сумма вероятностей, стоящих в столбце, соответствующем $X=x_1$).

Так же можно найти вероятности остальных возможных значений X. Для определения вероятностей возможных значений Y нужно сложить вероятности, стоящие в строке таблицы, соответствующей $Y=y_j$.

Пример 1. Дан закон распределения двумерной случайной величины:

Y	X		
	-2	3	6
-0,8	0,1	0,3	0,1
-0,5	0,15	0,25	0,1

Найти законы распределения составляющих.

Решение. Складывая стоящие в таблице вероятности «по столбцам», получим ряд распределения для X:

X	-2	3	6
p	0,25	0,55	0,2

Складывая те же вероятности «по строкам», найдем ряд распределения для Y:

Y	-0,8	-0,5
p	0,5	0,5

Функция распределения двумерной случайной величины

Функцией распределения F(x, y) двумерной случайной величины (X, Y) называется вероятность совместного выполнения неравенств $X<x$ и $Y<y$:

$$F(x, y)=P(X<x, Y<y). \qquad (8.1)$$

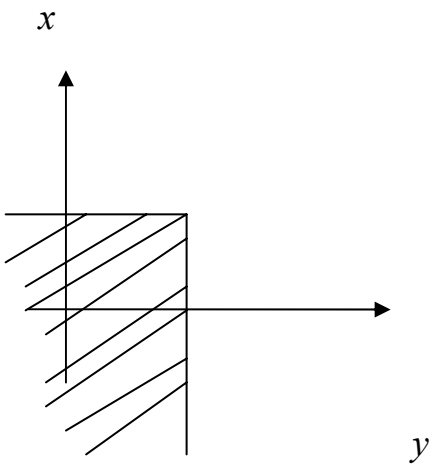


Рис.1.

Геометрически функция распределения $F(x, y)$ означает вероятность попадания случайной точка (X, Y) в область, заштрихованную на рис.1, если вершина прямого угла располагается в точке (x, y). Правая и верхняя границы в квадрант не включаются – это означает, что функция распределения *непрерывна слева* по каждому из аргументов.

Замечание. Определение функции распределения справедливо как для непрерывной, так и для дискретной двумерной случайной величины. В случае дискретной двумерной случайной величины ее функция распределения определяется по формуле:

$$F(x, y) = \sum_{\substack{i \\ x_i < x}} \sum_{\substack{j \\ y_j < y}} p_{ij} \tag{8.2}$$

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:

$F(x_2, y) \ge F(x_1, y)$, если $x_2 > x_1$;

$F(x, y_2) \ge F(x, y_1)$, если $y_2 > y_1$.

Доказательство. $F(x_2, y)=P(X<x_2, Y<y)=P(X<x_1, Y<y)+p(x_1 \le X<x_2, Y<y) \ge P(X < x_1, Y < y) = F(x_1, y)$. Аналогично доказывается и второе утверждение.

3) Имеют место предельные соотношения:

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

Доказательство. События a), b) и c) невозможны (так как невозможно событие $X<-\infty$ или $Y<-\infty$), а событие d) достоверно, откуда следует справедливость приведенных равенств.

4) При $y=+\infty$ функция распределения двумерной случайной величины становится функцией распределения составляющей X:

$$F(x, +\infty) = F_1(x). \tag{8.3}$$

При $x=+\infty$ функция распределения двумерной случайной величины становится функцией распределения составляющей Y:

$$F(+\infty, y)=F_2(y). \tag{8.4}$$

Доказательство. Так как событие $Y<+\infty$ достоверно, то $F(x, +\infty)=P(X<x)=F_1(x)$. Аналогично доказывается второе утверждение.

$$P[(x_1 < X < x_2)(y_1 < Y < y_2)] = F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1) \tag{8.5}$$

Двумерная случайная величина (X, Y) называется ***непрерывной***, если ее функция распределения $F(x, y)$ – непрерывная функция, дифференцируемая по

каждому из аргументов, и существует вторая смешанная производная: $\frac{\partial^2 F(x,y)}{\partial x \partial y}$

Плотность вероятности двумерной случайной величины

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

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y}. \qquad (8.6)$$

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) $$F(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f(x,y)dxdy \qquad (8.7)$$

(следует из определения двумерной плотности вероятности).

3) $$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)dxdy = 1 \qquad (8.8)$$

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

4) Вероятность попадания случайной точки в произвольную область.

Пусть в плоскости Oxy задана произвольная область D. Найдем вероятность того, что точка, координаты которой представляют собой систему двух случайных величин (двумерную случайную величину) с плотностью распределения $f(x,\ y)$, попадет в область D. Разобьем эту область прямыми, параллельными осям координат, на прямоугольники со сторонами Δx и Δy. Вероятность попадания в каждый такой прямоугольник равна $f(\xi_i, \eta_i)\Delta x \Delta y$, где (ξ_i, η_i) – координаты точки, принадлежащей прямоугольнику. Тогда

вероятность попадания точки в область D есть предел интегральной суммы $\sum_{i=1}^{n} f(\xi_i,\eta_i)\Delta x\Delta y$, то есть

$$P((X,Y)\subset D)=\iint_D f(x,y)dxdy. \qquad (8.9)$$

Функции распределения и плотности вероятностей одномерных составляющих двумерной случайной величины

Для нахождения функции распределения каждой составляющей, зная двумерную функцию плотности вероятности, скомбинируем (8.3) и (8.7) ((8.4) и (8.7)):

$$F_1(x)=F(x,+\infty)=\int_{-\infty}^{+\infty}\int_{-\infty}^{x} f(x,y)dxdy \qquad (8.10)$$

$$F_2(y)=F(+\infty,y)=\int_{-\infty}^{y}\int_{-\infty}^{+\infty} f(x,y)dxdy \qquad (8.11)$$

Тогда по определению одномерной плотности распределения

$$f_1(x)=\frac{dF_1(x)}{dx}=\frac{dF(x,\infty)}{dx}=\frac{d\left(\int_{-\infty}^{x}\int_{-\infty}^{\infty} f(x,y)\right)}{dx}=\int_{-\infty}^{\infty} f(x,y)dy. \qquad (8.12)$$

Аналогично найдем

$$f_2(y)=\int_{-\infty}^{\infty} f(x,y)dx. \qquad (8.13)$$

Условные законы распределения

Рассмотрим дискретную двумерную случайную величину и найдем закон распределения составляющей X при условии, что Y примет определенное значение (например, $Y=y_1$). Для этого воспользуемся формулой Байеса, считая гипотезами события $X=x_1$, $X=x_2,\ldots$, $X=x_n$, а событием A – событие $Y=y_1$. При такой постановке задачи нам требуется найти условные вероятности гипотез при условии, что A произошло. Следовательно,

$$p(x_i\mid y_1)=\frac{P((X=x_i)(Y=y_1))}{P(Y=y_1)}.$$

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

$$p(x_i \mid y_j) = \frac{P((X = x_i)(Y = y_j))}{P(Y = y_j)}. \qquad (8.14)$$

Аналогично находят условные законы распределения составляющей Y:

$$p(y_j \mid x_i) = \frac{P((X = x_i)(Y = y_j))}{P(X = x_i)}. \qquad (8.15)$$

Пример 2. Найдем закон распределения X при условии Y=-0,8 и закон распределения Y при условии X=3 для случайной величины, рассмотренной в примере 1.

Решение.

$$p(x_1 \mid y_1) = \frac{0,1}{0,5} = \frac{1}{5} = 0,2; \quad p(x_2 \mid y_1) = \frac{0,3}{0,5} = \frac{3}{5} = 0,6; \quad p(x_3 \mid y_1) = \frac{0,1}{0,5} = \frac{1}{5} = 0,2.$$

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

Условной плотностью вероятности $f(x \mid y) = f_y(x)$ распределения составляющих X при данном значении $Y=y$ называется

$$f(x \mid y) = f_y(x) = \frac{f(x,y)}{f_2(y)} = \frac{f(x,y)}{\int\limits_{-\infty}^{\infty} f(x,y)dx}. \qquad (8.16)$$

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

$$f(y \mid x) = f_x(y) = \frac{f(x,y)}{f_1(x)} = \frac{f(x,y)}{\int\limits_{-\infty}^{\infty} f(x,y)dy}. \qquad (8.17)$$

Числовые характеристики двумерных случайных величин

Такие характеристики, как начальные и центральные моменты, можно ввести и для системы двух случайных величин.

Начальным моментом порядка k, s двумерной случайной величины (X, Y) называется математическое ожидание произведения X^k на Y^s:

$$\alpha_{k,s} = M(X^k Y^s). \qquad (8.18)$$

Для дискретных случайных величин

$$\alpha_{k,s} = \sum_i \sum_j x_i^k y_j^s p_{ij,} \qquad (8.19)$$

Для непрерывных случайных величин

$$\alpha_{k,s} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x^k y^s f(x,y) dxdy. \qquad (8.20)$$

Центральным моментом порядка k, s двумерной случайной величины (X, Y) называется математическое ожидание произведения $(X - M(X))^k$ на $(Y - M(Y))^s$:

$$\mu_{k,s} = M(((X - M(X))^k (Y - M(Y))^s). \qquad (8.21)$$

Для дискретных случайных величин

$$\mu_{k,s} = \sum_i \sum_j (x_i - M(X))^k (y_j - M(Y))^s p_{ij} \qquad (8.22)$$

Для непрерывных случайных величин

$$\mu_{k,s} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - M(X))^k (y - M(Y))^s f(x,y) dxdy \qquad (8.23)$$

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

Наряду с числовыми характеристиками $M(X)$, $M(Y)$, $D(X)$, $D(Y)$ одномерных составляющих рассматриваются также числовые характеристики условных распределений: условные математические ожидания $M_y(X)$, $M_x(Y)$, $D_y(X)$, $D_x(Y)$.

Например,

$$M_y(X) = \int_{-\infty}^{+\infty} x f_y(x) dx$$

$$D_y(X) = \int_{-\infty}^{+\infty} (x - M(X))^2 f_y(x) dx$$

Корреляционный момент и коэффициент корреляции

Корреляционным моментом (или ковариацией) системы двух случайных величин называется центральный момент порядка 1,1:

$$K_{xy} = \mu_{1,1} = M(((X - M(X))(Y - M(Y))). \qquad (8.24)$$

Для дискретных случайных величин

$$K_{xy} = \sum_i \sum_j (x_i - M(X))(y_j - M(Y))\, p_{ij} \qquad (8.25)$$

Для непрерывных случайных величин

$$K_{xy} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - M(X))(y - M(Y))\, f(x,y)\, dxdy \qquad (8.26)$$

Ковариация может быть вычислена по формуле:

$$K_{xy} = M(XY) - M(X) \cdot M(Y) \qquad (8.27)$$

Действительно, $K_{xy} = M(XY - Y \cdot M(X) - X \cdot M(Y) + M(X)M(Y))$ по свойству математического ожидания:

$$K_{xy} = M(XY) - M(Y \cdot M(X)) - M(X \cdot M(Y)) + M(M(X)M(Y)) =$$
$$= M(XY) - M(Y) \cdot M(X) - M(X) \cdot M(Y) + M(X)M(Y) = M(XY) - M(X) \cdot M(Y)$$

Корреляционный момент описывает связь между составляющими двумерной случайной величины.

Случайные величины X и Y называются ***некоррелированными***, если $K_{xy}=0$.

Убедимся, что для независимых X и Y $K_{xy}=0$. Действительно, в этом случае $f(x,y)=f_1(x)f_2(y)$, тогда

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

Итак, две независимые случайные величины являются и некоррелированными. Однако понятия коррелированности и зависимости не эквивалентны, а именно, величины могут быть зависимыми, но при этом некоррелированными. Дело в том, что корреляционный момент характеризует не всякую зависимость, а только *линейную*. В частности, если $Y=aX+b$, то $K_{xy} = \pm\sigma_x\sigma_y$.

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

$$r_{xy} = \frac{K_{xy}}{\sigma_x \sigma_y} \ . \qquad (8.28)$$

Теорема 1. *Возможные значения коэффициента корреляции удовлетворяют неравенству* $|r_{xy}|\le 1$.

Доказательство. Докажем сначала, что $|K_{xy}|\le\sigma_x\sigma_y$. Действительно, если рассмотреть случайную величину $Z_1=\sigma_y X-\sigma_x Y$ и найти ее дисперсию, то получим: $D(Z_1)=2\sigma_x^2\sigma_y^2-2\sigma_x\sigma_y K_{xy}$. Так как дисперсия всегда неотрицательна, то $2\sigma_x^2\sigma_y^2-2\sigma_x\sigma_y K_{xy}\ge 0$, откуда $|K_{xy}|\le\sigma_x\sigma_y$. Отсюда $\left|\dfrac{K_{xy}}{\sigma_x\sigma_y}\right|=|r_{xy}|\le 1$, что и требовалось доказать.

Равномерное распределение на плоскости

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

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

Действительно, из формулы (8.8):

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)dxdy=\int_a^b\int_c^d Cdxdy=C(b-a)(d-c)=1\Rightarrow\ C=\frac{1}{(b-a)(d-c)}$$

Найдем двумерную функцию распределения:

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

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

КОНТРОЛЬНЫЕ ВОПРОСЫ

по курсу «Теория вероятностей»

1. Предметы и методы теории вероятностей и математической статистики.
2. Элементы комбинаторики, размещения, перестановки, сочетания.
3. Случайные события. Операции над событиями.
4. Классическая формула вероятности. Статистическая вероятность. Геометрические вероятности.
5. Теорема сложения вероятностей.
6. Условная вероятность. Теорема умножения вероятностей.
7. Формула полной вероятности. Формула Байеса.
8. Формула Бернулли.
9. Формула Пуассона.
10. Локальная и интегральная теоремы Лапласа.
11. Дискретные случайные величины.
12. Непрерывные случайные величины.
13. Функция распределения вероятностей.
14. Плотность распределения вероятностей.
15. Математическое ожидание случайной величины и его свойства.
16. Дисперсия случайной величины и ее свойства.
17. Моменты случайных величин.
18. Биномиальное распределение и его характеристики.
19. Распределение Пуассона.
20. Геометрическое и гипергеометрическое распределения.
21. Равномерное распределение в интервале.
22. Показательное распределение.
23. Нормальный закон распределения.
24. Неравенство Маркова. Неравенство Чебышева.
25. Теорема Чебышева. Теорема Бернулли.
26. Системы нескольких случайных величин. Дискретные двумерные случайные величины.
27. Системы нескольких случайных величин. Функция распределения и плотность вероятностей двумерной случайной величины.
28. Числовые характеристики двумерных случайных величин.
29. Корреляционный момент и коэффициент корреляции двумерных случайных величин.

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Тема 1

Задачи математической статистики. Генеральная и выборочная совокупности. Вариационный ряд и его основные числовые характеристики

Теория вероятностей (ТВ) — математическая наука, изучающая закономерности случайных явлений. Под случайными явлениями понимаются явления с неопределенным исходом, происходящие при неоднократном воспроизведении некоторого комплекса условий.

Математическая статистика (МС) — раздел математики, изучающий методы сбора, систематизации и обработки результатов наблюдений с целью выявления статистических закономерностей. МС опирается на ТВ. Если ТВ изучает закономерности случайных явлений на основе абстрактного определения действительности (теоретической вероятностной модели), то МС оперирует непосредственно результатами наблюдений над случайным явлением, представляющим выборку из некоторой конечной или гипотетической бесконечной генеральной совокупности. Используя результаты, полученные теорией вероятностей, МС позволяет не только оценить значения искомых характеристик, но и выявить степень точности выводов, получаемых при обработке данных. Коротко говоря, ТВ позволяет находить вероятности «сложных» событий через вероятности «простых» событий (связанных с ними каким-то образом), а МС по наблюдаемым значениям (выборке) оценивает вероятности этих событий либо осуществляет проверку предположений (гипотез) относительно этих вероятностей.

Задачи МС:

1. указать способы сбора и группировки статистических сведений, полученных в результате наблюдений или в результате специально поставленных экспериментов;
2. разработать методы анализа статистических данных в зависимости от целей исследования.

Ко второй задаче относятся:

а) оценка неизвестной вероятности события; оценка неизвестной функции распределения; оценка параметров распределения, вид которого известен; оценка зависимости случайной величины от одной или нескольких случайных величин и т.д.

б) проверка статистических гипотез о виде неизвестного распределения или о величине параметров распределения, вид которого известен.

Современная МС разрабатывает способы определения числа необходимых экспериментов до начала исследований (*планирование эксперимента*), в ходе исследования (*последовательный анализ*) и решает другие задачи. Современную МС определяют как ***науку о принятии решений в условиях неопределенности***.

Пусть требуется изучить совокупность однородных объектов относительно некоторого *качественного* или *количественного* признака, характеризующего эти объекты. Например, в партии изделий *качественным* признаком может служить стандартность деталей, а *количественным* – контролируемый размер деталей.

Предположим, что имеется некоторое множество $x_1, x_2, \ldots, x_N$ однородных предметов, определенный *признак* которых исследуется. Вся подлежащая изучению совокупность предметов называется ***генеральной совокупностью***. Предположим далее, что исследовать данный признак у всех предметов этой совокупности не представляется возможным (либо их очень много, либо они физически уничтожаются, либо по другим причинам). В этом случае используют *выборочный метод*, согласно которому из данной генеральной совокупности *случайным образом* выбираются n элементов $x_1, x_2, \ldots, x_n$. Та часть объектов, которая отобрана для непосредственного изучения из генеральной совокупности, называется ***выборочной совокупностью*** или ***выборкой***.

Размахом выборки ω называют разность между максимальным $x_{\max}$ и минимальным $x_{\min}$ значениями элементов выборки: $\omega = x_{\max} - x_{\min}$.

Объёмом совокупности (выборочной или генеральной) называют *число* (n или N соответственно) *объектов* этой совокупности. Если, например, из 1000 деталей для обследования отобрано 85, то объём генеральной совокупности N=1000, объём выборки n=85.

Для того, чтобы результаты обследования выборки отражали основные черты изучаемого признака, необходимо, чтобы объём выборки не был чрезвычайно мал. Выборка называется ***репрезентативной*** (представительной), если она достаточно хорошо представляет количественные соотношения генеральной совокупности. Например, о распределении жителей г. Минска по росту нельзя судить по результатам обследования одной квартиры. Ясно, что данные, относящиеся к одному высотному дому или группе домов, более показательны, репрезентативны.

Предположим, что проводится случайный эксперимент, например, измеряется некоторая величина ξ. На измерения могут влиять как систематические ошибки (погрешность прибора), так и случайные ошибки (внешние условия и т.п.), получаемые в результате воздействия различных (случайных) факторов. Таким образом, ξ можно интерпретировать как случайную величину (СВ). Предположим далее, что возможные значения

$$x_1, x_2, \cdots, x_n \qquad (1.1)$$

СВ ξ известны. Эти значения можно считать генеральной совокупностью. Если же известны и вероятности появления значений $x_1, x_2, \cdots, x_n$, то нам известен и закон распределения СВ ξ (*теоретический закон распределения* или *распределение генеральной совокупности*). Обобщая ситуацию на случай произвольной СВ (как дискретной, так и непрерывной), можно говорить **о *теоретической функции распределения*** (*функции распределения генеральной совокупности*). На практике она неизвестна. В этом случае производят

измерения (случайные) ξ, в результате получают значение $\tilde{x}_1$ СВ. Однако, как правило, судить о значениях СВ ξ по одному измеренному значению $\tilde{x}_1$ неубедительно. Поэтому на практике производят *n независимых совокупностей* случайных испытаний (измерений) данной СВ. В результате получают k реализовавшихся значений

$$\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_k \qquad (1.2)$$

данной СВ, которые называют *выборочными значениями* (данной СВ). Совокупность (1.2) можно интерпретировать как *выборку объёма* k. Выборочные значения $\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_k$ называют *вариантами*. Совокупность (1.2), расположенная в порядке неубывания, называется ***вариационным рядом***. Выберем из (1.2) только различные, расположенные в возрастающем порядке значения $x_1^*, x_2^*, , \cdots, x_r^*$.

Пусть признак x_1^* наблюдался n_1 раз, $x_2^* - n_2$ раза,…, $x_r^* - n_r$ раз. Тогда $\sum_{i=1}^{r} n_i = n$, где n – объём выборки, наблюдаемые величины $x_1^*, x_2^*, \cdots \cdots, x_r^*$ – варианты. Числа наблюдений n_i, $i = 1, 2, \cdots, p$, называют ***частотами***, а отношение n_i / n частот к объёму выборки – ***относительными частотами*** ω_i:

$$\omega_i = \frac{n_i}{n}, \quad i = \overline{1, p}; \quad \sum_{i=1}^{r} \omega_i = \sum_{i=1}^{r} \frac{n_i}{n} = \frac{1}{n} \underbrace{\sum_{i=1}^{r} n_i}_{=n} = 1. \qquad (1.3)$$

В теории вероятностей под распределением понимают соответствие между возможными значениями случайных величин и их вероятностями

ξ_i	x_1	x_2	…	x_r
p_i	p_1	p_2	⋯	p_r

$\sum_{i=1}^{r} p_i = 1$

В математической статистике под распределением понимают соответствие между наблюдаемыми вариантами и их частотами или относительными частотами.

Статистическим рядом или статистическим распределением выборки называют совокупность пар (x_i, n_i), $i = \overline{1,k}$, где x_1, $x_2, \ldots, x_k$ – различные элементы выборки, а $n_1, n_2, ..., n_k$ – частота выборочных значений x_1, $x_2, \ldots, x_k$, $\sum_{i=1}^{k} n_i = n$.

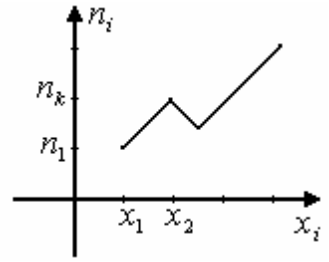
Статистический ряд записывается в виде таблицы

x_i	x_1	x_2	…	x_k
n_i	n_1	n_2	…	n_k

$\sum_{i=1}^{k} n_i = n$

Для наглядности принято использовать ***полигон частот*** как форму *графического представления статистических распределений.* ***Полигоном частот*** (относительных частот) выборки называется ломаная с вершинами в точках (x_i, n_i), $i = \overline{1,k}$, $(x_i, n_i / n)$, $i = \overline{1,k}$ (по оси координат откладываются

выборочные значения x_i, по оси ординат – соответствующие частоты n_i или относительные частоты ω_i):



Пример 1. Выборка, полученная в результате статистического наблюдения (единицы измерения опускаем) – 7, 17, 14, 17, 10, 7, 7, 14, 7, 14;

– ранжированный вариационный ряд –

$$x_j: \underbrace{7,7,7,7}_{n_1=4}, \underbrace{10,}_{n_2=1} \underbrace{14,14,14}_{n_3=3}, \underbrace{17,17}_{n_4=2}, \text{ где } j=1,2,\ldots,n \ , n=10;$$

– соответствующее статистическое распределение ($i=1,2,\ldots,k \ , k=4$):

x_i	7	10	14	17
n_i	4	1	3	2.

При большом объеме выборки ее элементы объединяют в группы (разряды, интервалы), представляя результаты опытов в виде *интервального статистического ряда*. Для этого весь диапазон значений случайной величины ξ (от $x_{\min}$ до $x_{\max}$) разбивают на k интервалов одинаковой длины h (обычно k меняется от 5 до 20) и подсчитывают частоты n_i (или относительные частоты ω_i) значений выборки, попавших в интервалы. Величина n_i/h называется *плотностью частоты*, а ω_i/h *– плотностью относительной частоты*.

Пусть x_i^* – середина i-го интервала, n_i – число элементов выборки, попавших в i-й интервал (при этом элемент, совпавший с верхней границей интервала, относится к последующему интервалу). Таким образом, *получим группированный статистический ряд*, в верхней строке которого записаны середины соответствующих интервалов x_i^*, а в нижней — частоты:

x_i^*	x_1^*	x_2^*	…	x_k^*
n_i	n_1	n_2	…	n_k

$\sum_{i=1}^{k} n_i = n$

Пример 2. Выборка, полученная в результате статистического наблюдения (единицы измерения опускаем) –3,14; 1,41; 2,87; 3,62; 2,71; 3,95;

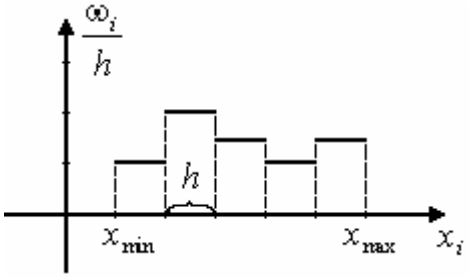
– ранжированный вариационный ряд –
x_j: 1,41; 2,71; 2,87; 3,14; 3,62; 3,95; где $j=1,2,\ldots,n \ , n=6$;

– соответствующее интервальное статистическое распределение ($i=1,2,\ldots,k \ , k=3$):

x_i	1–2	2–3	3–4
n_i	1	2	3.

Для *графического представления интервальных статистических распределений* принято использовать ***гистограмму относительных частот.***

Гистограммой относительных частот интервального статистического ряда называется ступенчатая фигура, составленная из прямоугольников, построенных на интервалах группировки длины h и высоты ω_i/h так, что площадь каждого прямоугольника равна относительной частоте ω_i.



Для построения гистограммы относительных частот на оси абсцисс откладывают частичные интервалы, а над ними проводят отрезки длиной ω_i/h параллельно оси ординат. Очевидно, площадь i-го частичного прямоугольника равна ω_i – относительной частоте вариант, попавших в i-ый интервал. Следовательно, ***площадь гистограммы относительных частот равна сумме всех относительных частот (т.е. равна 1), а площадь гистограммы частот равна объему выборки*** n**.**

Пример 3. Имеется распределение 80 предприятий по числу работающих на них (чел.):

x_i	150	250	350	450	550	650	750
n_i	1	3	7	30	19	15	5 .

Решение. Признак X – число работающих (чел.) на предприятии. В данной задаче признак X является дискретным. Поскольку различных значений признака сравнительно немного – $k = 7$, применять интервальный ряд для представления статистического распределения нецелесообразно (в прикладной статистике в подобных задачах часто используют именно интервальный ряд). Ряд распределения – дискретный. Построим полигон распределения частот (рис. 1).

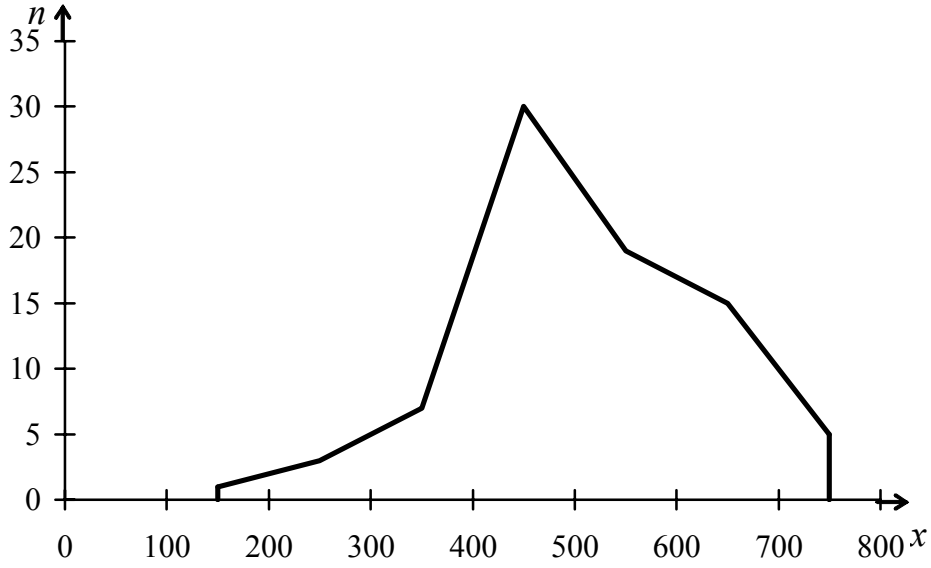


Рис. 1

Пример 4. Дано распределение 100 рабочих по затратам времени на обработку одной детали (мин):

x_{i-1}–x_i	22–24	24–26	26–28	28–30	30–32	32–34
n_i	2	12	34	40	10	2 .

Решение. Признак X – затраты времени на обработку одной детали (мин). Признак X – непрерывный, ряд распределения – интервальный. Построим гистограмму частот (рис. 2), предварительно определив $h = (x_k - x_0)/k = (34-22)/6 = 2$ $(k = 6)$ и плотность частоты n_i/h:

x_{i-1}–x_i	22–24	24–26	26–28	28–30	30–32	32–34
n_i/h	1	6	17	20	5	1 .

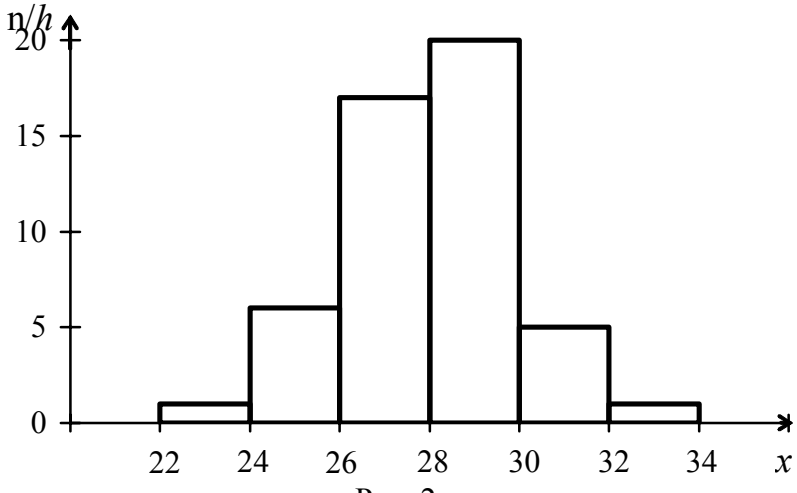


Рис. 2

Эмпирическая функция распределения

Пусть известно статистическое распределение (или статистический ряд) количественного признака ξ; n_x – число наблюдений, при которых наблюдалось значение признака, меньшее x, т.е. $\xi < x$; n – общее число наблюдений (объём выборки). Тогда относительная частота события $\xi < x$ есть n_x/n. При изменении x меняется и n_x/n, т.е. относительная частота n_x/n является функцией x. Так как эта функция находится эмпирическим (т.е. опытным) путём, то её называют ***эмпирической.***

Эмпирической функцией распределения (функцией распределения выборки) называется функция

$$F^*(x) = \frac{n_x}{n}, \qquad (1.4)$$

определяющая для каждого значения x относительную частоту события $\xi < x$. В (1.4) n_x – число вариант, меньших x ($n_x = \sum_{x_i < x} n_i$, x_i – варианты, n – объём выборки). Поэтому для расчетов удобна формула вида:

$$F^*(x) = \sum_{x_i < x} \frac{n_{x_i}}{n} \qquad (1.5)$$

Тогда, например, $F^*(x_3)$ означает $F^*(x_3) = n_{x_3}/n$, где n_{x_3} – число вариант, меньших x_3 или в табличной форме:

x_{i-1}–x_i	$-\infty$–x_0	x_0–x_1	x_1–x_2	…	x_{i-1}–x_i	…	x_{k-1}–x_k
$F^*(x_i)$	0	ω_1	$\omega_1+\omega_2$	…	$F^*(x_{i-1})+\omega_i$	…	$F^*(x_{k-1})+\omega_k = 1$;

Функцию распределения $F(x)$ *генеральной совокупности* ***называют теоретической функцией распределения***. Различие между эмпирической $F^*(x)$ и теоретической $F(x)$ функциями распределения состоит в том, что $F(x)$ определяет вероятность события $\xi < x$, а $F^*(x)$ – относительную частоту того же события. $F^*(x)$ обладает всеми свойствами $F(x)$.

Свойства эмпирической функции распределения $F^*(x)$*:*

1. значения $F^*(x)$ принадлежат $[0,1]$; $F^*(x) \in [0;1]$;
2. $F^*(x)$ – неубывающая функция;
3. если x_1 – наименьшая варианта, а x_k – наибольшая варианта, то $F^*(x) = 0$ для $x \le x_1$, $F^*(x) = 1$ для $x > x_k$;
4. $F^*(x)$ непрерывная слева функция.

Эмпирическая функция распределения выборки $F^*(x)$ ***служит для оценки теоретической функции распределения*** $F(x)$ ***генеральной совокупности***.

Пример 5. Построить эмпирическую функцию распределения по данному распределению выборки

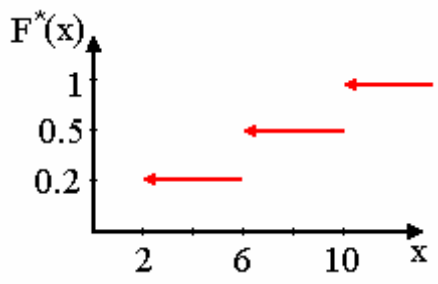
Варианты x_i	2	6	10
Частоты n_i	12	18	30

$n = \sum_{i=1}^{3} n_i = 60.$

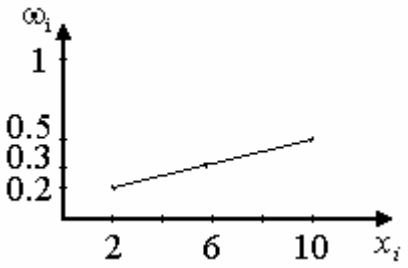
Решение. Здесь $x_1 = 2$ – наименьшая варианта, следовательно, $F^*(x) = 0$ для $x \le 2$; $x_3 = 10$ – наибольшая варианта, тогда $F^*(x) = 1$ при $x > 10$. Для $2 < x \le 6$ имеем $F^*(x) = n_x / n = \sum_{x_i < x} n_i / n = 12/60 = 0{,}2$, а для $6 < x \le 10$ следует $F^*(x) = (12+18)/60 = 0{,}5$.

Приведём аналитический вид полученной эмпирической функции распределения $F^*(x)$, её график и полигон частот:

$$F^*(x) = \begin{cases} 0, & x \le 2 \\ 0{,}2, & 2 < x \le 6 \\ 0{,}5, & 6 < x \le 10 \\ 1, & x > 10 \end{cases},$$



Полигон относительных частот имеет вид



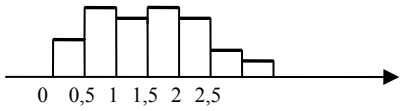
где координаты ω_i его вершин (x_i, ω_i) определяются по формулам

$$\omega_1 = \frac{12}{60} = 0{,}2, \quad \omega_2 = \frac{18}{60} = 0{,}3, \quad \omega_3 = \frac{30}{60} = 0{,}5, \quad \Sigma\omega_i = 1.$$

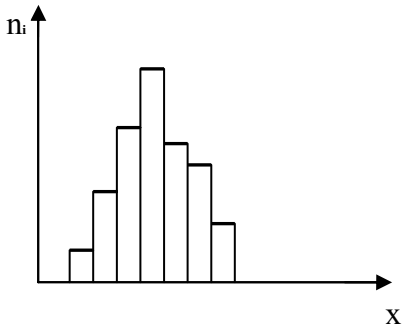
Тема 2

Точечное оценивание параметров распределения. Оценки математического ожидания и дисперсии для нормального закона распределения

Анализ полигона, гистограммы, эмпирической функции распределения даёт возможность сделать допущение о законе распределения случайных величин. По виду полученной гистограммы можно строить гипотезы об истинном характере распределения СВ ξ. Например, получив гистограмму вида:



можно заключить, что СВ ξ на отрезке [0,5;2,5] распределена равномерно. Из гистограммы вида



естественно предположить, что распределение СВ ξ является нормальным. На практике, однако, редко встречается такое положение, когда изучаемый закон распределения СВ ξ неизвестен полностью. Чаще всего из каких-либо теоретических соображений вид закона распределения ясен заранее и требуется найти только некоторые *параметры*, от которых он зависит. Например, если известно, что закон распределения СВ ξ нормальный (с плотностью распределения $f_\xi(x)=\frac{2}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}$), то задача сводится к нахождению значений двух параметров: a и σ. В некоторых задачах и сам вид закона распределения несущественен, а требуется найти только его числовые характеристики. Во всех подобных случаях можно обойтись сравнительно небольшим числом наблюдений – порядка одного или нескольких десятков. При *изучении числовых характеристик* СВ ξ мы рассматриваем математическое ожидание M_ξ, дисперсию D_ξ, среднее квадратичное отклонение σ_ξ. Эти числовые (т.е. точечные) характеристики играют большую роль в теории вероятностей. Аналогичные числовые характеристики существуют и для статистических распределений. Каждой числовой характеристике СВ ξ соответствует её статистическая аналогия.

Пусть закон распределения СВ ξ содержит некоторый параметр θ. Численное значение θ не указано, хотя оно и является вполне определенным числом. В связи с этим возникает следующая задача.

Исходя из набора наблюдаемых значений (выборки) $x_1, x_2, \ldots, x_n$ случайной величины ξ, полученного в результате n независимых испытаний, оценить значение параметра θ.

Оценкой (или ***статистикой***) θ_n^* неизвестного параметра θ теоретического распределения называют функцию $f(x_1, x_2, \ldots, x_n)$ от наблюдаемых (выборочных) значений случайных величин $x_1, x_2, \ldots, x_n$, обладающую свойством статистической устойчивости. Так как $x_1, x_2, \ldots, x_n$ рассматриваются как независимые случайные величины, то и оценка θ_n^* является случайной величиной, зависящей от закона распределения СВХ и числа n. Сам же оцениваемый параметр есть величина неслучайная (детерминированная).

Оценки параметров разделяются на *точечные* и *интервальные*.

Точечной называют статистическую оценку, определяемую одним числом $\theta_n^* = f(x_1, x_2, \ldots, x_n)$ (далее будем обозначать просто θ^*), где $x_1, x_2, \ldots, x_n$ – результаты n наблюдений (выборки) над количественным признаком ξ.

К оценке θ^* естественно предъявить ряд *требований:*

1. Желательно, чтобы, пользуясь величиной θ^* вместо θ, не делалось систематических ошибок ни в сторону занижения, ни в сторону завышения, т.е. чтобы выполнялось равенство

$$M(\theta^*) = \theta. \qquad (2.1)$$

Оценка, удовлетворяющая условию (2.1), называется ***несмещённой***.

Несмещённой называют точечную оценку, математическое ожидание которой равно оцениваемому параметру при любом объёме выборки. Требование несмещённости оценки особенно важно при малом числе испытаний.

Смещённой называют оценку, математическое ожидание которой не равно оцениваемому параметру θ.

2. Желательно, чтобы с увеличением числа n опытов значения случайной величины θ^* концентрировались около θ всё более тесно, т.е.

$$\theta^* \xrightarrow{P} \theta \quad при \quad n \to \infty \text{ или } \lim_{n\to\infty} P(|\theta^* - \theta| < \varepsilon) = 1 \qquad (2.2)$$

Оценку, обладающую свойством (2.2), называют ***состоятельной***.

Если оценка θ^* параметра θ является несмещенной, а ее дисперсия

$$D(\theta^*) \to 0 \quad при \quad n \to \infty, \qquad (2.3)$$

то оценка θ^* является и состоятельной. Это непосредственно вытекает из неравенства Чебышева $P(|\theta^* - \theta| < \varepsilon) \geq 1 - \dfrac{D(\theta^*)}{\varepsilon^2}$

3. Если θ_1^* и θ_2^* – различные несмещённые оценки параметра θ, то оценка θ_1^* называется более эффективной, чем оценка θ_2^*, если

$$D_{\theta_1^*} < D_{\theta_2^*}. \qquad (2.4)$$

Поэтому разумно самой эффективной оценкой назвать оценку, на которой достигается $\min D$.

Эффективной называют статистическую оценку, которая при заданном объёме выборки n имеет наименьшую возможную дисперсию.

Пусть изучается дискретная генеральная совокупность относительно количественного признака ξ.

Генеральной средней $\overline{x}$ называют среднее арифметическое значений признака генеральной совокупности.

Если все значения $x_1, x_2, \ldots, x_N$ признака генеральной совокупности объёма N различны, то

$$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N}. \qquad (2.5)$$

Если же значения признака $x_1, x_2, \ldots, x_k$ имеют соответственно частоты $N_1, N_2, \ldots, N_k$, причём $\sum_{i=1}^{k} N_i = N,$ то

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i N_i}{N}. \qquad (2.6)$$

т.е. ***генеральная средняя есть средняя взвешенная значений признака с весами, равными соответствующим частотам.***

Пусть для изучения *генеральной совокупности* относительного количественного признака ξ извлечена *выборка* объема n. Наиболее распространёнными оценками в математической статистике являются *выборочное среднее* $\overline{x}_в$ – оценка математического ожидания M_ξ, *выборочная дисперсия* $D_в$ – оценка дисперсии D_ξ, *выборочное среднеквадратичное отклонение* S – оценка среднеквадратичного отклонения σ.

Если все значения $x_1, x_2, \ldots, x_n$ признака ξ выборки объема n различны, то

$$\overline{x}_в = \frac{\sum_{i=1}^{n} x_i}{n}, \qquad (2.7)$$

т.е. ***выборочная средняя*** $\overline{x}_в$ есть среднее арифметическое значение признака выборочной совокупности. Если же значения признака $x_1, x_2, \ldots, x_k$ имеют соответственно частоты $n_1, n_2, \ldots, n_k,$ причем $\sum_{i=1}^{k} n_i = n,,$ то выборочным средним является величина

$$\bar{x}_в = \frac{\sum_{i=1}^{k} n_i x_i}{n}, \qquad (2.8)$$

т.е. *выборочная средняя есть средняя взвешенная значений признака с весами, равными соответствующим частотам.* Покажем, что

В качестве точечной оценки M^* для M_ξ – математического ожидания СВ ξ – может служить выборочное среднее $\bar{x}_в$, т. е. $M^* = \bar{x}_в$

Действительно, т.к. СВ $x_1, x_2, \ldots, x_n$ имеют один и тот же закон распределения, совпадающий с законом распределения СВ ξ, то

$$M(M^*) \overset{(2.7)}{=} M\frac{\sum_{i=1}^{n} x_i}{n} = \frac{1}{n} M\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n} M(x_i) = \frac{1}{n}\cdot nM_\xi = M_\xi,$$

т.е. оценка M^* для математического ожидания СВ ξ согласно (2.1) является несмещенной.

Рассмотрим дисперсию DM^*:

$$DM^* = D\bar{x}_в = D\frac{\sum_{i=1}^{n} x_i}{n} = \frac{1}{n^2} D\sum_{i=1}^{n} x_i = \frac{1}{n^2}\sum_{i=1}^{n} D(x_i) = \frac{1}{n^2}\cdot nD_\xi = \frac{D_\xi}{n},$$

где D_ξ – дисперсия СВ ξ. Так как $DM^* \to 0$ при $n \to \infty$, то из последнего равенства в силу (2.3) следует, что оценка M^* является состоятельной и несмещённой.

По определению: $D_\xi = M[\xi - M_\xi]^2$. Так как D_ξ есть математическое ожидание СВ $[\xi - M\xi]^2$, то естественной *оценкой* для D_ξ представляются выражения:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N} \qquad (2.9)$$

$$\sigma^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 N_i}{N}, \quad \sum_{i=1}^{k} N_i = N \qquad (2.9^*)$$

$$D_в = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_в)^2}{n}. \qquad (2.10)$$

$$D_в = \frac{\sum_{i=1}^{k} n_i(x_i - \bar{x}_в)^2}{n}, \qquad \sum_{i=1}^{k} n_i = n. \qquad (2.10^*)$$

Статистическую оценку $D_в$ целесообразно использовать для оценки σ^2 дисперсии генеральной совокупности. Ее называют *выборочной дисперсией*.

Выборочная дисперсия $D_в$ ***есть среднее арифметическое квадратов отклонений наблюдаемых значений признака*** ξ ***от их выборочного среднего***.

Преобразуем формулу (2.10):

$$D_в = \frac{1}{n}\cdot\sum_{i=1}^{n}(x_i - \overline{x}_в)^2 = \frac{1}{n}\cdot\sum_{i=1}^{n}[(x_i - M_\xi) - (\overline{x}_в - M_\xi)]^2 =$$

$$= \frac{1}{n}\cdot\sum_{i=1}^{n}(x_i - M_\xi)^2 - 2\cdot\frac{1}{n}\cdot(\overline{x}_в - M_\xi)\cdot\sum_{i=1}^{n}(x_i - M_\xi) + \frac{1}{n}\cdot n\cdot(\overline{x}_в - M_\xi)^2 =$$

$$= \frac{1}{n}\cdot\sum_{i=1}^{n}\left(x_i - M_\xi\right)^2 - 2\cdot(\overline{x}_в - M_\xi)^2 + (\overline{x}_в - M_\xi)^2 = \frac{1}{n}\cdot\sum_{i=1}^{n}\left(x_i - M_\xi\right)^2 - (\overline{x}_в - M_\xi)^2.$$

Найдем математическое ожидание оценки $D_в$**:**

$$MD_в = \frac{1}{n}\cdot\sum_{i=1}^{n}M\left(x_i - M_\xi\right)^2 - M(\overline{x}_в - M_\xi)^2 = \frac{1}{n}\cdot n\cdot D_\xi - D_в =$$

$$= D_\xi - \frac{D_\xi}{n} = D_\xi - \frac{D_\xi}{n} = \frac{n-1}{n}D_\xi. \qquad (2.11)$$

Полученная оценка (2.11) является смещенной, т.к. $MD_в \neq D_\xi$, а именно: $MD_в = \frac{n-1}{n}D_\xi$.

Несмещенную оценку для D_ξ можно получить, если положить

$$D_\xi \approx \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_в)^2 = S^2. \qquad (2.12)$$

Действительно, тогда из (2.1) следует

$$MS^2 = \frac{1}{n-1}M\sum_{i=1}^{n}(x_i - \overline{x}_в)^2 = \frac{n}{n-1}\cdot M\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_в)^2 = \frac{n}{n-1}MD_в = \frac{n}{n-1}\frac{n-1}{n}D_\xi = D_\xi.$$

Данная оценка является *несмещенной оценкой* дисперсии. Её называют *исправленной дисперсией* и обозначают

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_в)^2. \qquad (2.13)$$

Если значения $x_1, x_2, \ldots, x_k$ встречаются с частотами $n_1, n_2, \ldots, n_k$, то *исправленная выборочная дисперсия* имеет вид

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{k}n_i(x_i - \overline{x}_в)^2. \qquad (2.13^*)$$

Для оценки среднего квадратичного отклонения генеральной совокупности используют *исправленное среднее квадратичное отклонение*

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x}_в)^2}{n-1}}, \qquad (2.15)$$

которое, однако, ***не является несмещенной оценкой***.

Из сопоставления оценок дисперсии (2.10) и (2.13), (2.10*) и (2.13*) видно, что они отличаются лишь знаменателями. Очевидно, что при достаточно

большом объеме выборки n выборочная дисперсия (2.10), (2.10*) и исправленная дисперсия (2.13), (2.13*) различаются незначительно. Исправленная дисперсия используется на практике при объеме выборки $n < 30$.

Для вычислений $D_в$ часто используется формула:

$$D_в = \overline{x_в^2} - (\overline{x}_в)^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2 = \frac{\sum_{i=1}^{k} n_i \cdot x_i^2}{n} - \left(\frac{\sum_{i=1}^{k} n_i x_i}{n}\right)^2. \qquad (2.16)$$

К показателям вариации относят также $R = x_{\max} - x_{\min}$ – *размах вариации* и $v = \frac{\sigma_в}{\overline{x}_в} \cdot 100\ \%$ $(\overline{x}_в \neq 0)$ – *коэффициент вариации*.

Коэффициент вариации – безразмерная характеристика. На практике считают, что если $v < 33\ \%$, то совокупность однородная.

Пример 1. Найти дисперсию и исправленную дисперсию по данному распределению выборки:

x_i	1	2	3	4
n_i	20	15	10	5

n=50

Решение. Вычислим сначала

$$\overline{x}_в = \frac{20 \cdot 1 + 15 \cdot 2 + 10 \cdot 3 + 5 \cdot 4}{50} = 2; \text{ и } \overline{x_в^2} = \frac{20 \cdot 1 + 15 \cdot 4 + 10 \cdot 9 + 5 \cdot 16}{50} = 5.$$

Тогда

$$D_в = \overline{x_в^2} - (\overline{x}_в)^2 = 1; \quad S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{k} n_i \cdot (x_i - \overline{x}_в)^2 = \frac{50}{49}.$$

В случае, когда первоначальные варианты x_i – большие числа, то целесообразно вычесть из всех вариант одно и то же число c, равное $\overline{x}_в$ или близкое к нему, т.е. перейти к условным вариантам $u_i = x_i - c$. При этом $Du_i = Dx_i$.

$$D_в(u_i) = \overline{u_i^2} - (\overline{u}_i)^2 = \frac{\sum_{i=1}^{k} n_i u_i^2}{n} - \left(\frac{\sum_{i=1}^{k} n_i u_i}{n}\right)^2 \qquad (2.17)$$

$$Mu_i = Mx_i - Mc = \overline{x}_в - c. \qquad (2.18)$$

Таким образом,

$$\overline{x}_в = c + \frac{\sum_{i=1}^{k} u_i \cdot n_i}{n}. \qquad (2.19)$$

Пример 2. По данному распределению выборки

x_i	1250	1270	1280
n_i	2	5	3

n=10

найти выборочную среднюю.

Решение. Поскольку первоначальные варианты – большие числа, то перейдем к условным вариантам. Выберем в качестве c=1270, тогда новые варианты u_i вычисляются по формулам $u_i = x_i - 1270$. Распределение выборки в условных вариантах принимает вид

u_i	-20	0	10
n_i	2	5	3

а выборочное среднее равно

$$\overline{x}_в = 1270 + \frac{-20 \cdot 2 + 10 \cdot 3}{10} = 1269.$$

Пример 3. Имеется распределение 80 предприятий по числу работающих на них (чел.):

x_i	150	250	350	450	550	650	750
n_i	1	3	7	30	19	15	5 .

Найти числовые характеристики распределения предприятий по числу работающих.

Решение. Признак X – число работающих (чел.) на предприятии. Для расчета характеристик данного распределения удобнее использовать таблицу:

Число работающих на предприятии, (x_i,чел.)	Число предприятий (n_i)	$x_i\, n_i$	$(x_i - \overline{x}_в)^2\, n_i$	$x_i^2\, n_i$
150	1	150	129600	22500
250	3	750	202800	187500
350	7	2450	179200	857500
450	30	13500	108000	6045000
550	19	10450	30400	5747500
650	15	9750	294000	6337500
750	5	3750	288000	2812500
Итого	80	40800	1232000	22040000 .

$\overline{x}_в = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i = \frac{40800}{80} = 510$ (чел.) – среднее число работающих на предприятии.

Дисперсию рассчитываем двумя способами.

1) по формуле (2.10) $D_в = \frac{1}{n}\sum_{i=1}^{k} (x_i - \overline{x}_в)^2 \cdot n_i = \frac{123200}{80} = 15400.$

2) по формуле (2.16)

$D_в = \overline{x^2} - \overline{x}_в^2$, где $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{k} x_i^2 \cdot n_i = \frac{22040000}{80} = 275500.$

$D_в = 275500 - (510)^2 = 15400.$

$\sigma_в = \sqrt{D_в} = \sqrt{15400} \approx 124$ (численность работающих на каждом предприятии отклоняется от средней численности в среднем на 124 чел.)

$R = x_{max} - x_{min} = 750 - 150 = 600$ (чел.).

$v = \frac{\sigma_в}{\overline{x}_в} \cdot 100\,\% = \frac{124}{510} \cdot 100\,\% \approx 24{,}3\,\%.$

Так как $\nu \approx 24{,}3\,\% < 33\%$, то исследуемая совокупность однородная.

Пример 4. Найти числовые характеристики распределения затрат времени на обработку одной детали (пример 2).

Решение. Признак X – затраты времени на обработку одной детали (мин) – непрерывный. Распределение задано интервальным рядом. Характеристики такого ряда находят по тем же формулам, что и для дискретного ряда, предварительно заменив интервальный ряд дискретным. Для этого для каждого интервала x_{i-1}–x_i вычисляют его середину x_i'. Расчеты представим в таблице:

Затраты времени на обработку 1 детали (X, мин): x_{i-1}–x_i	Число рабочих (n_i)	x_i'	$x_i' n_i$	$(x_i' - \overline{x}_в)^2 n_i$	$(x_i')^2 n_i$
22–24	2	23	46	50	1058
24–26	12	25	300	108	7500
26–28	34	27	918	34	24786
28–30	40	29	1160	40	33640
30–32	10	31	310	90	9610
32–34	2	33	66	50	2178
Итого	100	-	2800	372	78772 .

$\overline{x}_в = \frac{1}{n}\sum_{i=1}^{k} x_i' \cdot n_i = \frac{2800}{100} = 28$ (мин) – среднее время на обработку одной детали.

Дисперсию рассчитываем двумя способами.

1) $D_в = \frac{1}{n}\sum_{i=1}^{k}(x_i' - \overline{x}_в)^2 \cdot m_i = \frac{372}{100} = 3{,}72;$

2) $D_в = \overline{(x')^2} - \overline{x}_в^2$, где $\overline{(x')^2} = \frac{1}{n}\sum_{i=1}^{k}(x_i')^2 \cdot m_i = \frac{78772}{100} = 787{,}72;$

$D_в = 787{,}72 - (28)^2 = 3{,}72.$

$\sigma_в = \sqrt{D_в} = \sqrt{3{,}72} \approx 1{,}93$ (мин), то есть затраты времени на обработку одной детали каждым рабочим отклоняются от средних затрат времени в среднем на 1,93 мин.

$R = x_{max} - x_{min} = 34 - 22 = 12$ (мин).

$v = \frac{\sigma_в}{\overline{x}_в} \cdot 100\,\% = \frac{1{,}93}{28} \cdot 100\,\% \approx 6{,}9\,\%$ – совокупность однородная.

Интервальное оценивание (доверительные интервалы)

Выше был рассмотрен вопрос об оценке неизвестного параметра θ одним числом θ^*, т.е. о *точечной* оценке. В ряде задач требуется не только найти для параметра θ подходящее численное значение, но и оценить его *точность* и *надежность*. Требуется знать, к каким ошибкам может привести замена θ его точечной оценкой θ^*, и с какой степенью уверенности можно ожидать, что эти оценки не выйдут за известные пределы.

Такого рода задачи особенно актуальны при малом числе наблюдений, когда точечная оценка θ^* в значительной мере случайна и приближенная замена θ на θ^* может привести к серьезным ошибкам. Для определения точности и надежности θ^* в МС вводят понятие *доверительного интервала* и *доверительной вероятности*. Часто из физических соображений делается вывод, что ξ распределена по нормальному закону с плотностью вероятности

$$f_\xi = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}, \quad a = M_\xi, \quad \sigma = \sqrt{D\xi}.$$

Возникает задача оценки параметров a и σ или одного из них, если известны наблюдаемые значения $x_1, x_2, ..., x_n$ СВ ξ.

Пусть для параметра θ из опыта получена несмещенная оценка θ^*.
Оценим возможную при этом ошибку. Назначим некоторую достаточно большую вероятность γ ($\gamma = 0{,}95;\ 0{,}99;\ 0{,}9$) такую, что событие с вероятностью γ можно считать практически достоверным. Найдём такое значение ε, $\varepsilon > 0$, для которого вероятность отклонения оценки на величину, не превышающую ε, равна γ:

$$P\left(\left|\theta^* - \theta\right| < \varepsilon\right) = \gamma. \qquad (2.20)$$

Тогда диапазон практически возможных значений ошибки, возника-ющей при замене θ на θ^*, будет равен $\pm\varepsilon$. Большие по абсолютной величине ошибки будут появляться с малой вероятностью $\alpha = 1 - \gamma$.

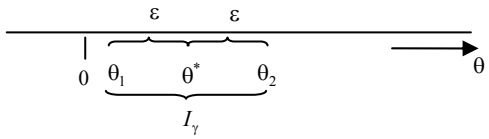
Перепишем уравнение (2.20) в виде:

$$P(\theta^* - \varepsilon < \theta < \theta^* + \varepsilon) = \gamma. \qquad (2.21)$$

Равенство (2.21) означает, что с вероятностью γ неизвестное значение параметра θ попадает в интервал I_γ, равный

$$I_\gamma = (\theta^* - \varepsilon;\ \theta^* + \varepsilon), \qquad (2.22)$$

который является случайным, т.к. случайным является центр θ^* интервала I_γ. Случайной является и его длина, равная 2ε, т.к. ε, как правило, вычисляется по опытным данным. Поэтому в (2.21) величину γ лучше толковать не как вероятность γ попадания точки θ в интервал I_γ, а как вероятность того, что случайный интервал I_γ накроет точку θ:



θ^* – центр доверительного интервала, $\theta_1 = \theta^* - \varepsilon$, $\theta_2 = \theta^* + \varepsilon$.

Вероятность γ принято называть *доверительной вероятностью* (надежностью), а интервал I_γ – *доверительным интервалом*.

Интервал (θ_1, θ_2) будем называть *доверительным* для оценки параметра θ при заданной доверительной вероятности γ или при заданном уровне значимости $\alpha = 1 - \gamma$, если он с вероятностью γ "накрывает" оцениваемый параметр θ, т.е.

$$P(\theta \in (\theta_1, \theta_2)) = P(\theta_1 < \theta < \theta_2) = \gamma. \qquad (2.23)$$

Границы интервала θ_1 и θ_2 называют *доверительными границами*. Доверительный интервал можно рассматривать как интервал значений параметра θ, совместимых с опытными данными и не противоречащих им. Метод доверительных интервалов был разработан Ю.Нейманом*, который использовал идеи Р.Фишера**.

Рассмотрим вопрос о нахождении доверительных границ θ_1 и θ_2. Пусть для параметра θ имеется несмещённая оценка θ^*. Если бы был известен закон распределения величины θ^*, задача нахождения доверительного интервала была бы весьма простой. Для этого достаточно было бы найти такое значение ε, для которого выполнено соотношение (2.20). Сложность состоит в том, что закон распределения оценки θ^* зависит от закона распределения СВ ξ, следовательно, от его неизвестных параметров, в частности, от параметра θ.

Доверительные интервалы для оценки неизвестного математического ожидания нормального распределения при известном σ

Пусть количественный признак ξ генеральной совокупности распределен нормально, причем известно σ – среднее квадратичное отклонение этого распределения. Оценим неизвестное математическое ожидание a по выборочной средней $\bar{x}_в$, т.е. найдем доверительные интервалы, покрывающие параметр a с надежностью γ. Будем рассматривать $\bar{x}_в$ как СВ $\bar{\xi}$ (т.е. $\bar{x}_в$ меняется от выборки к выборке), а выборочные значения признака $x_1, x_2, \ldots x_n$ – как одинаково распределенные (т.е. имеющие одну и ту же функцию распределения $F(x)$) независимые в совокупности случайные величины $\xi_1, \xi_2, \ldots \xi_n$ (эти числа также изменяются от выборки к выборке). Тогда

* Ю.Нейманом (1894-1981) – американский математик-статистик
** Р.Фишера (1890-1962) – английский статистик и генетик

математические ожидания каждой из величин $x_1, x_2, \ldots x_n$ одинаковы и равны a, т.е. $M_{x_i} = a, \quad \sigma_{x_i} = \sigma, \ i = \overline{1,n}.$

Известно, что если СВ ξ распределена нормально, то выборочная средняя также распределена нормально и

$$M_{\bar{\xi}} = a, \ D_{\bar{\xi}} = \frac{\sigma^2}{n}, \ \sigma_{\bar{\xi}} = \frac{\sigma}{\sqrt{n}}.$$

Потребуем выполнение соотношения $P\left(|\xi - a| < \delta\right) = \gamma$, где γ – заданная надежность. Поскольку для нормально распределенной СВ ξ $P(|\xi - a| < \delta) = \Phi(\frac{\delta}{\sigma})$, то, сделав замену $\xi \to \bar{\xi}, \quad \sigma \to \sigma_{\bar{\xi}} = \frac{\sigma}{\sqrt{n}}$, получим $P(|\bar{\xi} - a| < \delta) = \Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right) = \Phi(t_\gamma)$, где $t_\gamma = \frac{\delta\sqrt{n}}{\sigma}$. Следовательно, $\delta = \frac{t_\gamma \sigma}{\sqrt{n}}$. Таким образом, $P(|\bar{\xi} - a| < \frac{t_\gamma \sigma}{\sqrt{n}}) = \Phi(t_\gamma)$. Так как вероятность задана и равна γ, то, заменив $\bar{\xi}$ на $\bar{x}_в$, получим

$$P\left(\bar{x}_в - \frac{t_\gamma \sigma}{\sqrt{n}} < a < \bar{x}_в + \frac{t_\gamma \sigma}{\sqrt{n}}\right) = \Phi\left(t_\gamma\right) = \gamma. \qquad (2.24)$$

Таким образом, с доверительной вероятностью γ (надежностью γ) можно утверждать, что доверительный интервал $\left(\bar{x}_в - \frac{t_\gamma \sigma}{\sqrt{n}}; \bar{x}_в + \frac{t_\gamma \sigma}{\sqrt{n}}\right)$ покрывает неизвестное математическое ожидание a нормально распределенной СВ с известным среднеквадратичным отклонением σ с точностью $\delta = \frac{t_\gamma \sigma}{\sqrt{n}}$. Число t_γ определяется из соотношения $\Phi(t_\gamma) = \gamma$, где $\Phi(x)$ – функция Лапласа. По таблице значений функции Лапласа находим аргумент t_γ, которому соответствует значение функции Лапласа, равное γ.

Пример 3. СВ ξ распределена нормально с $\sigma = 3$. Найти доверительный интервал для оценки неизвестного математического ожидания a по выборочной средней $\bar{x}_в$, если объем выборки $n = 36$ и задана надежность оценки $\gamma = 0.95$.

Решение. Вычислим $\Phi\left(t_\gamma\right) = \gamma = 0.95$. По таблице значений функции Лапласа находим $t_\gamma = 1.96$; следовательно, точность оценки $\delta = \frac{t_\gamma \sigma}{\sqrt{n}} = \frac{1.96 \cdot 3}{6} = 0.98$. Тогда доверительный интервал $(\bar{x}_в - \delta, \bar{x}_в + \delta)$ имеет вид

$(\bar{x}_в - 0.98, , \bar{x}_в + 0.98)$. Таким образом, с вероятностью 0.95 СВ ξ попадет в интервал $(\bar{x}_в - 0.98, \bar{x}_в + 0.98)$.

Смысл заданной надежности: надежность $\gamma = 0.95$ означает, что если проведено достаточно большое число выборок, то 95% из них определяют такие доверительные интервалы, в которых действительно заключен параметр a; в 5% случаев параметр a может выйти за границы доверительного интервала.

Пример 4. С целью определения среднего трудового стажа на предприятии методом случайной повторной выборки проведено обследование трудового стажа рабочих. Из всего коллектива рабочих завода случайным образом выбрано 400 рабочих, данные о трудовом стаже которых и составили выборку. Средний по выборке стаж оказался равным 9,4 года. Считая, что трудовой стаж рабочих имеет нормальный закон распределения, определить с вероятностью 0,97 границы, в которых окажется средний трудовой стаж для всего коллектива, если известно, что $\sigma = 1{,}7$ года.

Решение. Признак X – трудовой стаж рабочих. Этот признак имеет нормальный закон распределения с известным параметром $\sigma = 1{,}7$, параметр a неизвестен. Сделана выборка объемом $n = 400$, по данным выборки найдена точечная оценка параметра a: $\bar{x}_в = 9{,}4$. С надежностью $\gamma = 0{,}97$ найдем интервальную оценку параметра a по формуле:

$$\bar{x}_в - \frac{t \cdot \sigma}{\sqrt{n}} < a < \bar{x}_в + \frac{t \cdot \sigma}{\sqrt{n}}.$$

По таблице значений функции Лапласа $\Phi(t) \approx \frac{0{,}97}{2} = 0{,}485$ находим $t = 2{,}17$; тогда: $9{,}4 - \frac{2{,}17 \cdot 1{,}7}{\sqrt{400}} < a < 9{,}4 + \frac{2{,}17 \cdot 1{,}7}{\sqrt{400}}$,

$9{,}4 - 0{,}18 < a < 9{,}4 + 0{,}18$. Итак, $9{,}22 < a < 9{,}58$, то есть средний трудовой стаж рабочих всего коллектива лежит в пределах от 9,22 года до 9,58 года (с надежностью $\gamma = 0{,}97$).

С изменением надежности γ изменится и интервальная оценка.

Пусть $\gamma = 0{,}99$, тогда $\Phi(t) = 0{,}495$, отсюда $t = 2{,}58$. Тогда:

$9{,}4 - \frac{2{,}58 \cdot 1{,}7}{20} < a < 9{,}4 + \frac{2{,}58 \cdot 1{,}7}{20}$, или $9{,}4 - 0{,}22 < a < 9{,}4 + 0{,}22$.

Окончательно: $9{,}18 < a < 9{,}62$.

Доверительные интервалы для оценки математического ожидания нормального распределения при неизвестном σ

Пусть количественный признак ξ генеральной совокупности распределен по нормальному закону, причем σ неизвестно. Требуется оценить неизвестное математическое ожидание a с помощью доверительных интервалов с заданной доверительной вероятностью (надежностью) γ.

Воспользоваться результатами предыдущего раздела нельзя, т.к. параметр σ в данном случае неизвестен. Для решения задачи по выборке $x_1, x_2, \ldots x_n$ вычислим $\bar{x}_в$ и исправленную дисперсию S^2. Используя распределение Стьюдента, можно показать, что

$$P\left(\bar{x}_в - \frac{t_{\gamma,n}S}{\sqrt{n}} < a < \bar{x}_в + \frac{t_{\gamma,n}S}{\sqrt{n}}\right) = \gamma. \qquad (2.25)$$

Таким образом, доверительный интервал $\left(\bar{x}_в - \frac{t_{\gamma,n}S}{\sqrt{n}}; \bar{x}_в + \frac{t_{\gamma,n}S}{\sqrt{n}}\right)$ покрывает неизвестный параметр a с надежностью γ. По таблице распределения Стьюдента по заданным n и γ можно найти $t_{\gamma,n}$ ($t_{\gamma,n} = t(\gamma, n)$).

Пример 5. Количественный признак ξ генеральной совокупности распределен нормально. По выборке объема n=16 найдены $\bar{x}_в = 20.2$ и исправленное среднее квадратичное отклонение $S = 0.8$. Оценить неизвестное математическое ожидание a при помощи доверительного интервала с надежностью γ=0.95.

Решение. Для нахождения доверительного интервала следует по таблице t-распределения Стьюдента по n=16, γ=0.95 найти $t_{\gamma,n}$=2.13. Тогда границы доверительного интервала имеют вид

$$\bar{x}_в - t_{\gamma,n}\frac{S}{\sqrt{n}} = 20.2 - \frac{2.13 \cdot 0.8}{4} = 19.774; \; \bar{x}_в + t_{\gamma,n}\frac{S}{\sqrt{n}} = 20.2 + \frac{2.13 \cdot 0.8}{4} = 20.626.$$

Таким образом, с надежностью γ=0.95 неизвестный параметр a заключается в доверительном интервале (19.774, 20.626).

Пример 6. С целью определения средней продолжительности рабочего дня на предприятии методом случайной повторной выборки проведено обследование продолжительности рабочего дня сотрудников. Из всего коллектива завода случайным образом выбрано 30 сотрудников. Данные табельного учета о продолжительности рабочего дня этих сотрудников и составили выборку. Средняя по выборке продолжительность рабочего дня оказалась равной 6,85 часа, а $S = 0{,}7$ часа. Считая, что продолжительность рабочего дня имеет нормальный закон распределения, с надежностью $\gamma = 0{,}95$ определить, в каких пределах находится действительная средняя продолжительность рабочего дня для всего коллектива данного предприятия.

Решение. Признак X – продолжительность рабочего дня. Признак имеет нормальное распределение с неизвестными параметрами. Сделана выборка объемом $n = 30$, по выборочным данным найдены точечные оценки параметров распределения: $\bar{x}_в = 6{,}85$; $S = 0{,}7$. С надежностью $\gamma = 0{,}95$ найдем интервальную оценку параметра a по формуле:

$$\bar{x}_в - \frac{t_{\gamma,n} \cdot S}{\sqrt{n}} < a < \bar{x}_в + \frac{t_{\gamma,n} \cdot S}{\sqrt{n}},$$

$t_{\gamma,n}$ находим по таблице t-распределения Стьюдента $t_{\gamma,n} = t(0{,}95;\ 30)=2{,}045$. Тогда:

$$6{,}85-\frac{2{,}045\cdot 0{,}7}{\sqrt{30}}<a<6{,}85+\frac{2{,}045\cdot 0{,}7}{\sqrt{30}},\ \text{или}\ 6{,}85-0{,}26<a<6{,}85+0{,}26\ .$$

Итак, $6{,}59<a<7{,}11$, то есть с надежностью $\gamma=0{,}95$ средняя продолжительность рабочего дня для всего коллектива лежит в пределах от 6,59 до 7,11 ч.

Определение объема выборки

Для определения необходимого объема выборки, при котором с заданной вероятностью γ можно утверждать, что выборочная средняя отличается от генерального по абсолютной величине меньше чем на δ, пользуются формулами:

а) в случае известной дисперсии из формулы (2.24):

$$n=\frac{t_\gamma{}^2\sigma^2}{\delta^2},\qquad (2.26)$$

где $\Phi\left(t_\gamma\right)=\gamma$.

б) в случае неизвестной дисперсии организуют специальную «пробную» выборку небольшого объема, находят оценку S^2 и, полагая $\sigma^2\approx S^2$, находят объем «основной» выборки:

$$n=\frac{t_\gamma{}^2S^2}{\delta^2},\qquad (2.27)$$

Пример 7. Найти минимальный объем выборки, на основе которой можно было бы оценить математическое ожидание СВ с ошибкой, которая не превышает 0.2 и надежностью 0.98, если допускается что СВ имеет нормальное распределение с $\sigma=4$.

Решение. Из равенства $\Phi(t_\gamma)=0.98$ по таблице определяют $t_\gamma=2.33$. По формуле (2.30) находим: $n=\frac{t_\gamma{}^2\sigma^2}{\delta^2}=\frac{2.33^2\cdot 16}{0.2^2}\approx 2171$.

Тема 3

Проверка статистических гипотез

С теорией статистического оценивания параметров тесно связана проверка статистических гипотез. Она используется в том случае, когда необходим обоснованный вывод о преимуществах того или иного способа вложения инвестиций, об уровне доходности ценных бумаг, об эффективности лекарственных препаратов, о значимости построенной математической модели и т.д.

При изучении многих статистических данных необходимо знать *закон* распределения генеральной совокупности. Если закон распределения неизвестен и есть основания предположить, что он имеет определенный вид (например, *A*), то выдвигают гипотезу: генеральная совокупность распределена по закону *A*. В данной гипотезе речь идет о *виде* предполагаемого распределения.

Возможен случай, когда закон распределения известен, а его параметры неизвестны. Если есть основания предположить, что неизвестный параметр θ равен определенному значению θ_0, то выдвигают гипотезу: $\theta = \theta_0$. Здесь речь идет о *предполагаемой величине параметра* одного известного распределения. Возможны гипотезы о равенстве параметров
двух или нескольких распределений, о независимости выборок и др.

Все выводы, которые делаются в МС, вообще говоря, являются гипотезами, т.е. предположениями о неизвестных параметрах известных распределений, об общем виде неизвестного теоретического распределения или функции распределения изучаемой СВ. Такие гипотезы называют *статистическими гипотезами*.

Различают *простые* и *сложные, параметрические* и *непараметрические* статистические гипотезы.

Статистическая гипотеза называется *простой*, если она однозначно определяет закон распределения СВ. *Сложной* называют гипотезу, состоящую из конечного или бесконечного числа простых гипотез. Например, гипотезы "вероятность появления события A в схеме Бернулли равна ⅓", "закон распределения СВ – нормальный с параметрами $a = 0,\ \sigma^2 = 1$" являются *простыми* в отличие от *сложных* гипотез: "вероятность появления события A в схеме Бернулли заключена между ⅓ и ½", "закон распределения СВ не является нормальным". Гипотеза называется *параметрической*, если в ней содержится некоторое условие о значении параметра известного распределения. Гипотезу, в которой сформулированы предположения относительно вида распределения, называют *непараметрической*.

Если исследовать всю генеральную совокупность, то, естественно, можно было бы наиболее точно установить справедливость выдвигаемой гипотезы. Однако такое исследование не всегда возможно, и суждение об истинности статистических гипотез проверяется на основании выборки.

Выдвигаемую (проверяемую) гипотезу называют *основной* или *нулевой* гипотезой H_0. Если, например, по полигону или гистограмме частот, построенным по некоторой выборке, можно предположить, что СВ распределена по нормальному закону, то может быть выдвинута гипотеза $H_0 : a = a_0,\ \sigma = \sigma_0$. Одновременно с гипотезой H_0 выдвигается *альтернативная* (*конкурирующая*) гипотеза H_1. Если гипотеза H_0 будет отвергнута, то имеет место конкурирующая ей гипотеза.

Конкурирующей (*альтернативной*) называют гипотезу H_1, являющуюся логическим отрицанием H_0. *Нулевая* H_0 и *альтернативная* H_1 гипотезы представляют собой две возможности выбора, осуществляемого в задачах проверки статистических гипотез. Например, если $H_0 : \theta = \theta_0$, то альтернативная гипотеза может иметь вид $H_1 : \theta \neq \theta_0$, $H_1 : \theta > \theta_0$, или $H_1 : \theta < \theta_0$.

Выдвинутая гипотеза может быть правильной или неправильной, в связи с чем возникает необходимость ее проверки. Поскольку проверку осуществляют статистическими методами, ее называют *статистической.* В результате статистической проверки гипотезы неправильное решение может быть принято в двух случаях: с одной стороны, на основании результатов опыта можно отвергнуть правильную гипотезу; с другой – можно принять неверную гипотезу. Очевидно, последствия этих ошибок могут оказаться различными. Отметим, что правильное решение может быть принято также в двух случаях:

1) гипотеза принимается, и она в действительности является правильной;
2) гипотеза отвергается, и она в действительности не верна.

По полученным значениям статистики основная гипотеза принимается или отклоняется. При этом, так как выборка носит случайный характер, могут быть допущены два вида ошибок:

–может быть отвергнута правильная гипотеза, в этом случае допускается *ошибка первого рода,*

–может быть принята неверная гипотеза, тогда допускается *ошибка второго рода* (см. схему).

H_0	H_0 – принимается	H_0 – отвергается
верна	правильное решение	ошибка I рода
ошибочна	ошибка II рода	правильное решение

Вероятность α *совершить ошибку I рода*, т.е. отвергнуть гипотезу H_0, когда она верна, называется *уровнем значимости* критерия.

Обычно принимают $\alpha = 0.1, 0.05, \cdots, 0.01$. Смысл α: при $\alpha = 0.05$ в 5 случаях из 100 имеется риск допустить ошибку I рода, т.е. отвергнуть правильную гипотезу. *Вероятность допустить ошибку II рода*, т.е. принять гипотезу H_0, когда она неверна, обозначают β.

Вероятность $1 - \beta$ не допустить ошибку II рода, т.е. отвергнуть гипотезу H_0, когда она ошибочна, называется *мощностью критерия.*

Используя терминологию статистического контроля качества продукции можно сказать, что вероятность α представляет "*риск поставщика*" (или *"риск производителя"*), связанный с вероятностью признать негодной по результатам выборочного контроля всю партию годных изделий, удовлетворяющих стандарту, а вероятность β – "*риск потребителя*", связанный с вероятностью принять по анализу выборки негодную партию, не удовлетворяющую стандарту. В некоторых прикладных исследованиях ошибка I рода α означает вероятность того, что сигнал, предназначенный наблюдателю, не будет принят, а ошибка II рода β – вероятность того, что наблюдатель примет ложный сигнал.

Для проверки справедливости нулевой гипотезы H_0 используют специально подобранную СВ K, точное или приближенное распределение которой известно. Эту СВ K, которая служит для проверки нулевой гипотезы H_0, называют *статистическим критерием* (или просто *критерием*).

Для проверки статистической гипотезы по данным выборок вычисляют частные значения входящих в критерий величин и получают частное (*наблюдаемое*) значение критерия $K_{набл}$.

После выбора определенного статистического критерия для решения вопроса о принятии или непринятии гипотезы множество его возможных значений разбивают на два непересекающихся подмножества, одно из которых называется *областью принятия гипотезы* (или *областью допустимых значений критерия*), а второе – *критической областью*.

Критической областью называется совокупность значений статистического критерия K, при которых нулевую гипотезу H_0 отвергают.

Областью принятия гипотезы (*областью допустимых значений критерия)* называется совокупность значений статистического критерия K, при которых нулевую гипотезу H_0 принимают.

Основной принцип проверки статистических гипотез. *Если наблюдаемое значение $K_{набл}$ статистического критерия K принадлежит критической области, то основная гипотеза отвергается в пользу альтернативной; если оно принадлежит области принятия гипотезы, то гипотезу принимают*.

Поскольку статистический критерий K – одномерная СВ, то все ее возможные значения принадлежат некоторому интервалу. Следовательно, и критическая область, и область принятия гипотезы – также интервалы. Тогда должны существовать точки, их разделяющие.

Критическими точками (границами) $k_{кр}$ называют точки, отделяющие критическую область от области принятия гипотезы.

В отличие от рассмотренного в 1.4 интервального оценивания параметров, в котором имелась лишь одна возможность ошибки – получение доверительного интервала, не накрывающего оцениваемый параметр – при проверке статистических гипотез возможна двойная ошибка (как I рода α, так и II рода β). Вероятности оценок I и II рода (α и β) однозначно определяются выбором

критической области. Естественным является желание сделать α и β сколь угодно малыми. Однако эти требования являются противоречивыми, ибо при фиксированном объеме выборки можно сделать сколь угодно малой лишь одну из величин – α или β, что сопряжено с неизбежным увеличением другой. *Одновременное уменьшение вероятностей* α и β возможно лишь при увеличении объема выборки. При разработке статистических критериев необходимо уменьшать как ошибку I рода, так и ошибку II рода.

Поскольку одновременное уменьшение ошибок I и II рода невозможно, то при нахождении критических областей для данной статистики уровень значимости задают, стараясь подобрать такой критерий, чтобы вероятность ошибки II рода была наименьшей.

Различают *одностороннюю* (правостороннюю и левостороннюю) и *двустороннюю* критические области.

Правосторонней называют критическую область, определяемую неравенством $K > k_{кр}$, где $k_{кр} > 0$.

Левосторонней называют критическую область, определяемую неравенством $K < k_{кр}$, где $k_{кр} < 0$.

Двусторонней называют критическую область, определяемую неравенствами $K < k_1$, $K > k_2$, где $k_2 > k_1$.

Если критические точки симметричны относительно нуля, то двусторонняя критическая область определяется неравенствами $K < -k_{кр}$, $K > k_{кр}$, где $k_{кр} > 0$ или, что равносильно, $|K| > k_{кр}$.

Как найти критическую область? Пусть $K = K(x_1, x_2, \ldots, x_n)$ – статистический критерий, выбранный для проверки нулевой гипотезы H_0, k_0 – некоторое число, $k_0 \in R$. Найдем правостороннюю критическую область, определяемую неравенством $K > k_{кр}$, где $k_{кр} > 0$. Для ее отыскания достаточно найти критическую точку $k_{кр}$. Рассмотрим вероятность $P(K > k_0)$ в предположении, что гипотеза H_0 верна. Очевидно, что с ростом k_0 вероятность $P(K > k_0)$ уменьшается. Тогда k_0 можно выбрать настолько большим, что вероятность $P(K > k_0)$ станет ничтожно малой. Другими словами, при заданном уровне значимости α можно определить критическое значение $k_{кр}$ из неравенства $P(K > k_{кр}) = \alpha$.

Критическую точку $k_{кр}$ *ищут из требования, чтобы при условии справедливости нулевой гипотезы* H_0 *вероятность того, что критерий* K *примет значение, большее* $k_{кр}$*, была равна принятому уровню значимости* α*:*

$$P(K > k_{кр}) = \alpha. \qquad (3.1)$$

Для каждого из известных статистических критериев (нормального, Стьюдента, критерия Пирсона χ^2, Фишера-Снедекора, Кочрена и др.) имеются

соответствующие таблицы, по которым находят $k_{кр}$, удовлетворяющее этим требованиям. После нахождения $k_{кр}$ по данным выборок вычисляют реализовавшееся (наблюдаемое) значение $K_{набл}$ критерия K. Если окажется, что $K_{набл} > k_{кр}$, (т.е. реализовалось маловероятное событие), то нулевая гипотеза H_0 отвергается. Следовательно, принимается конкурирующая гипотеза H_1. Если же $K_{набл} < k_{кр}$, то в этом случае нет оснований отвергнуть выдвинутую гипотезу H_0. Следовательно, гипотеза H_0 принимается. Другими словами, *выдвинутая статистическая гипотеза согласуется с результатами эксперимента* (выборочными данными).

Левосторонняя критическая область определяется неравенством $K < k_{кр}$, где $k_{кр} < 0$. *Критическую точку* $k_{кр}$ *находят из требования, чтобы при условии справедливости нулевой гипотезы* H_0 *вероятность того, что критерий* K *примет значение, меньшее* $k_{кр}$*, была равна принятому уровню значимости* α:

$$P(K < k_{кр}) = \alpha. \qquad (3.2)$$

Двусторонняя критическая область определяется неравенствами $K < k_1$, $K > k_2$, где $k_2 > k_1$. *Критические точки* k_1, k_2 *находят из требования, чтобы при условии справедливости нулевой гипотезы* H_0 *сумма вероятностей того, что критерий* K *примет значение, меньшее* k_1 *или большее* k_2*, была равна принятому уровню значимости* α*:*

$$P(K < k_1) + P(K > k_2) = \alpha. \qquad (3.3)$$

Если распределение критерия симметрично относительно нуля, и для увеличения его мощности выбрать симметричные относительно нуля точки $-k_{кр}$ и $k_{кр}$, $k_{кр} > 0$, то $P(K < -k_{кр}) = P(K > k_{кр})$, и из $P(K < k_1) + \ + P(K > k_2) = \alpha$ следует

$$P(K > k_{кр}) = \alpha/2. \qquad (3.4)$$

Это соотношение и служит для отыскания критических точек двусторонней критической области.

Отметим, что *принцип проверки статистической гипотезы не дает логического доказательства ее верности или неверности. Принятие гипотезы* H_0 *следует расценивать не как раз и навсегда установленный, абсолютно верный содержащийся в ней факт, а лишь как достаточно правдоподобное, не противоречащее опыту утверждение.*

Если проверка статистических гипотез основана на предположении об известном законе распределения генеральной совокупности, из которого следует определенное распределение критерия, то критерии проверки таких гипотез называют *параметрическими критериями*. Если закон распределения генеральной совокупности неизвестен, то соответствующие критерии называются *непараметрическими.* Понятно, что непараметрические критерии обладают значительно меньшей мощностью, чем параметрические. Отсюда

следует, что для сохранения той же мощности при использовании непараметрического критерия по сравнению с параметрическим необходимо иметь значительно больший объем наблюдений.

Наиболее распространенным критерием проверки статистических гипотез о виде распределения генеральной совокупности (т.е. непараметрическим критерием) является *критерий Пирсона* χ^2.

Проверка гипотез о среднем значении нормально распределенной СВ при известной и неизвестной дисперсии

Пусть имеется генеральная совокупность X, распределенная по нормальному закону с *известной дисперсией* $D(X)=\sigma^2$ (т.е. σ известно). Генеральная средняя a неизвестна, но есть основания предполагать, что она равна гипотетическому (предполагаемому) значению a_0. Например, если X – совокупность размеров x_i партии деталей, изготавливаемых станком-автоматом, то можно предполагать, что генеральная средняя a этих размеров равна проектному размеру a_0. Для проверки этого предположения (гипотезы) делают выборку, находят $\overline{x}_в$ и устанавливают, *значимо* или *незначимо* различаются $\overline{x}_в$ и a_0. Если различие окажется незначимым, то станок в среднем обеспечивает проектный размер; если же различие значимое, то станок требует наладки.

Из нормальной генеральной совокупности X извлечем выборку $x_1,\ldots,x_n$ объема n, по которой найдем $\overline{x}_в$. При этом дисперсия σ^2 известна. Поскольку предполагается, что $x_1,\ldots,x_n$ как СВ $X_1,\ldots,X_n$ взаимно независимы, то они имеют одинаковые нормальные распределения, а следовательно, и одинаковые характеристики (математическое ожидание, дисперсию, и т.д.).

Необходимо по известному $\overline{x}_в$ при заданном уровне значимости α проверить гипотезу $H_0: a=a_0$ о равенстве генеральной средней a гипотетическому значению a_0.

Поскольку $\overline{x}_в$ является несмещенной оценкой генеральной средней, т.е. $M(\overline{X})=a$, то гипотезу $H_0: a=a_0$ можно записать в виде $H_0: M(\overline{X})=a_0$. Таким образом, требуется проверить, что математическое ожидание выборочной средней $\overline{X}$ равно гипотетической генеральной средней a_0, т.е. *значимо* или *незначимо* различаются выборочная $\overline{X}$ и генеральная a_0 средние.

В качестве *критерия проверки* гипотезы H_0 примем СВ $U=\frac{\overline{X}-a_0}{\sigma(\overline{X})}$. В силу свойства $\sigma(\overline{X})=\frac{\sigma}{\sqrt{n}}$ одинаково распределенных взаимно независимых СВ (см. раздел 1.4.1) критерий проверки гипотезы H_0 принимает вид $U=\frac{\overline{X}-a_0}{\sigma}\cdot\sqrt{n}$. Случайная величина U распределена по стандартному нормальному закону $N(0;1)$ (т.е. с $a=0, \sigma=1$). Критическая область строится в зависимости от вида

конкурирующей гипотезы H_1. Сформулируем правила проверки гипотезы H_0, обозначив через $U_{набл}$ значение критерия U, вычисленное по данным наблюдений.

Правило 1. Для того чтобы при заданном уровне значимости α проверить гипотезу $H_0: a = a_0$ о равенстве неизвестной генеральной средней a нормальной совокупности с известной дисперсией σ^2 гипотетическому значению a_0 при конкурирующей гипотезе $H_1: a \neq a_0$, необходимо вычислить

$$U_{набл} = \frac{\bar{x}_в - a_0}{\sigma} \cdot \sqrt{n} \qquad (3.5)$$

и по таблице значений функции Лапласа найти критическую точку *двусторонней критической области* из равенства

$$\Phi\left(u_{кр}\right) = 1 - \alpha\,. \qquad (3.6)$$

Если $\left|U_{набл}\right| < u_{кр}$ – нет оснований отвергнуть гипотезу H_0; если $\left|U_{набл}\right| > u_{кр}$ – гипотезу H_0 отвергают.

Правило 2. При конкурирующей гипотезе $H_1: a > a_0$ критическую точку $u_{кр}$ *правосторонней критической области* находят из равенства

$$\Phi\left(u_{кр}\right) = 1 - 2\alpha\,. \qquad (3.7)$$

Если $U_{набл} < u_{кр}$ – нет оснований отвергнуть гипотезу H_0; если $U_{набл} > u_{кр}$ – гипотезу H_0 отвергают.

Правило 3. При конкурирующей гипотезе $H_1: a < a_0$ критическую точку $u_{кр}$ находят по правилу 2, а затем полагают границу *левосторонней критической области* $u'_{кр} = -u_{кр}$. Если $U_{набл} > -u_{кр}$ – нет оснований отвергнуть гипотезу H_0; если $U_{набл} < -u_{кр}$ – гипотезу H_0 отвергают.

Замечание. Из правила 1 следует, что если область принятия гипотезы H_0 есть интервал $-u_{кр} < U_{набл} < u_{кр}$, то область ее отклонения – $U \in (-\infty; u_{кр}) \cup (u_{кр}; +\infty)$

Пример 1. Из нормальной генеральной совокупности с известным $\sigma = 0.49$ извлечена выборка объема $n = 49$ и по ней найдено выборочное среднее $\bar{x}_в = 21.7$. При уровне значимости $\alpha = 0.05$ проверить гипотезу $H_0: a = a_0 = 21$ при конкурирующей гипотезе $H_1: a > 21$.

Решение. По данным задачи найдем $U_{набл} = \frac{\bar{x}_в - a_0}{\sigma} \cdot \sqrt{n} = \frac{21.7 - 21}{0.49} \cdot \sqrt{49} = 10$. Поскольку конкурирующая гипотеза H_1 имеет вид $H_1: a > 21$, то критическая область – правосторонняя. По правилу 2 критическую точку $u_{кр}$ находим из равенства $\Phi\left(u_{кр}\right) = 1 - 2\alpha = 1 - 2 \cdot 0.05 = 0.9$. По таблице значений функции Лапласа находим $u_{кр} = 1.65$. Так как $U_{набл} = 10 > 1.65$, то гипотезу H_0 отвергаем. Таким образом, различие между выборочной и гипотетической генеральной средней значимое.

Рассмотрим случай, когда *дисперсия* $D(X)=\sigma^2$ генеральной совокупности, распределенной по нормальному закону, *неизвестна* (т.е. σ неизвестно). Такая ситуация может возникнуть, например, в случае малых выборок. В качестве проверки гипотезы H_0 принимают СВ

$$T=\frac{\overline{X}_в-a_0}{S}\cdot\sqrt{n-1}\,, \qquad (3.8)$$

где S – "исправленное" среднее квадратическое отклонение. Случайная величина T имеет распределение Стьюдента с $k=n-1$ степенями свободы. Критическая область, как и в рассмотренном выше случае с известной дисперсией $D(X)=\sigma^2$, строится в зависимости от вида конкурирующей гипотезы.

Правило 1. Для того, чтобы при заданном уровне значимости α проверить гипотезу $H_0: a=a_0$ о равенстве неизвестной генеральной средней a нормальной совокупности с неизвестной дисперсией σ^2 гипотетическому значению a_0 при конкурирующей гипотезе $H_1: a\neq a_0$, необходимо вычислить

$$T_{набл}=\frac{\overline{x}_в-a_0}{s}\cdot\sqrt{n-1} \qquad (3.9)$$

и по таблице критических точек распределения Стьюдента, по заданному уровню значимости α, помещенному в верхней строке таблицы, и числу степеней свободы $k=n-1$ найти критическую точку $t_{двуст.кр.}(\alpha;k)$ *двусторонней критической области*. Если $|T_{набл}|<t_{двуст.кр}$ – нет оснований отвергнуть гипотезу H_0; если $|T_{набл}|>t_{двуст.кр}$ – гипотезу H_0 отвергают.

Правило 2. При конкурирующей гипотезе $H_1: a>a_0$ по заданному уровню значимости α, помещенному в нижней строке таблицы критических точек распределения Стьюдента, и числу степеней свободы $k=n-1$ найти критическую точку $t_{правост.кр.}(\alpha;k)$ *правосторонней критической области*. Если $T_{набл}<t_{правост.кр}$ – нет оснований отвергнуть гипотезу H_0; если $T_{набл}>t_{правост.кр}$ – гипотезу H_0 отвергают.

Правило 3. При конкурирующей гипотезе $H_1: a<a_0$ сначала по правилу 2 находят "вспомогательную" критическую точку $t_{правост.кр.}(\alpha;k)$, а затем полагают границу *левосторонней критической области* $t_{левост.кр.}(\alpha;k)=-t_{правост.кр.}(\alpha;k)$. Если $T_{набл}>-t_{правост.кр}$ – нет оснований отвергнуть гипотезу H_0; если $T_{набл}<-t_{правост.кр}$ – гипотезу H_0 отвергают.

Для нахождения критической области необходимо знать критическое значение выборочной средней, которое можно найти из формулы для статистики (3.8):

$$\overline{x}_{кр}=a_0+\frac{t_{кр}}{\sqrt{n-1}}\cdot s \qquad (3.10)$$

Пример 2. По выборке объема $n=20$, извлеченной из нормальной генеральной совокупности, найдены выборочное среднее значение $\overline{x}_в=18$ и "исправленное" среднее квадратическое отклонение $s=4.5$. При уровне значимости 0.05 проверить гипотезу $H_0: a=a_0=17$ при конкурирующей гипотезе $H_1: a\neq 17$.

Решение. Вычислим наблюдаемое значение критерия:

$$T_{набл}=\frac{\overline{x}_в-a_0}{s}\cdot\sqrt{n-1}=\frac{18-17}{4.5}\cdot\sqrt{19}=0.99.$$

Поскольку конкурирующая гипотеза $H_1: a\neq 17$ – двусторонняя, то по таблице критических точек распределения Стьюдента по уровню значимости $\alpha=0.05$, помещенному в верхней строке таблицы, и по числу степеней свободы $k=20-1=19$, согласно правилу 1, находим критическую точку $t_{двуст.кр.}(\alpha;k)=t_{двуст.кр.}(0.05;19)=2.09$. Так как $|T_{набл}|=0.99<<t_{двуст.кр}=2.09$, то нет оснований отвергнуть гипотезу $H_0: a=a_0=17$. Следовательно, выборочное среднее *незначимо* отличается от гипотетической генеральной средней.

Пример 3. На основании сделанного прогноза средняя дебиторская задолженность однотипных предприятий региона должна составить $a_0=120$ ден.ед. Выборочная проверка 10 предприятий дала среднюю задолженность $\overline{x}_в=135$ ден.ед. и среднее квадратическое отклонение $s=20$ ден.ед. При уровне значимости 0.05 выяснить, можно ли принять данный прогноз. Найти критическую область для $\overline{x}$, если в действительности средняя дебиторская задолженность всех предприятий региона равна 130 ден.ед.

Решение. Проверяемая гипотеза $H_0: a=\overline{x}_в=120$ при конкурирующей гипотезе $H_1: a>120$. Так как генеральная дисперсия σ^2 неизвестна, то используем $t-$критерий Стьюдента. Вычислим наблюдаемое значение критерия:

$$T_{набл}=\frac{\overline{x}_в-a_0}{s}\cdot\sqrt{n-1}=\frac{135-120}{20}\cdot\sqrt{10-1}=2.25.$$

Поскольку конкурирующая гипотеза $H_1: a>120$ – правосторонняя, то по таблице критических точек распределения Стьюдента по уровню значимости $\alpha=0.05$, помещенному в нижней строке таблицы, и по числу степеней свободы $k=10-1=9$, согласно правилу 2, находим критическую точку $t_{правостор.кр.}(\alpha;k)=t_{правостор.кр.}(0.05;9)=1.83$. Так как $T_{набл}=2.25>1.83$, то гипотеза H_0 отвергается, т.е. на 5%-ом уровне значимости сделанный прогноз должен быть отвергнут.

Так как выдвинутая альтернативная гипотеза $H_1: a>120$, то критическая область – правосторонняя и критическое значение выборочной средней можно найти из формулы (3.10):

$$\bar{x}_{кр} = a_0 + \frac{t_{кр}}{\sqrt{n-1}} \cdot s = 120 + 1.83\frac{20}{\sqrt{10-1}} = 132.2 \text{ ден.ед.}$$

Таким образом, критическая область значений для $\bar{x}$ есть интервал $(132.2; +\infty)$.

Пример 4. На основании исследований одного залегания ученым-археологам стало известно, что диаметр раковин ископаемого моллюска равен 18.2 мм. В распоряжении ученых оказалась выборка из 50 раковин моллюсков из другого залегания, для которой было вычислено $\bar{x}_в = 18.9$ мм, $S = 2.18$ мм. Можно ли сделать предположение при $\alpha = 0.05$, что конкретное местообитание раковин не оказало влияние на диаметр их раковин?

Решение. Нулевая гипотеза $H_0 : a = a_0 = 18.2$ при конкурирующей гипотезе $H_1 : a \neq 18.2$. По правилу 1 имеем двустороннюю критическую область. Найдем и сравним $T_{набл}$ и $t_{двуст.кр.}$. Имеем:

$$T_{набл} = \frac{\bar{x}_в - a_0}{s} \cdot \sqrt{n-1} = \frac{18.9 - 18.2}{2.18} \cdot \sqrt{50-1} \approx 0.5$$

$$t_{двуст.кр.}(\alpha; k) = t_{двуст.кр.}(0.05; 49) = 2.02$$

Так как $T_{набл} = 0.5 < t_{двуст.кр.} = 2.02$, то нет оснований принимать гипотезу H_0. Т.е. с 95% уверенностью можно утверждать, что конкретное местообитание раковин оказало влияние на диаметр их раковин.

Исключение грубых ошибок наблюдений

Грубые ошибки могут возникнуть из-за ошибок показаний измерительных приборов, ошибок регистрации, случайного сдвига запятой в десятичной записи числа и т.д.

Пусть, например, $x^*, x_1, x_2, ..., x_n$ – совокупность имеющихся наблюдений, причем x^* резко выделяется. Необходимо решить вопрос о принадлежности резко выделяющегося значения к остальным наблюдениям.

Для ряда наблюдений $x_1, x_2, ..., x_n$ рассчитывают $\bar{x}_в$ и исправленное среднее квадратическое отклонение S. При справедливости гипотезы $H_0 : \bar{x} = x^*$ о принадлежности x^* к остальным наблюдениям статистика

$$T = \frac{\bar{x} - x^*}{S} \qquad (3.11)$$

имеет t – распределение Стьюдента с $\nu = n - 1$ степенями свободы. Конкурирующая гипотеза H_1 имеет вид: $\bar{x} > x^*$ или $\bar{x} < x^*$ – в зависимости от того, является ли резко выделяющееся значение больше или меньше остальных наблюдений. Гипотеза H_0 отвергается, если $|T_{набл.}| > t_{кр.}$, и принимается, если $|T_{набл.}| < t_{кр.}$

Пример 5. Имеются следующие данные об урожайности пшеницы на 8 опытных участках одинакового размера (ц/га):

26.5 26.2 35.9 30.1 32.3 29.3 26.1 25.0

Есть основание предполагать, что значение урожайности третьего участка $x^* = 35.9$ зарегистрировано неверно. Является ли это значение аномальным (резко выделяющимся) на 5%-ом уровне значимости?

Решение. Исключив значение $x^* = 35.9$, найдем для оставшихся наблюдений

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{26.5 + 26.2 + 30.1 + 32.3 + 29.3 + 26.1 + 25.0}{7} = 27.93 \text{ (ц/га)}$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{6}\Big((26.5 - 27.93)^2 + (26.2 - 27.93)^2 + (30.1 - 27.93)^2 +$$

$$+(32.3 - 27.93)^2 + (29.3 - 27.93)^2 + (26.1 - 27.93)^2 + (25.0 - 27.93)^2\Big) = 7.109$$

$$S = 2.67 \text{ (ц/га)}$$

Фактически наблюдаемое значение $T_{набл.} = \dfrac{\bar{x} - x^*}{S} = \dfrac{27.93 - 35.9}{2.67} \approx -2.98$.

Табличное значение $t_{кр.} = t_{1-2\alpha;\, n-1} = t_{0.9;\, 6} = 1.94$.

Сравнивая $|T_{набл.}|$ и $t_{кр.}$ ($|T_{набл.}| > t_{кр.}$), гипотезу $H_0 : \bar{x} = x^*$ (о принадлежности x^* к остальным наблюдениям) отвергаем. Следовательно, значение $x^* = 35.9$ является аномальным, и его следует отбросить.

Критерий для проверки гипотезы о вероятности события.

Пусть проведено n независимых испытаний (n – достаточно большое число), в каждом из которых некоторое событие A появляется с одной и той же, но неизвестной вероятностью p, и найдена относительная частота $\frac{m}{n}$ появлений A в этой серии испытаний. Проверим при заданном уровне значимости α нулевую гипотезу H_0, состоящую в том, что вероятность p равна некоторому значению p_0.

Примем в качестве статистического критерия случайную величину

$$U = \frac{\left(\frac{M}{n} - p_0\right)\sqrt{n}}{\sqrt{p_0 q_0}}, \qquad (3.12)$$

имеющую нормальное распределение с параметрами $M(U)=0$, $\sigma(U)=1$ (то есть нормированную). Здесь $q_0=1-p_0$. Вывод о нормальном распределении критерия следует из теоремы Лапласа (при достаточно большом n относительную частоту можно приближенно считать нормально распределенной с математическим ожиданием p и средним квадратическим отклонением $\sqrt{\frac{pq}{n}}$).

Критическая область строится в зависимости от вида конкурирующей гипотезы.

1) Если H_0: $p = p_0$, а H_1: $p \neq p_0$, то критическую область нужно построить так, чтобы вероятность попадания критерия в эту область равнялась заданному уровню значимости α. При этом наибольшая мощность критерия достигается тогда, когда критическая область состоит из двух интервалов, вероятность попадания в каждый из которых равна $\frac{\alpha}{2}$. Поэтому $u_{кр}$ определяется по таблице значений функции Лапласа из условия $\Phi(u_{кр}) = 1 - \alpha$, а критическая область имеет вид $(-\infty; -u_{кр}) \cup (u_{кр}; +\infty)$.

Далее нужно вычислить наблюдаемое значение критерия:

$$U_{набл} = \frac{\left(\frac{m}{n} - p_0\right)\sqrt{n}}{\sqrt{p_0 q_0}}. \qquad (3.13)$$

Если $|U_{набл}| < u_{кр}$, то нулевая гипотеза принимается.
Если $|U_{набл}| > u_{кр}$, то нулевая гипотеза отвергается.

2) Если конкурирующая гипотеза H_1: $p > p_0$, то критическая область определяется неравенством $U > u_{кр}$, то есть является правосторонней, причем $p(U > u_{кр}) = \alpha$. Тогда $p(0 < U < u_{кр}) = 1 - 2\alpha$. Следовательно, $u_{кр}$ можно найти по таблице значений функции Лапласа из условия, что $\Phi(u_{кр}) = 1 - 2\alpha$. Вычислим наблюдаемое значение критерия по формуле (3.13).
Если $U_{набл} < u_{кр}$, то нулевая гипотеза принимается.
Если $U_{набл} > u_{кр}$, то нулевая гипотеза отвергается.

3) Для конкурирующей гипотезы H_1: $p < p_0$ критическая область является левосторонней и задается неравенством $U < -u_{кр}$, где $u_{кр}$ вычисляется 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 6. Пусть проведено 50 независимых испытаний, и относительная частота появления события A оказалась равной 0,12. Проверить при уровне значимости α=0,01 нулевую гипотезу H_0: p=0,1 при конкурирующей гипотезе H_1: p>0,1.

Решение. Найдем $U_{набл} = \frac{(0{,}12 - 0{,}1)\sqrt{50}}{\sqrt{0{,}1 \cdot 0{,}9}} = 0{,}471$. Критическая область является правосторонней, а $u_{кр}$ находим из равенства $\Phi(u_{кр}) = 1 - 2 \cdot 0{,}01 = 0{,}98$. Из таблицы значений функции Лапласа определяем $u_{кр}$=2,33. Итак, $U_{набл} < u_{кр}$, и гипотеза о том, что p=0,1, принимается.

Непараметрические критерии

В разделе 1.6 закон распределения генеральной совокупности предполагался известным. Если же он неизвестен, но имеются основания предположить, что предполагаемый закон имеет определенный вид (например, A), то проверяют нулевую гипотезу: H_0 : генеральная совокупность распределена по закону A. Проверка гипотезы о предполагаемом законе

распределения так же, как и проверка гипотезы о неизвестных параметрах известного закона распределения, производится при помощи специально подобранной случайной величины – *критерия согласия.* Как бы хорошо ни был подобран теоретический закон распределения, между эмпирическим и теоретическим распределениями неизбежны расхождения. Поэтому возникает вопрос: объясняются ли эти расхождения только случайными обстоятельствами, связанными с ограниченным числом наблюдений, или они являются существенными и связаны с тем, что теоретический закон распределения подобран неудачно. Для ответа на этот вопрос и служат критерии согласия.

Критерием согласия называют критерий проверки гипотезы о предполагаемом законе неизвестного распределения.

Одним из основных критериев согласия является *критерий* χ^2 *(критерий Пирсона*).

Критерий Пирсона

Критерий Пирсона позволяет, в частности, проверить гипотезу о нормальном распределении генеральной совокупности. Для проверки этой гипотезы будем сравнивать *эмпирические* n_i (т.е. наблюдаемые) и *теоретические* n_i' (т.е. вычисленные в предположении нормального закона распределения) *частоты*, которые, как правило, различаются. Случайно (*незначимо*) или неслучайно (*значимо*) это расхождение? Ответ на этот вопрос и дает критерий согласия Пирсона.

Предположим, что генеральная совокупность X распределена нормально. Приведем ***алгоритм нахождения теоретических частот***.

1. Весь интервал наблюдаемых значений СВ X (выборки объема n) делят на k частичных интервалов (x_i, x_{i+1}) одинаковой длины, находят $x_i^* = (x_i + x_{i+1})/2$ – середины частичных интервалов. В качестве частоты n_i варианты x_i^* принимают число вариант, попавших в i-ый интервал. Получают последовательность равноотстоящих вариант и соответствующих им частот:

x_i	x_1^*	x_2^*	…	x_k^*	
n_i	n_1	n_2	…	n_k	$\sum_{i=1}^{k} n_i = n$

2. Вычисляют $\overline{x}_в^*$ и выборочное среднее квадратическое отклонение σ^*.

3. Нормируют СВ X, т.е. переходят к величине $Z = \frac{(X - \overline{x}_в^*)}{\sigma^*}$ и вычисляют концы интервалов (z_i, z_{i+1}): $z_i = (x_i - \overline{x}_в^*)/\sigma^*$, $z_{i+1} = (x_{i+1} - \overline{x}_в^*)/\sigma^*$, причем полагают наименьшее значение $z_1 = -\infty$, а наибольшее $z_k = +\infty$.

4. Вычисляют теоретические вероятности p_i попадания СВ X в интервалы (x_i, x_{i+1}) по формуле $p_i = \frac{1}{2}\left(\Phi(z_{i+1}) - \Phi(z_i)\right)$ ($\Phi(z)$–функция Лапласа). Находят *теоретические частоты* $n_i' = np_i$.

Пусть по выборке объема n нормально распределенной генеральной совокупности X получено эмпирическое распределение

x_i	x_1	x_2	…	x_k	
n_i	n_1	n_2	…	n_k	$\sum_{i=1}^{k} n_i = n$

и вычислены теоретические частоты n_i'.

Необходимо при уровне значимости α проверить справедливость нулевой гипотезы H_0: {генеральная совокупность распределена нормально}.

В качестве критерия проверки гипотезы H_0 примем СВ

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(n_i - n_i'\right)^2}{n_i'} \qquad (3.14)$$

Это – СВ, т.к. в различных опытах она принимает различные, не известные заранее значения. Ясно, что $\chi^2 \to 0$ при $n_i \to n_i'$, т.е. чем меньше различаются эмпирические n_i и теоретические n_i' частоты, тем меньше значение критерия χ^2. Таким образом, критерий (3.14) характеризует близость эмпирического и теоретического распределения.

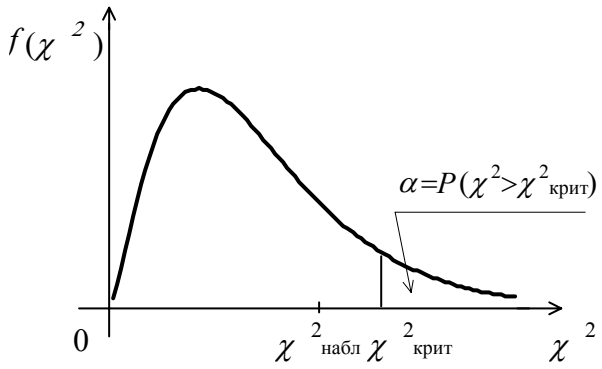
Известно, что при $n \to \infty$ закон распределения СВ (3.14) стремится к закону распределения χ^2 с ν степенями свободы. Поэтому СВ в (3.14) обозначается через χ^2, а сам критерий называют *критерием согласия* χ^2. Число степеней свободы ν находят по равенству $\nu = k - r - 1$, где k – число групп (частичных интервалов), r – число параметров предполагаемого распределения, которые оценены по данным выборки (для нормального закона распределения $r = 2$, поэтому $\nu = k - 3$).

Построим *правостороннюю критическую область* (т.к. односторонний критерий более "жестко" отвергает гипотезу H_0), исходя из требования, чтобы, в предположении справедливости гипотезы H_0, вероятность попадания критерия в эту область была равна принятому уровню значимости α: $P\left[\chi^2 > \chi^2_{кр}(\alpha;\nu)\right] = \alpha$. Следовательно, правосторонняя критическая область определяется неравенством $\chi^2 > \chi^2_{кр}(\alpha;\nu)$, а область принятия гипотезы H_0 – неравенством $\chi^2 < \chi^2_{кр}(\alpha;\nu)$.

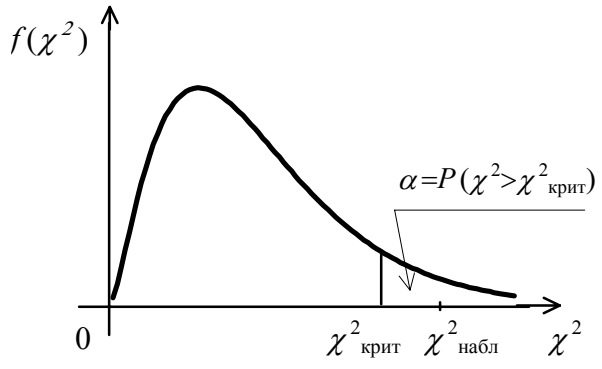
Значение критерия (3.14), вычисленное по данным наблюдений, обозначим $\chi^2_{набл}$. Сформулируем

Правило проверки нулевой гипотезы H_0**.** *Для того чтобы при заданном уровне значимости* α *проверить нулевую гипотезу* H_0 *: {генеральная совокупность распределена нормально}, необходимо вычислить теоретические частоты* n_i' *и наблюдаемое значение критерия согласия* χ^2 *Пирсона* $\chi^2_{набл} = \sum_{i=1}^{k} \frac{(n_i - n_i')^2}{n_i'}$. *По таблице критических точек распределения* χ^2 *по заданному уровню значимости* α *и числу степеней свободы* $\nu = k - 3$ *найти критическую точку* $\chi^2_{кр}(\alpha;\nu)$.

- *Если наблюдаемое значение критерия* $\chi^2_{\text{набл}}$ *попало в область принятия гипотезы* $\chi^2_{набл} < \chi^2_{кр}(\alpha;\nu)$, *то нет оснований отвергнуть нулевую гипотезу* H_0 *(рис. 5а)).*
- *Если наблюдаемое значение критерия* $\chi^2_{\text{набл}}$ *попало в критическую область* $\chi^2_{набл} > \chi^2_{кр}(\alpha;\nu)$, *то нулевую гипотезу* H_0 *отвергают (рис. 5б)).*



а).



б)

Рис. 5

Покажем, что для контроля вычислений наблюдаемого критерия $\chi^2_{набл}$ можно использовать равенство

$$\sum_{i=1}^{k} \frac{n_i^2}{n_i'} - n = \chi^2_{набл}. \qquad (3.15)$$

Действительно, из (3.14) вытекает:

$$\chi^2 = \sum_{i=1}^{l} \frac{\left(n_i - n_i'\right)^2}{n_i'} = \sum_{i=1}^{l} \left(\frac{n_i^2}{n_i'} - \frac{2n_i n_i'}{n_i'} + \frac{n_i'^2}{n_i'} \right) = \sum_{i=1}^{l} \left(\frac{n_i^2}{n_i'} - 2n_i + n_i' \right) =$$

$$= \sum_{i=1}^{l} \frac{n_i^2}{n_i'} - 2\sum_{i=1}^{l} n_i + \sum_{i=1}^{l} np_i = \sum_{i=1}^{l} \frac{n_i^2}{n_i'} - 2n + n\sum_{i=1}^{l} p_i = \sum_{i=1}^{l} \frac{n_i^2}{n_i'} - 2n + n = \sum_{i=1}^{l} \frac{n_i^2}{n_i'} - n$$

Пример 7. Используя критерий Пирсона при уровне значимости 0,05, установить, случайно или значимо расхождение между эмпирическими и теоретическими частотами, которые вычислены, исходя из предположения о нормальном распределении признака X генеральной совокупности:

n_i	14	18	32	70	20	36	10
n_i'	10	24	34	80	18	22	12.

Решение. Выдвигаем нулевую гипотезу H_0 и ей конкурирующую H_1.

H_0: признак X имеет нормальный закон распределения.

H_1: признак X имеет закон распределения, отличный от нормального.

В данном случае рассматривается правосторонняя критическая область. Проверим гипотезу с помощью случайной величины $\chi^2_{набл} = \sum_{i=1}^{k} \frac{\left(n_i - n_i'\right)^2}{n_i'}$, которая имеет распределение χ^2 с $\nu = k - 3 = 7 - 3 = 4$ степенями свободы. Вычислим наблюдаемое значение критерия χ^2 по выборочным данным. Расчеты представим в таблице:

n_i	n_i'	$\frac{\left(n_i - n_i'\right)^2}{n_i'}$
14	10	1,6
18	24	1,5
32	34	0,118
70	80	1,25
20	18	0,222
36	22	8,909
10	12	0,333
Итого 200	200	13,932 .

$\chi^2_{\text{набл}} \approx 13{,}93$; $\chi^2_{\text{крит}}(0{,}05;\ 4) = 9{,}5$ (по таблице). Сравниваем $\chi^2_{\text{набл}}$ и $\chi^2_{\text{крит}}(0{,}05;\ 4)$. Так как $\chi^2_{\text{набл}} > \chi^2_{\text{крит}}(0{,}05;\ 4)$, то есть наблюдаемое значение критерия попало в критическую область (см. рис. 5б)), нулевая гипотеза отвергается, справедлива конкурирующая гипотеза, то есть признак X имеет закон распределения, отличный от нормального, расхождение между эмпирическими и теоретическими частотами значимо.

Тема 4

Элементы теории корреляции. Линейная регрессия

Одной из основных задач МС является нахождение зависимости между двумя или несколькими СВ. В естественных науках часто речь идет о ***функциональной зависимости*** между величинами X и Y, когда *каждому значению одной переменной соответствует вполне определенное значение другой* (например, скорость свободного падения тела в вакууме зависит от времени падения). Однако строгая функциональная зависимость реализуется редко, т.к. обе СВ или одна из них подвержены действию случайных факторов. В этом случае возникает ***статистическая зависимость***.

Статистической (*вероятностной, стохастической*) называют зависимость, при которой *изменение одной из величин влечет изменение распределения другой*. Возникновение понятия статистической зависимости обусловлено тем, что зависимая переменная подвержена влиянию ряда неконтролируемых или неучтенных факторов, а также тем, что измерение значений переменных, как правило, сопровождается некоторыми случайными ошибками. Примером статистической зависимости может служить зависимость всхожести семян некоторых культур от количества микроэлементов при их обработке, зависимость производительности труда на предприятии от его энерговооруженности и.т.д. Таким образом, статистическая зависимость между двумя СВ Y и X неоднозначна. Статистическая зависимость, в частности, проявляется в том, что при изменении одной из величин изменяется *среднее значение* другой. Для исследователя представляет интерес *усредненная* по x схема зависимости, т.е. закономерность в изменении *условного математического ожидания* $M_x(Y)$ (или в других обозначениях – $M(Y|X=x)$, т.е. математического ожидания СВ Y, вычисленной в предположении, что СВ X приняла значение x) в зависимости от x. Такую статистическую зависимость называют ***корреляционной***.

Корреляционной (или ***регрессионной***) ***зависимостью*** между двумя переменными величинами называется функциональная зависимость между значениями одной из них и условным математическим ожиданием другой.

Корреляционная зависимость, например, имеется:

- между ростом и весом человека – с увеличением роста *средний вес* также возрастает;
- между надежностью автомобиля и его возрастом – чем больше возраст, тем *в среднем* меньше его надежность.

Рассмотрим пример СВ Y, которая связана с другой СВ X не функционально, а корреляционно.

Пример 1. Пусть Y – успеваемость студентов, X – посещаемость ими учебных занятий. У одинаковых студенческих групп (по количеству студентов и количеству часов лекционных и практических занятий) по результатам экзаменационной сессии успеваемость разная, т.е. Y не является функцией от X – посещаемости учебных занятий: $Y \neq f(X)$. Однако, как показывает опыт,

результаты экзаменационной сессии лучше у тех студентов, которые систематически посещали учебные занятия. Это означает, что Y связано с X корреляционной зависимостью (связью).

Для уточнения определения корреляционной зависимости введем понятие *условной средней.*

Условной средней $\bar{Y}_x = M(Y|X=x)$ называется среднее значение СВ Y при $X=x$.

В качестве *оценок* условных математических ожиданий принимают *условные средние*, которые находят по данным наблюдений, т.е. по выборке. Так например, если при $x_1=2$ СВ Y приняла значения $y_1=3,\ y_2=7, y_3=5,$ то *условное среднее* $\bar{y}_{x_1} = \dfrac{y_1+y_2+y_3}{3} = \dfrac{3+7+5}{3} = 5$. Условное математическое ожидание $M_x(Y)$ СВ Y есть функция от x: $M_x(Y)=f(x)$, которую называют *функцией регрессии Y на X*. Поскольку каждому значению x соответствует одно значение условного среднего, т.е. $\bar{Y}_x = f(x)$ является функцией от x, то можно сказать, что СВ Y зависит от СВ X *корреляционно***.**

Корреляционной зависимостью Y от X называется функциональная зависимость условной средней $\bar{Y}_x$ от x.

Уравнение $y=f(x)$ называется *уравнением регрессии Y на X*. Функция $f(x)$ называется *регрессией Y на X,* а ее график – *линией регрессии СВ Y на* СВ X.

Аналогично для СВ X определяются *условное математическое ожидание* $M_y(X)$ (или в других обозначениях – $M(X|Y=y)$, *условное среднее* $\bar{X}_y = M(X|Y=y)$, *корреляционная зависимость* СВ X от СВ Y, *функция регрессии* СВ X на СВ Y: $\varphi(y)=\bar{X}_y$.

Основными задачами теории корреляции являются:

1. Установление формы корреляционной связи, т.е. вида функции регрессии (линейная, квадратичная, показательная и т.д.).

2. Оценка тесноты корреляционной связи Y от X, которая оценивается величиной рассеяния значений Y около $\bar{Y}_x$. Большое рассеяние означает слабую зависимость Y от X либо вообще отсутствие таковой. Малое рассеяние указывает на существование достаточно сильной зависимости Y от X.

Важной с точки зрения приложений является ситуация, когда обе функции регрессии $f(x), \varphi(y)$ являются линейными. Тогда говорят, что СВ X и Y связаны линейной корреляционной зависимостью (*линейной корреляцией*). Такая ситуация возникает, например, если система СВ (X,Y) имеет совместное нормальное распределение. Тогда модельные уравнения регрессии являются линейными, а их графики – прямыми.

Рассмотрим методы нахождения линейной регрессии, представляющей наибольший интерес. Пусть даны результаты n измерений двух СВ X и Y:

(x_1, y_1), (x_2, y_2), ..., (x_n, y_n). Предварительное представление о характере зависимости между X и Y можно получить, если элементы выборки $(x_i, y_i), i = \overline{1,n}$, изобразить графически точками координатной плоскости в выбранной системе координат. В результате получим *точечную диаграмму* статистической зависимости, которая называется ***корреляционным полем***. По его виду можно составить предварительное мнение о степени и типе зависимости двух СВ.

Как известно, для описания системы двух СВ (X, Y) вводят математические ожидания $M(X)$, $M(Y)$ и дисперсии $D(X) = \sigma_X^2, D(Y) = \sigma_Y^2$ составляющих, а также *корреляционный момент* (*ковариацию*) $K(X,Y) = M\{[X - M(X)][Y - M(Y)]\}$ и *коэффициент корреляции* $r = \dfrac{K(X,Y)}{\sigma(X)\sigma(Y)}$

Корреляционный момент $K(X,Y)$ служит для характеристики связи между СВ X и Y: если X и Y независимы, то $K(X,Y) = 0$, а следовательно, и $r = 0$.

Две случайные величины X и Y называются *коррелированными*, если их корреляционный момент (или, что то же самое, коэффициент корреляции) отличен от нуля. СВ X и Y называются *некоррелированными*, если их корреляционный момент равен нулю.

Рассмотрим вопрос о силе связи между признаками X и Y. Для этой цели введем *выборочный коэффициент корреляции* $r_в$. На основе определения *теоретического коэффициента корреляции* $r = \dfrac{K(X,Y)}{\sigma(X)\sigma(Y)}$ и оценок параметров теоретического распределения через выборочные, *выборочный коэффициент корреляции* $r_в$ может быть представлен в виде

$$r_в = \frac{K_в(X,Y)}{\sigma_в(X)\sigma_в(Y)}, \qquad (4.1)$$

где $K_в(X,Y)$ – выборочный корреляционный момент,

$$K_в(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_в)(y_i - \bar{y}_в) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}_в \cdot \bar{y}_в = \overline{xy}_в - \bar{x}_в \cdot \bar{y}_в,$$

$\sigma_в(X) = \sqrt{D_в(X)} = \sqrt{\overline{x_в^2} - (\bar{x}_в)^2}$, $\sigma_в(Y) = \sqrt{D_в(Y)} = \sqrt{\overline{y_в^2} - (\bar{y}_в)^2}$ – выборочные среднеквадратические отклонения признаков X и Y. Следовательно, из (4.1) имеем

$$r_в = \frac{\overline{xy}_в - \bar{x}_в \cdot \bar{y}_в}{\sqrt{\overline{x_в^2} - (\bar{x}_в)^2}\sqrt{\overline{y_в^2} - (\bar{y}_в)^2}} \qquad (4.2)$$

Формула (4.2) симметрична относительно двух переменных, т.е. x и y можно менять местами. Выборочный коэффициент корреляции $r_в$ обладает теми же свойствами, что и теоретический коэффициент корреляции $r_Г$, и является мерой линейной зависимости между двумя наблюдаемыми величинами.

Отметим ***основные свойства выборочного коэффициента корреляции*** (при достаточно большом объеме выборки):

1. Коэффициент корреляции $r_в$ принимает значения на отрезке $[-1;1]$, т.е. $-1 \le r_в \le 1$.

2. Если все значения переменных увеличить (уменьшить) на одно и то же число или в одно и то же число раз, то величина выборочного коэффициента корреляции не изменится.

3. При $r_в = \pm 1$ корреляционная связь представляет *линейную функциональную зависимость*. При этом линии регрессии Y на X и X на Y совпадают, все наблюдаемые значения располагаются на общей прямой.

4. Если с ростом значений одной СВ значения второй возрастают, то $r_в > 0$, если убывают, то $r_в < 0$.

5. При $r_в = 0$ *линейная* корреляционная связь отсутствует, групповые средние переменных совпадают с их общими средними, а линии регрессии Y на X и X на Y параллельны осям координат.

Замечание. Равенство $r_в = 0$ говорит лишь об отсутствии *линейной* корреляционной зависимости (т.е. о некоррелированности СВ X и Y), но не вообще об отсутствии корреляционной, а тем более статистической зависимости.

Выборочный коэффициент корреляции $r_в$ является *оценкой генерального коэффициента корреляции* $r_Г = \dfrac{K(X,Y)}{\sigma(X)\sigma(Y)}$, характеризующего тесноту связи между СВ X и Y генеральной совокупности. На практике о тесноте корреляционной зависимости между рассматриваемыми СВ судят не по величине $r_Г$, который, как правило, неизвестен, а по величине его выборочного аналога $r_в$. Так как $r_в$ вычисляется по значениям переменных, случайно попавших в выборку из генеральной совокупности, то в отличие от параметра $r_Г$ параметр $r_в$ – *величина случайная*.

Проверка значимости выборочного коэффициента корреляции в случае выборки из нормального двумерного распределения

Пусть двумерная генеральная совокупность (X,Y) распределена нормально. Из этой совокупности извлечена выборка объема n и по ней найден выборочный коэффициент корреляции $r_в$, причем оказалось, что $r_в \neq 0$. Поскольку выборка произведена случайно, нельзя утверждать, что $r_Г \neq 0$. Требуется проверить, значимо или незначимо отличие выборочного коэффициента корреляции $r_в$ от нуля либо это вызвано только случайными изменениями выборочных значений.

Задача. При заданном уровне значимости α проверить справедливость гипотезы $H_0: r_Г = 0$ о равенстве нулю генерального коэффициента корреляции при конкурирующей гипотезе $H_1: r_Г \neq 0$.

Для проверки гипотезы H_0 задается уровень значимости α – допустимая вероятность ошибки (α=0.05; 0.01; 0.1). Вычисляют статистику $T_{набл} = \dfrac{|r_в|}{\sqrt{1-r_в^2}} \cdot \sqrt{n-2},$ где n – объем выборки. По таблице распределения Стьюдента по заданному уровню значимости α и числу степеней свободы $k = n-2$ находят $t_{кр} = t_{\alpha;n-2}$. Если $T_{набл} < t_{кр}$ – нет оснований отвергнуть гипотезу H_0. Если $T_{набл} > t_{кр}$, то гипотезу H_0 о равенстве коэффициента корреляции нулю отвергают. Другими словами, $r_в$ значимо отличается от нуля, т.е. СВ X и Y коррелированны. В этом случае считают, что зависимость между наблюдаемыми величинами можно приблизить линейной зависимостью.

Пример 2. По выборке объема $n = 102$, извлеченной из нормальной двумерной совокупности, найден выборочный коэффициент корреляции $r_в = 0.3$. При уровне значимости 0.05 проверить нулевую гипотезу о равенстве нулю генерального коэффициента $r_Г$ при конкурирующей гипотезе $H_1: r_Г \neq 0$.

Решение. Найдем наблюдаемое значение критерия $T_{набл} = \dfrac{|r_в|}{\sqrt{1-r_в^2}} \cdot \sqrt{n-2} = \dfrac{|0.3|}{\sqrt{1-0.3^2}} \cdot \sqrt{102-2} = \dfrac{30}{0.954} = 28.62.$ По условию, конкурирующая гипотеза имеет вид $r_Г \neq 0$, следовательно, критическая область – двусторонняя. По уровню значимости $\alpha = 0.05$ и числу степеней свободы $k = n-2 = 100$ по таблице критических точек распределения Стьюдента для двусторонней критической области находим критическую точку $t_{кр} = t_{\alpha;n-2} = t_{кр}(0.05;100) = 1.98$. Так как $T_{набл} = 28.62 > t_{кр} = 1.98,$ то нулевую гипотезу H_0 отвергаем, т.е. выборочный коэффициент корреляции значимо отличается от нуля. Таким образом, СВ X и Y коррелированны.

Определение коэффициентов уравнения линейной регрессии

Рассмотрим систему двух СВ (X,Y). Если обе функции регрессии $f(x), \varphi(y)$ Y на X и X на Y линейны, то говорят, что СВ X и Y связаны *линейной корреляционной зависимостью.* Тогда графиками линейных функций регрессии являются прямые линии. Пусть, например, приближенное представление СВ Y представлено в виде линейной функции СВ X: $Y \approx f(x) = a + bX$, где a, b – параметры, подлежащие определению. Их можно определить различными способами. Наиболее употребительным является метод наименьших квадратов (МНК). Функцию $f(x)$ называют "*наилучшим приближением*" Y в смысле МНК, если математическое ожидание

$M[Y-f(X)]^2$ принимает наименьшее возможное значение. Поэтому $f(x)$ называют *среднеквадратической регрессией* Y на X.

Теорема 1. *Линейная средняя квадратическая регрессия* Y *на* X *имеет вид*

$$f(x)=m_y+r\frac{\sigma_y}{\sigma_x}(x-m_x), \qquad (4.3)$$

где $m_x=M(X), m_y=M(Y), \sigma_x=\sqrt{D(X)}, \sigma_y=\sqrt{D(Y)}, r=\dfrac{K(X,Y)}{\sigma_x\sigma_y}$ –

коэффициент корреляции СВ X *и* Y.

Из теоремы 1 следует, что в силу (4.3) параметры a, b имеют вид: $a=m_y-r\frac{\sigma_y}{\sigma_x}m_x$, $b=r\frac{\sigma_y}{\sigma_x}$. Коэффициент $b=r\frac{\sigma_y}{\sigma_x}$ называется *коэффициентом регрессии* Y *на* X, а прямая $y-m_y=r\frac{\sigma_y}{\sigma_x}(x-m_x)$ – *прямой среднеквадратической регрессии* Y *на* X. Аналогично можно получить прямую среднеквадратической регрессии X на Y: $x-m_x=r\frac{\sigma_x}{\sigma_y}(y-m_y)$, где $r\frac{\sigma_x}{\sigma_y}$ *– коэффициент регрессии* X *на* Y. Из сопоставления уравнений линейной регрессии Y на X и X на Y видно, что при $r=\pm1$ они совпадают. Кроме того очевидно, что обе прямые $y-m_y=r\frac{\sigma_y}{\sigma_x}(x-m_x)$, $x-m_x=r\frac{\sigma_x}{\sigma_y}(y-m_y)$ проходят через точку $(m_x;m_y)$, называемую *центром совместного распределения СВ* X *и* Y.

Выше были введены уравнения регрессии Y на X: $M(Y|X=x)=f(x)$ и X на Y: $M(X|Y=y)=\varphi(y)$. Поскольку условное математическое ожидание $M(Y|X=x)$ является функцией от x, то и его оценка $M(M(Y|X=x))$, т.е. условное среднее $\bar{y}_x$, также является функцией от x: $\bar{y}_x=f^*(x)$; где $f^*(x)$ в силу (4.3) имеет вид $f^*(x)=\bar{y}_в+r_в\frac{s_y}{s_x}(x-\bar{x}_в)$. Полученное уравнение называют *выборочным уравнением регрессии* Y *на* X; функцию $f^*(x)$ *– выборочной регрессией* Y *на* X, а ее график *– выборочной линией регрессии* Y *на* X. Аналогично, уравнение $\bar{x}_y=\varphi^*(y)$, где $\varphi^*(y)=\bar{x}_в+r_в\frac{s_x}{s_y}(y-\bar{y}_в)$, называют *выборочным уравнением регрессии* X *на* Y; функцию $\varphi^*(y)$ *– выборочной регрессией* X *на* Y, а ее график *– выборочной линией регрессии* X *на* Y.

Как по данным наблюдений $(x_1;y_1),(x_2;y_2),\ldots,(x_n;y_n)$, полученным в результате n независимых опытов, найти параметры функций $f^*(x)$, $\varphi^*(y)$,

если их вид известен? Как оценить тесноту связи между СВ X и Y и установить, коррелированны ли эти величины?

Пусть принята гипотеза о линейной зависимости между величинами X и Y. По данным наблюдений найдем, например, выборочное уравнение прямой линии среднеквадратической регрессии Y на X: $f^*(x) = Y = a + bx$. Будем полагать, что все результаты измерений $(x_i; y_i), i = \overline{1,n}$ различны. Подберем параметры a,b так, чтобы точки $(x_1; y_1),(x_2; y_2),\ldots,(x_n; y_n)$, построенные на плоскости xOy по данным наблюдений, лежали как можно ближе к прямой $Y = a + bx$ в смысле МНК. Сформулированное требование означает, что параметры a,b будем выбирать из условия, чтобы сумма квадратов отклонений $Y_i - y_i, i = \overline{1,n}$, была минимальной. Здесь Y_i – ордината, вычисленная по эмпирическому (выборочному) уравнению $Y_i = a + bx_i$, соответствующая наблюдаемому значению x_i, y_i – наблюдаемая ордината, соответствующая x_i. Следовательно, рассмотрим функцию

$$F(a,b) = \sum_{i=1}^{n} (Y_i - y_i)^2 = \sum_{i=1}^{n} ((a + bx_i) - y_i)^2 \to \min_{a,b} \qquad (4.4)$$

Необходимое условие экстремума сводится к условиям

$$\begin{cases} \dfrac{\partial F(a,b)}{\partial a} = 0, \\ \dfrac{\partial F(a,b)}{\partial b} = 0. \end{cases}$$

Находя соответствующие частные производные и приравнивая их нулю, получаем:

$$\begin{cases} n \cdot a + b\sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} y_i, \\ a\sum\limits_{i=1}^{n} x_i + b\sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} x_i y_i. \end{cases}, \qquad (4.5)$$

или, после деления обеих частей уравнений системы на n:

$$\begin{cases} a + b\overline{x}_в = \overline{y}_в, \\ a\overline{x}_в + b\overline{x_в^2} = \overline{xy}_в \end{cases},$$

где $\overline{x}_в = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i$, $\overline{y}_в = \dfrac{1}{n}\sum\limits_{i=1}^{n} y_i$, $\overline{x_в^2} = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i^2$, $\overline{xy}_в = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i y_i$.

Вычисляя определитель Δ данной системы

$$\Delta = \begin{vmatrix} 1 & \overline{x}_в \\ \overline{x}_в & \overline{x_в^2} \end{vmatrix} = \overline{x_в^2} - (\overline{x}_в)^2 = D_в,$$

находим неизвестные параметры a,b:

$$a=\frac{\Delta a}{\Delta}=\frac{\begin{vmatrix}\overline{y}_в & \overline{x}_в\\ \overline{xy}_в & \overline{x^2_в}\end{vmatrix}}{\overline{x^2_в}-\left(\overline{x}_в\right)^2}=\frac{\overline{y}_в\cdot\overline{x^2_в}-\overline{x}_в\cdot\overline{xy}_в}{\overline{x^2_в}-\left(\overline{x}_в\right)^2}=\frac{(\overline{y}_в\cdot\overline{x^2_в}-\overline{y}_в\cdot\left(\overline{x}_в\right)^2)+\overline{y}_в\cdot\left(\overline{x}_в\right)^2-\overline{x}_в\cdot\overline{xy}_в}{\overline{x^2_в}-\left(\overline{x}_в\right)^2}=$$

$$=\overline{y}_в-\frac{\overline{x}_в\cdot\overline{xy}_в-\overline{y}_в\cdot\left(\overline{x}_в\right)^2}{\overline{x^2_в}-\left(\overline{x}_в\right)^2}=\overline{y}_в-\frac{\overline{x}_в\cdot(\overline{xy}_в-\overline{y}_в\cdot\overline{x}_в)}{\overline{x^2_в}-\left(\overline{x}_в\right)^2}\overset{(4.1)}{=}$$

$$=\overline{y}_в-\frac{\overline{x}_в\cdot K_в(X,Y)}{\sigma^2_в(X)}=\overline{y}_в-\frac{\overline{x}_в\cdot r_в(X,Y)\cdot\sigma_в(X)\cdot\sigma_в(Y)}{\sigma^2_в(X)}=\overline{y}_в-r_в(X,Y)\frac{\overline{x}_в\cdot\sigma_в(Y)}{\sigma_в(X)},$$

$$b=\frac{\Delta b}{\Delta}=\frac{\begin{vmatrix}1 & \overline{y}_в\\ \overline{x}_в & \overline{xy}_в\end{vmatrix}}{\overline{x^2_в}-\left(\overline{x}_в\right)^2}=\frac{\overline{xy}_в-\overline{y}_в\cdot\overline{x}_в}{\overline{x^2_в}-\left(\overline{x}_в\right)^2}=\frac{K_в(X,Y)}{\sigma^2_в(X)}\overset{(4.1)}{=}$$

$$=\frac{r_в(X,Y)\cdot\sigma_в(X)\cdot\sigma_в(Y)}{\sigma^2_в(X)}=r_в(X,Y)\frac{\sigma_в(Y)}{\sigma_в(X)}.$$

Таким образом, выборочное уравнение линейной регрессии $y=a+bx$ Y на X имеет вид:

$$y=\overline{y}_в-r_в(X,Y)\frac{\overline{x}_в\cdot\sigma_в(Y)}{\sigma_в(X)}+r_в(X,Y)\frac{\sigma_в(Y)}{\sigma_в(X)}x,$$

или в иной записи – $y-\overline{y}_в=r_в(X,Y)\frac{\sigma_в(Y)}{\sigma_в(X)}(x-\overline{x}_в)$.

Аналогичным образом можно получить уравнение линейной регрессии X на Y: $x-\overline{x}_в=r_в(X,Y)\frac{\sigma_в(X)}{\sigma_в(Y)}\left(y-\overline{y}_в\right)$. На практике совместное распределение СВ (X,Y) зачастую неизвестно, а известны только результаты наблюдений, т.е. выборка пар (x_i,y_i), $i=\overline{1,n}$ значений СВ (X,Y). Тогда все рассмотренные величины $m_x,m_y,\sigma_x,\sigma_y,r$ заменяем их выборочными аналогами: в полученных уравнениях $\sigma_в(X)$, $\sigma_в(Y)$ – их несмещенными оценками

$$S^2_x=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x}_в)^2=\frac{1}{n-1}\sum_{i=1}^{n}x_i^2-\frac{n}{n-1}\overline{x}_в^2,$$

$$S^2_y=\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y}_в)^2=\frac{1}{n-1}\sum_{i=1}^{n}y_i^2-\frac{n}{n-1}\overline{y}_в^2,$$

а x,y в левых частях – на соответствующие условные средние $\overline{y}_x,\overline{x}_y$, получим эмпирические функции линейной регрессии в виде

$$\overline{y}_x-\overline{y}_в=r_в\frac{S_y}{S_x}(x-\overline{x}_в),\qquad(4.6)$$

$$\overline{x}_y-\overline{x}_в=r_в\frac{S_x}{S_y}\left(y-\overline{y}_в\right).\qquad(4.7)$$

Заметим, что если нанести обе линии регрессии на корреляционное поле, то прямые должны пересечься в точке $\left(\overline{x}_в, \overline{y}_в\right)$.

Уравнения линейной регрессии получены в предположении, что все измерения встречаются по одному разу. При большом числе наблюдений одно и то же значение СВ X может повторяться n_x раз, а СВ Y – n_y раз. Одинаковая пара чисел (x, y) может наблюдаться n_{xy} раз. Поэтому результаты наблюдений группируют, подсчитывая *частоты* n_x, n_y, n_{xy}. Все данные записывают в *корреляционную таблицу*. Построение корреляционной таблицы следующее: в клетки верхней строки записывают наблюдаемые значения $x_i, i=\overline{1,k}$, а в первый столбец – наблюдаемые значения $y_j, j=\overline{1,m}$. На пересечении строк и столбцов записывают *кратности* $n_{x_i y_j}, i=\overline{1,k}$, $j=\overline{1,m}$ наблюдаемых пар значений признаков. В правом нижнем углу расположена сумма всех частот n_{x_i}, $i=\overline{1,k}$ и n_{y_j}, $j=\overline{1,m}$, равная общему числу всех наблюдений n (объему выборки).

x_i / y_i	x_1	x_2	…	x_k	n_{y_j}
y_1	n_{11}	n_{12}	…	n_{1k}	$\sum_{i=1}^{k} n_{1i}$
y_2	n_{21}	n_{22}	…	n_{2k}	$\sum_{i=1}^{k} n_{2i}$
…	…	…	…	…	…
y_m	n_{m1}	n_{m2}	…	n_{mk}	$\sum_{i=1}^{k} n_{mi}$
n_{x_i}	$\sum_{j=1}^{m} n_{j1}$	$\sum_{j=1}^{m} n_{j2}$	…	$\sum_{j=1}^{m} n_{j1}$	$n=\sum_{i=1}^{k} n_{x_i}=\sum_{j=1}^{m} n_{y_j}$

Для непрерывных СВ корреляционная таблица имеет вид

x_i / y_i	$[x_1, x_2)$	$[x_2, x_3)$	…	$[x_{k-1}, x_k]$	n_{y_j}
$[y_1, y_2)$	n_{11}	n_{12}	…	n_{1k}	$\sum_{i=1}^{k} n_{1i}$
$[y_2, y_3)$	n_{21}	n_{22}	…	n_{2k}	$\sum_{i=1}^{k} n_{2i}$
…	…	…	…	…	…
$[y_{m-1}, y_m]$	n_{m1}	n_{m2}	…	n_{mk}	$\sum_{i=1}^{k} n_{mi}$
n_{x_i}	$\sum_{j=1}^{m} n_{j1}$	$\sum_{j=1}^{m} n_{j2}$	…	$\sum_{j=1}^{m} n_{j1}$	$n = \sum_{i=1}^{k} n_{x_i} = \sum_{j=1}^{m} n_{y_j}$

где n_{ij} – частоты (кратности) наблюдаемых пар значений признаков, попавших в соответствующие интервалы $[x_i, x_{i+1})$, $[y_j, y_{j+1})$, $i = \overline{1, k-1}$, $j = \overline{1, m-1}$. В этом случае таблица сводится к предыдущей путем перехода к серединам интервалов группировки статистических данных.

Если на основании наблюдаемых значений (x_i, y_i), $i = \overline{1, n}$ СВ (X, Y) можно предположить, что зависимость y_i от x_i квадратичная ($y = ax^2 + \ + bx + c$, то применение МНК

$$S(a, b, c) = \sum_{i=1}^{n} \left(ax_i^2 + bx_i + c - y_i \right)^2 \to \min$$

дает возможность найти неизвестные параметры a, b, c. Отметим, что и в этом случае схема для нахождения параметров a, b, c является линейной.

Если же рассматривается нелинейная зависимость наблюдаемых значений y_i от x_i, то обычно используют методы линеаризации, т.е. переходят к условным переменным, где зависимость от параметров становится линейной, а затем применяют МНК. Пусть, например, на основании наблюдаемых значений (x_i, y_i), $i = \overline{1, n}$ СВ (X, Y) выдвинута гипотеза H_0: зависимость y_i от x_i имеет вид $a, b \quad y = ae^{bx}$. Прологарифмировав данное нелинейное уравнение, получим $\ln y = \ln a + bx \ln e$. Введя обозначения $Y = \ln a, \quad A = \ln a, B = b,$ получим линейную зависимость $Y = A + Bx,$ для которой можно применить описанный выше МНК нахождения неизвестных параметров A, B. Из введенной замены переменных находим $a = e^A, b = B$, а следовательно, и предполагаемую зависимость $y = e^{A+Bx}$.

Помимо зависимости (корреляционной) между двумя СВ можно рассматривать корреляционную зависимость одной СВ от двух и более СВ. В таких случаях говорят о *множественной регрессии*. Например, множественная регрессия от двух переменных: $z = ax + by + c$.

В этом случае параметры уравнения находятся по МНК и рассматривают корреляционную связь между каждыми признаками и отдельно тесноту связи между признаком z и общими признаками x и y. Для этого вычисляется совместный выборочный коэффициент корреляции, который выражается через выборочные коэффициенты корреляции компонент. При нахождении выборочного уравнения регрессии необходимо проверять статистическую гипотезу о значимости коэффициента корреляции, т.е. о том, как связано выборочное уравнение регрессии с регрессией, изучаемой генеральной совокупностью.

Уравнение регрессии можно использовать для *прогнозирования* (*предсказания*).

Пример 3**.** Изучается зависимость себестоимости одного изделия (Y, р.) от величины выпуска продукции (X, тыс. шт.) по группе предприятий за отчетный период. Получены следующие данные:

X	2	3	4	5	6
Y	1,9	1,7	1,8	1,6	1,4 .

Провести корреляционно-регрессионный анализ зависимости себестоимости одного изделия от выпуска продукции.

Решение. Признак X – объем выпускаемой продукции, тыс. шт. (факторный признак). Признак Y – себестоимость одного изделия, р. (результативный признак). Предполагаем, что признаки имеют нормальный закон распределения. Признаки находятся в статистической зависимости, так как себестоимость одного изделия зависит не только от объема выпускаемой продукции, но и от многих других факторов, которые в данном случае не учитываются. Определим форму связи. Построим точки с координатами (x_i , y_i) и по их расположению определим форму связи (рис. 6).

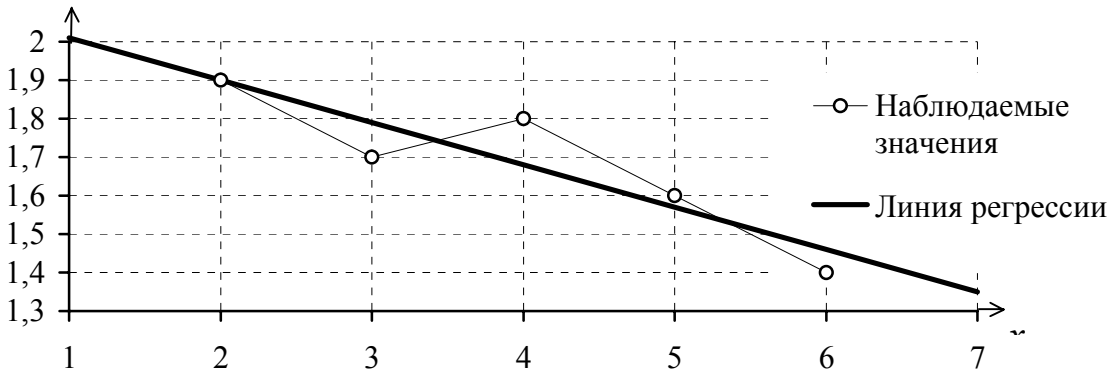


Рис. 6

Итак, форма связи линейная.

Проведем корреляционный анализ. Рассчитаем выборочный линейный коэффициент корреляции:

$$r_в = \frac{\overline{xy}_в - \bar{x}_в \cdot \bar{y}_в}{\sqrt{\overline{x_в^2} - (\bar{x}_в)^2}\sqrt{\overline{y_в^2} - (\bar{y}_в)^2}}$$

Расчеты представим в таблице:

	x_i	y_i	$x_i \cdot y_i$	x_i^2	y_i^2
	2	1,9	3,8	4	3,61
	3	1,7	5,1	9	2,89
	4	1,8	7,2	16	3,24
	5	1,6	8,0	25	2,56
	6	1,4	8,4	36	1,96
Итого	20	8,4	32,5	90	14,26.

$$\bar{x}_в = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{20}{5} = 4; \quad \bar{y}_в = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{8,4}{5} = 1,68;$$

$$\overline{xy}_в = \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \frac{32,5}{5} = 6,5;$$

$$\overline{x_в^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{90}{5} = 18; \quad \overline{y_в^2} = \frac{1}{n}\sum_{i=1}^{n} y_i^2 = \frac{14,26}{5} = 2,852;$$

$$\sigma_x^2 = \overline{x_в^2} - \bar{x}_в^2 = 18 - 16 = 2; \quad \sigma_y^2 = \overline{y_в^2} - \bar{y}_в^2 = 2,852 - (1,68)^2 = 0,0296;$$

$$r_в = \frac{6,5 - 4 \cdot 1,68}{\sqrt{2 \cdot 0,0296}} \approx -\ 0,90.$$

Проверим значимость выборочного коэффициента корреляции. Для этого выдвигаем гипотезы:

H_0: $r_Г = 0$,

H_1: $r_Г \neq 0$. Примем уровень значимости $\alpha = 0,05$.

Для проверки нулевой гипотезы используем случайную величину $T_{набл} = \frac{|r_в|}{\sqrt{1 - r_в^2}} \cdot \sqrt{n-2}$, имеющую распределение Стьюдента с $k = n - 2 = 3$ степенями свободы. По выборочным данным находим наблюдаемое значение критерия $T_{набл} = \frac{|-0,90| \cdot \sqrt{3}}{\sqrt{1 - 0,81}} \approx 3,58$. По таблице критических точек распределения Стьюдента находим $t_{кр.дв}(0,05; 3) = 3,18$. Сравниваем $T_{набл}$ и $t_{кр}(0,05; 3)$. Так как $T_{набл} > t_{кр}$, то есть $T_{набл}$ попало в критическую область, нулевая гипотеза отвергается, справедлива конкурирующая гипотеза: $r_Г \neq 0$, значит, $r_в$ значим. Признаки X и Y коррелированны. Так как $|r_в|$ близок к единице, следовательно, себестоимость одного изделия и объем выпускаемой продукции находятся в тесной корреляционной зависимости.

Найдем коэффициент детерминации. $D = r_в^2 \cdot 100\,\% = 0{,}81\,\%$, то есть вариация себестоимости единицы продукции в среднем на 81 % объясняется вариацией объема выпускаемой продукции.

Выразим эту связь аналитически приблизительно в виде линейного уравнения регрессии:

$$\bar{y}_x - \bar{y}_в \approx b(x - \bar{x}_в),$$

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x^2} = \frac{-0{,}22}{2} = -0{,}11.$$

$$\bar{y}_x - 1{,}68 = -0{,}11\,(x - 4) \text{ или } \bar{y}_x \approx -0{,}11x + 2{,}12.$$

Из уравнения следует, что с увеличением выпуска продукции на 1 тыс. шт. себестоимость одного изделия снизится в среднем на 0,11 р.

Найдем по уравнению регрессии себестоимость одного изделия, если выпуск продукции составит 5,2 тыс. шт. :

$$\bar{y}_x \approx -0{,}11 \cdot 5{,}2 + 2{,}12 = 1{,}55 \text{ (р.)}$$

Пример 4. Для нормирования труда проведено статистическое исследование связи между количеством изготавливаемых изделий (X, шт.) и затратами времени на обработку одного изделия (Y, мин). Сделана выборка объемом $n = 51$ и получены следующие данные: $r_в = 0{,}8$, $\bar{x} = 8$, $\sigma_x = 3{,}2$, $\bar{y} = 40$, $\sigma_y = 8$. Проверить значимость коэффициента корреляции при $\alpha = 0{,}02$. Построить уравнение регрессии.

Решение. Признак X – количество изготавливаемых изделий, шт. Признак Y – затраты времени на обработку одного изделия, мин.

Предполагаем, что признаки имеют нормальный закон распределения. Они находятся в статистической зависимости, так как затраты времени зависят не только от количества изготавливаемых изделий, но и от многих других факторов, которые в данном случае не учитываются. В данном случае связь линейная, теснота связи характеризуется линейным коэффициентом корреляции $r_в = 0{,}8$. Но прежде чем делать вывод о тесноте взаимосвязи, необходимо проверить значимость коэффициента корреляции. Выдвигаем нулевую гипотезу и ей конкурирующую:

H_0: $r_Г = 0$,

H_1: $r_Г \neq 0$.

Проверяем нулевую гипотезу с помощью случайной величины, имеющей распределение Стьюдента с $k = n - 2 = 49$ степенями свободы: $T = \dfrac{|r_в|}{\sqrt{1 - r_в^2}} \cdot \sqrt{n - 2}$.

По выборочным данным найдем наблюдаемое значение критерия $T_{набл} = \dfrac{0{,}8 \cdot \sqrt{49}}{\sqrt{1 - 0{,}64}} \approx 9{,}33$. По таблице критических точек распределения Стьюдента находим $t_{кр.дв}(\alpha, k) = t_{кр.дв}(0{,}02;\ 49) = 2{,}40$. Сравниваем $T_{набл}$ и

$t_{кр.дв}(0{,}02;\ 49)$. Так как $|T_{набл}| > t_{кр.дв}(0{,}02;\ 49)$, то есть наблюдаемое значение критерия попало в критическую область, нулевая гипотеза отвергается, справедлива конкурирующая гипотеза: $r_г \neq 0$, признаки X и Y коррелированны, $r_в$ значим.

$D = r_в^2 \cdot 100\,\% = 64\,\%$, то есть вариация затрат времени на обработку одного изделия в среднем на 64 % объясняется за счет вариации количества изготавливаемых изделий.

Выразим эту взаимосвязь аналитически в виде уравнения регрессии вида:

$$\bar{y}_x - \bar{y} \approx b(x - \bar{x}).$$

Коэффициент b выразим через парный линейный коэффициент корреляции:

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x^2}; \quad r_в = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y}.$$

Сравнивая эти две формулы, можем записать:

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y} \cdot \frac{\sigma_y}{\sigma_x} = r_в \cdot \frac{\sigma_y}{\sigma_x}.$$

Тогда по выборочным данным будем иметь:

$$b = 0{,}8 \cdot 8/32 = 2; \ \ \bar{y}_x - 40 \approx 2(x - 8) \ \text{ или } \ \bar{y}_x \approx 24 + 2x\ .$$

Из уравнения следует, что с увеличением количества выпускаемых изделий на 1 шт., затраченное время в среднем увеличится на 2 мин.

КОНТРОЛЬНЫЕ ВОПРОСЫ

по курсу «Математическая статистика»

1. Предметы и методы математической статистики. Задачи математической статистики. Генеральная и выборочная совокупности.
2. Выборочные аналоги интегральной функций распределения.
3. Выборочные аналоги дифференциальной функций распределения.
4. Статистические характеристики вариационных рядов.
5. Понятие о точечной оценке числовой характеристики случайной величины. Свойства точечных оценок.
6. Точечные оценки математического ожидания и их свойства.
7. Точечные оценки дисперсии и их свойства.
8. Частость как точечная оценка вероятности.
9. Понятие об интервальной оценке параметров распределения.
10. Доверительный интервал для математического ожидания при известном σ.
11. Доверительный интервал для математического ожидания при неизвестном σ.
12. Интервальная оценка вероятности.
13. Определение объема выборки.
14. Понятие статистических гипотез их виды. Понятие ошибки первого и второго рода.

15. Основной принцип проверки статистических гипотез
16. Понятие односторонней и двусторонней критической области. Правило нахождения критических точек.
17. Проверка гипотез о среднем значении нормально распределенной СВ при известной дисперсии
18. Проверка гипотез о среднем значении нормально распределенной СВ при неизвестной дисперсии
19. Исключение грубых ошибок наблюдений
20. Построение теоретического закона распределения по опытным данным.
21. Критерий согласия Пирсона.
22. Понятие функциональной, стохастической и корреляционной зависимости. Функция регрессии.
23. Генеральное и выборочное корреляционные отношения.
24. Линейное уравнение регрессии.
25. Генеральный и выборочный коэффициенты корреляции.
26. Нелинейные функции регрессии.
27. Понятие о множественной регрессии.

ЛИТЕРАТУРА

1. *Гмурман В.Е.* Теория вероятностей и математическая статистика. –М.: Высшая школа, 1977.
2. *Гмурман В.Е.* Руководство к решению задач по теории вероятностей и математической статистике. –М.: Высшая школа, 1997.
3. *Калинина В.Н., Панкин В.Ф.* Математическая статистика. –М.: Высшая школа, 1994.
4. *Мацкевич И.П., Свирид Г.П., Булдык Г.М.* Сборник задач и упражнений по высшей математике (Теория вероятностей и математическая статистика).– Минск: Вышейша школа, 1996.
5. *Тимофеева Л.К., Суханова Е.И.* Математика для экономистов. Сборник задач по теории вероятностей и математической статистике. –М.: УМиИЦ «Учебная литература», 1998. –182 с.
6. *Кремер Н.Ш.* Теория вероятностей и математическая статистика. –М.: ЮНИТИ, 2001.–542с.
7. *Гусак А.А., Бричикова Е.А.* Справочное пособие к решению задач. Теория вероятностей.– Минск: ТетраСистемс, 2000.–287с.
8. *Гурский Е.И.* Сборник задач по теории вероятностей и математической статистике.–Минск: Вышэйшая школа, 1975.–250с.
9. *Вентцель Е.С.* Теория вероятностей.–М: из-во физико-математической литературы, 1962–564с.