Брянская государственная сельскохозяйственная академия

Комогорцев В. Ф.

ЛЕКЦИИ

по теории вероятностей и математической статистике

для студентов экономических специальностей

Брянск – 2009

ОГЛАВЛЕНИЕ

Введение

В окружающей нас жизни есть много чего такого, что невозможно предсказать заранее. Например, невозможно точно предсказать:

- погоду в предстоящий день;
- исход футбольного матча;
- сторону, которой выпадет подброшенная монета;
- оценку на предстоящем экзамене;
- Процент выполнения производственного задания, ход его выполнения, качество выполнения, и так далее.

Такого рода *события, величины, процессы* называются *случайными*. Их случайность связана с тем, что многие факторы, их определяющие, нам не известны. Или этих факторов слишком много, что не позволяет все их учесть. Случайные события, величины, процессы изучает специальная наука - *теория вероятностей.*

Отметим сразу, что теория вероятностей не ставит задачу предсказать, произойдет или не произойдет интересующее нас случайное событие (скажем, выигрыш в лотерею), или какой будет случайная величина (сумма выигрыша), или как пойдет интересующий нас процесс (производственный, политический, и т. д.), да и не может эту задачу решить. Но при большом числе опытов (испытаний) по наблюдению за случайными явлениями в результатах испытаний обнаруживаются определенные закономерности. Изучением этих закономерностей и занимается теория вероятностей.

Исторически теория вероятностей зарождалась как теория азартных игр. В частности, как теория игры в кости (Гюйгенс, Паскаль, Ферма, 17 век).

Следующий этап развития теории вероятностей связан с именем швейцарского математика Якова Бернулли (вторая половина 17-го века), который доказал основную теорему вероятности - «закон больших чисел». После открытия этого закона теория вероятностей становится уже наукой.

Крупнейшими представителями этой науки в 18 веке и первой половине 19-го века были французские математики Лаплас и Пуассон и немецкий математик Гаусс.

Особенно быстро тория вероятностей развивалась во второй половине 19-го и 20-м веке в связи с возникновением и развитием другой вероятностной науки – математической статистики, основы которой мы тоже рассмотрим (во второй части этого курса).

В развитие теории вероятностей и математической статистики крупный вклад внесли и наши отечественные ученые. В первую очередь это члены знаменитой Петербургской математической школы П.Л.Чебышев, А.М.Ляпунов, А.А.Марков (конец 19-го - начало 20-го века). А также работавшие уже в советское время академики А.Н.Колмогоров, С.Н.Бернштейн, А.Н. Хинчин, Н.В. Смирнов и др. Из западных ученых развитие теории вероятностей и математической статистики в наибольшей степени связано с именами английских и американских ученых Стьюдента, Фишера, Пирсона и некоторых других.

В настоящее время теория вероятностей и математическая статистика принадлежат к числу важнейших разделов математики и широко применяются в самых различных отраслях науки и техники. Например, в биологии (исследование процессов гибели и размножения живых организмов); в теоретической физике (физика микромира, статистическая физика); в военном деле (теория стрельб); в машиностроении (теория надежности механизмов и машин); в экономике (организация и планирование производства), и т.д. В связи с такой широтой области применения теории вероятностей и математической статистики и в связи с их все возрастающим значением эти науки входят в настоящее время в учебные программы практически всех инженерно-технических, технологических, экономических, сельскохозяйственных и многих других вузов. И если не каждому специалисту в своей практической деятельности приходится применять вероятностные методы, то уж знакомство с основными понятиями и идеями вероятностных методов необходимо каждому – хотя бы для того, чтобы понимать выводы и рекомендации, вытекающие из использования этих методов.

Часть I
ТЕОРИЯ ВЕРОЯТНОСТЕЙ

Глава 1
Случайные события и их вероятности

§ 1. Случайное событие и вероятность его появления

Определение. *Событие A называется случайным, если заранее, до производства испытания (эксперимента, наблюдения) неизвестно, произойдет оно или нет.*

Примеры:

1) Испытание - покупка лотерейного билета; случайное событие A – выигрыш по этому билету.

2) Испытание - выстрел стрелка по мишени; случайное событие A – попадание стрелка в мишень.

3) Испытание - подбрасывание монеты; случайное событие A - выпадение орла.

4) Испытание - приход студента на экзамен; случайное событие A – сдача им экзамена.

5) Испытание- посадка в землю семени; случайное событие A – прорастание семени.

Интересующее нас случайное событие A может иметь как большую возможность (много шансов) для своего появления, так и не очень. Если известно, что в повторяемых испытаниях событие A наступает часто – у него, естественно, будет большая возможность появления в каждом отдельном испытании. А если редко - то небольшая.

Возникает естественная задача *численной оценки* (оценки числом) степени возможности появления в отдельном испытании любого интересующего нас случайного события A. Для решения этой задачи в теории вероятностей разработано понятие *вероятности* $p(A)$ появления случайного события A в единичном испытании. Связывается это понятие *со средней долей* появления события A в сериях повторных испытаний.

Пусть n – число повторных испытаний (число купленных лотерейных билетов, число выстрелов по мишени, и т.д.), а $k_{ср}$ – среднее число появлений события A в этих n повторных испытаний (среднее число выигрышных билетов из n купленных билетов, среднее число попаданий в мишень при n выстрелах по

мишени, и.т.д.). Тогда, по определению, в качестве *вероятности p(A) появления события A в одном испытании* принимается величина

$$p(A) = \frac{k_{ср}}{n} \qquad (1.1)$$

То есть $p(A)$ – это *средняя доля* появлений события A в повторных испытаниях.

Кстати, если известен средний процент $C_{ср}\%$ появления события A в повторных испытаниях, то среднюю долю появления события A, а следовательно, и его вероятность $p(A)$ появления в каждом отдельном испытании можно получить и по такой, вытекающей из (1.1), формуле:

$$p(A) = \frac{C_{ср}\%}{100\%} \qquad (1.2)$$

Пусть, например, случайное событие A - это попадание стрелка в мишень. И пусть из опытов известно, что из каждых 10 выстрелов (испытаний) стрелок попадает в цель в среднем 7 раз. Тогда средняя доля попадания стрелка в мишень при повторении выстрелов равна $\frac{k_{ср}}{n} = \frac{7}{10} = 0{,}7$. Это и есть, согласно (1.1), вероятность $p(A)$ попадания стрелка в мишень с одного выстрела: $p(A) = 0{,}7$.

Можно рассудить и по-другому: если стрелок попадает в мишень в среднем 7 раз из 10, то это значит, что он в среднем попадает 70 раз из 100, то есть в среднем в 70% выстрелов. А значит, $C_{ср}\% = 70\%$, откуда по формуле (1.2) получаем: $p(A) = 0{,}7$. То есть получаем тот же самый результат.

Обратно, зная, что вероятность $p(A)$ попадания стрелка в мишень с одного выстрела (в одном испытании) равна 0,7, на основании формул (1.1) и (1.2) делаем вывод, что средняя доля попаданий такого стрелка в мишень при повторных выстрелах равна 0,7. А это значит, что этот стрелок попадает в мишень *в среднем* 7 раз из 10, или 70 раз из 100, или 700 раз из 1000 , то есть *в среднем* в 70 % выстрелов.

На практике, то есть с помощью реально проводимых испытаний, среднюю долю появления события A , а следовательно, и вероятность $p(A)$ появлений события A в одном испытании можно, очевидно, найти лишь приближенно, ибо реальная доля появлений события A в сериях повторных испытаний от серии к серии меняется. Для достаточно надежного определения этой доли должно быть произведено много повторных испытаний (чем больше, тем лучше). Но в некоторых случаях указанную долю, а значит, и вероятность $p(A)$ появления события A в одном испытании можно найти чисто умозрительно, без производства повторных испытаний. Причем найти не приближенно, а точно.

Пусть, например, нам известны все возможные исходы одного испытания, а, следовательно, мы знаем и их общее число n (полагаем, что n - число конечное). И пусть все эти исходы заведомо *равновозможны*. Равновозможность исходов означает отсутствие каких-либо преимуществ по появлению у каждого исхода испытания по сравнению с любым другим возможным исходом. Равно-

возможные исходы испытания при повторении испытаний появляются в среднем одинаково часто.

Далее, пусть известно число m тех исходов, которые благоприятствуют появлению интересующего нас события A (это значит, что известно общее число m тех исходов испытания, при осуществлении которых автоматически появляется событие A). При повторении испытаний равновозможные исходы испытания будут наступать в среднем одинаково часто. А значит, если испытаний будет n, то каждый из возможных исходов испытания появится в среднем один раз. Тогда событие A в любых n испытаниях будет появляться в среднем m раз. То есть $k_{ср} = m$. А тогда из формулы (1.1) получаем:

$$p(A) = \frac{m}{n} \qquad (1.3)$$

Формула (1.3) называется *классической формулой*. Напомним, что в ней n *– число всех возможных исходов испытания*, а m *– число благоприятствующих событию A исходов*. Ее можно применять лишь в том случае, когда все n возможных исходов испытания *равновозможны*. Классическая формула (1.3) позволяет находить вероятности случайных событий без производства испытаний, чисто умозрительно. Причем находить эти вероятности *точно*.

Пример1. Пусть испытание – это подбрасывание монеты, а событие A - это выпадение орла. В этом испытании два возможных исхода – орел и решка. То есть n = 2. В силу симметрии монеты эти два исхода равновозможны. Из них лишь один исход благоприятствует событию A (m=1).Тогда по классической формуле(1.3) получаем: $p(A) = \frac{1}{2} = 0{,}5.$

Полученный результат очевидным образом верен. Действительно, при повторных бросаниях симметричной монеты орел будет выпадать *в среднем* в 50% бросаний (эта цифра будет другой, если только монета не симметрична, например, погнута). И по формуле (1.2) получаем то же самое: $p(A) = 0{,}5$.

Кстати, если подбрасываемую монету нельзя считать симметричной (она как–то деформирована, так что симметрия ее сторон нарушена), то и в этом случае n = 2 – число всех возможных исходов испытания (орел и решка), а m = 1- единственный благоприятствующий событию A исход (орел). Но в данном случае классическую формулу (1.3) применять нельзя, ибо исходы испытания (орел и решка) не равновозможны. Тогда для определения вероятности $p(A)$ выпадения орла при подбрасывании монеты остается один путь: бросать монету много раз (повторять испытания) и искать опытным путем $C_{ср.}$% - средний процент появления события A (выпадения орла). После нахождения этого среднего процента можно будет по формуле (1.2) нейти искомую вероятность $p(A)$ – вероятность выпадения герба при одном бросании данной монеты. Естественно, что таким путем $p(A)$ можно найти лишь приближенно (тем точнее, чем больше будет произведено повторных бросаний монеты).

Пример2. Пусть испытание – это подбрасывание игральной кости (кубика с пронумерованными гранями), а событие A – выпадение цифры 5. В указанном испытании 6 возможных исходов, которые все равновозможны (n=6). А m=1 –

один благоприятствующий событию A исход (выпадение пятерки). Тогда по классической формуле (1.3) получаем: $p(A)=\frac{1}{6}$. Это цифра, в соответствии с формулой (1.1), означает, что из каждых 6 испытаний (из каждых 6 бросаний кубика) пятерка будет выпадать *в среднем* один раз. Это совершенно согласуется и со здравым смыслом.

Пример3. Пусть испытание – это вынимание наудачу одной карты из колоды в 36 карт. А событие A – это появление туза (любого). В указанном испытании n= 36 возможных исходов (любая из 36 карт может оказаться вынутой, и вынимание каждой из них – это возможный исход испытания). Все эти исходы, очевидно, равновозможны. А m=4 – число благоприятствующих событию A исходов (в колоде 4 туза). Тогда по классической формуле получаем: $p(A)=\frac{4}{36}=\frac{1}{9}$. Эта цифра, в соответствии с формулой (1.1), означает, что из каждых 36 испытаний (опытов по выниманию наудачу одной карты из колоды) событие A (туз) будет появляться *в среднем* 4 раза. Или, что одно и то же, из 9 испытаний событие A (туз) будет появляться *в среднем* 1 раз. Это совершенно согласуется и со здравым смыслом.

Примечание. Проведем, для сравнения, и неправильное решение этой же задачи. В указанном испытании всего два возможных исхода (n=2): вынимание туза и вынимание любой другой карты. А m=1 - один благоприятствующий событию А исход. И тогда, по классической формуле, получаем: $p(A)=\frac{1}{2}=0{,}5.$

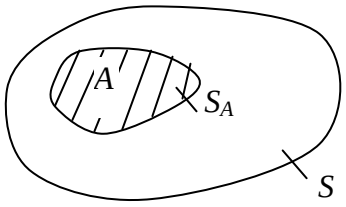


Рис. 1.1

Но этот результат неверен, да он и не согласуется со здравым смыслом. Ошибка решения состоит в том, что оба указанные выше возможные исходы испытания не равновозможны, а такие исходы в классической формуле (1.3) применять нельзя.

Вероятности появления любого события A можно дать *наглядную геометрическую интерпретацию*. Для этого условимся представлять себе любое испытание как бросание наудачу точки на некоторую плоскую область площадью S, а событием A будем считать попадание брошенной точки на некоторую часть этой области площадью S_A (рис.1.1):

При повторных бросаниях средняя доля попаданий брошенной наудачу точки на область A, а значит и вероятность $p(A)$ появления события A при каждом отдельном бросании (испытании) будет определяться, очевидно, отношением площадей S_A и S. То есть

$$p(A)=\frac{S_A}{S} \qquad (1.4)$$

Эту очень полезную в силу ее наглядности геометрическую интерпретацию вероятности появления любого случайного события A мы используем позже, при выводе некоторых важных формул теории вероятностей.

Основные свойства вероятности случайного события.

1. Вероятность $p(A)$ появления случайного события A представляет собой *число*. Это число является *мерой возможности* появления события A при производстве испытания. В частности, в соответствии с (1.2), вероятность $p(A)$ определяет *средний процент* $C_{ср}\%$ появления события A в повторных испытаниях:

$$C_{ср}\% = p(A)\cdot 100\ \% \qquad (1.5)$$

Чем больше $p(A)$, тем возможнее (вероятнее) появление события A в одном испытании, и тем чаще оно будет появляться при повторении испытаний. Например, если $p(A)=0{,}3$, то это, в соответствии с формулой (1.5), означает, что событие A будет появляться *в среднем* в 30% испытаний (в среднем 30 раз из 100 или 3 раза из 10 испытаний). То есть в каждом испытании у события A имеется 3 шанса из 10 (30 шансов из 100) за то, что оно появится.

2. Для любого случайного события A

$$0 \le p(A) \le 1 \qquad (1.6)$$

Это очевидным образом вытекает из всех приведенных выше формул (1.1) – (1.4), определяющих величину $p(A)$ (продумайте это).

3. Если появление события A в испытании невозможно, то вероятность $p(A)$ его появления при осуществлении испытания равна нулю: $p(A) = 0$.

Действительно, сколько бы раз мы ни повторяли испытание, невозможное событие не появится ни разу. Таким образом, $C_{ср}\% = 0\%$, а значит, по формуле (1.2) получаем: $p(A) = 0$. Этот же результат для невозможного события (продумайте это) следует и из формул (1.1), (1.3) и (1.4).

4. Если $p(A) = 0$, то событие A невозможно (или возможно, но лишь чисто теоретически). Наиболее очевидным образом это утверждение следует из формулы (1.4) и рис.1. Действительно, если $p(A) = 0$, то $\frac{S_A}{S} = 0$, откуда $S_A = 0$. А это будет в том случае, если область A отсутствует вообще (и значит, событие A невозможно), или эта область A имеется, но ее площадь S_A равна нулю (например, область A – это некоторая фиксированная точка на области S). В последнем случае событие A, имея нулевую вероятность, все – таки возможно. Но за его появление имеется один шанс, а против его появления – бесконечное количество шансов (одна точка области S против бесчисленного количества остальных точек этой области). Ясно, что практических возможностей для своего появления у такого события A нет. Событие A возможно, но лишь чисто теоретически.

5. Если появление события A в испытании достоверно (событие A произойдет обязательно), то вероятность $p(A)$ его появления при осуществлении испытания равна единице: $p(A) = 1$.

Действительно, достоверное событие будет появляться в каждом испытании, а значит, в 100% испытаний. То есть в этом случае $C_{ср}\% = 100\%$. Тогда по формуле (1.2) получаем: $p(A) = 1$. Это же результат для достоверного события следует и из формул (1.1), (1.3) и (1.4) (продумайте это).

6. Если $p(A) = 1$, то событие A или достоверно (произойдет обязательно), или практически достоверно (непоявление события A возможно, но лишь чисто теоретически). Например, таким практически достоверным событием A, имеющим единичную вероятность, будет непопадание брошенной наудачу точки на некоторую заданную точку области S (см. рис.1)

Упражнения

1. Какой должна быть $p(A)$, чтобы можно было утверждать, что появление события A более вероятно, чем его непоявление?

Ответ: $0{,}5 < p(A) \le 1$

2. Доказать свойства (1) – (6) вероятности $p(A)$ для тех событий, чьи вероятности можно находить по классической формуле (1.3).

3. Какова вероятность того, что при двукратном подбрасывании монеты оба раза выпадает орел?

Ответ: $\frac{1}{4}$

4. Какова вероятность того, что при подбрасывании двух игральных костей сумма выпавших очков будет больше 10?

Ответ: $\frac{1}{12}$

5. Шифр замка автоматической камеры хранения состоит из одной буквы русского алфавита (в нем 33 буквы) и трех цифр (каждая от 0 до 9). Найти вероятность открыть камеру, если набрать шифр наудачу.

Ответ: $\frac{1}{33000}$

6. На круг наудачу брошена точка. Найти вероятность того, что она попадет во вписанный в круг квадрат.

Ответ: $\frac{2}{\pi} \approx 0{,}64$

7. В трапеции с основаниями 6 см и 2 см проведена диагональ, и на трапецию наудачу брошена точка. Каковы вероятности попадания брошенной точки в треугольники, на которые разбилась трапеция?

Ответ: $\frac{1}{4}$ и $\frac{3}{4}$

8. Какова вероятность того, что первый встречный прохожий родился не весной?

Ответ: $\frac{3}{4}$

§ 2. Некоторые сведения из комбинаторики

Комбинаторика - это часть так называемой *дискретной математики*, изучающая разнообразные соединения элементов. Под элементами понимаются любые однотипные вещи: предметы, буквы, числа, живые существа и т.д. Различают 3 вида соединений элементов:

- *Размещения;*
- *Перестановки;*
- *Сочетания*.

1. *Размещения.* Пусть рассматривается совокупность из n упорядоченных, т.е. пронумерованных, элементов ($a_1; a_2;\ldots a_n$). Будет составлять из этих n элементов всевозможные *упорядоченные группы* по m элементов в каждой группе, где m – любое натуральное число, не превосходящее n. Эти группы будем считать различными, если они отличаются друг от друга хотя бы одним элементом или даже только порядком следования элементов в группе (у каждого элемента в упорядоченной группе есть свое учитываемое место). Такие группы называются *размещениями из n элементов по m элементов в каждом размещении*. Их общее число обозначается символом A_n^m, и находится оно по формуле:

$$A_n^m = \frac{n!}{(n-m)!} \qquad (2.1)$$

Напомним, что выражение $n!$ называется эн–факториал, и определяется оно так:

$0!=1;\quad 1!=1;\quad 2!=1\cdot 2=2;\quad 3!=1\cdot 2\cdot 3=6;\quad 4!=1\cdot 2\cdot 3\cdot 4=24;\ldots\quad n!=1\cdot 2\cdot 3\cdot\cdot n.$ (2.2)

Доказательство формулы (2.1).

1) Составим сначала все возможные размещения из n элементов ($a_1; a_2;\ldots a_n$) по одному элементу в каждом размещении и подсчитаем их число A_n^1. Эти размещения – просто отдельно взятые элементы $a_1; a_2;\ldots a_n$. Их количество равно n. То есть

$$A_n^1=n \qquad (2.3)$$

2) Составим теперь все возможные размещения из n элементов по два элемента в каждом размещении и подсчитаем их число A_n^2. Эти размещения тоже очевидны:

$a_1\ a_2$	$a_2\ a_1$	$a_3\ a_1$	$a_n\ a_1$
$a_1\ a_3$	$a_2\ a_3$	$a_3\ a_2$	$a_n\ a_2$
$a_1\ a_4$	$a_2\ a_4$	$a_3\ a_4$	$a_n\ a_3$
…..…	…..…	…..…	…..…
$a_1\ a_n$	$a_2\ a_n$	$a_3\ a_n$	$a_n\ a_{n-1}$

(2.4)

В каждом из столбцов (2.4) n -1 размещение, а всех столбцов n, поэтому

$$A_n^2=n(n-1) \qquad (2.5)$$

3) Составим все возможные размещения из n элементов по три элемента в каждом размещении и подсчитаем их число A_n^3. Эти размещения получим, если возьмем за основу все возможные размещения (2.4) по два элемента и добавим

по очереди справа к каждому из них любой из оставшихся n-2 элементов. Таким образом, из каждого размещения (2.4) по два элемента можно образовать n-2 размещения по три элемента. Например, из одного размещение $a_1\ a_2$, содержащего два элемента, можно образовать следующее n-2 размещений по три элемента:

$$a_1\ a_2\ a_3;\quad a_1\ a_2\ a_4;\quad a_1\ a_2\ a_5;\ldots\ldots\ldots\ a_1\ a_2\ a_n \qquad (2.6)$$

Следовательно,

$$A_n^3 = A_n^2 \cdot (n\text{-}2) = n(n\text{-}1)(n\text{-}2) \qquad (2.7)$$

Аналогично:

$$A_n^4 = A_n^3 \cdot (n\text{-}3) = n(n\text{-}1)(n\text{-}2)(n\text{-}3) \qquad (2.8)$$

$$A_n^5 = A_n^4 \cdot (n\text{-}4) = n(n\text{-}1)(n\text{-}2)(n\text{-}3)(n\text{-}4)$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

То есть

$$A_n^m = n(n\text{-}1)(n\text{-}2)(n\text{-}3)\ldots..(n\text{-}m\text{+}1) \qquad (2.9)$$

Итак, мы получили формулу для A_n^m при произвольном m (m=1,2,…n).

Преобразуем ее к более удобному виду. Для этого домножим и разделим выражение (2.9) на убывающие недостающие множители $(n\text{-}m)(n\text{-}m\text{-}1)\cdot\cdots 3\cdot 2\cdot 1$ так, чтобы последний множитель в (2.9) стал 1:

$$A_n^m = \frac{n(n-1)(n-2)(n-3)\cdots(n-m+1)(n-m)(n-m-1)\cdots 3\cdot 2\cdot 1}{(n-m)(n-m-1)\cdots 3\cdot 2\cdot 1} =$$

$$= \frac{1\cdot 2\cdot 3\cdots n}{1\cdot 2\cdot 3\cdots(n-m)} = \frac{n!}{(n-m)!}$$

Формула (2.1) доказана.

Для примера подсчитаем общее количество всех возможных размещений из трех элементов (a_1; a_2; a_3) по одному, по два и по три элемента. Причем сделаем это и непосредственно, составив все эти размещения и пересчитав их, и по формуле (2.1):

$$\left.\begin{matrix} a_1 \\ a_2 \\ a_3 \end{matrix}\right] A_3^1 = 3; \qquad \left.\begin{matrix} a_1a_2 \\ a_1a_3 \\ a_2a_1 \\ a_2a_3 \\ a_3a_1 \\ a_3a_2 \end{matrix}\right] A_3^2 = 6; \qquad \left.\begin{matrix} a_1a_2a_3 \\ a_1a_3a_2 \\ a_2a_1a_3 \\ a_2a_3a_1 \\ a_3a_1a_2 \\ a_3a_2a_1 \end{matrix}\right] A_3^3 = 6 \qquad (2.11)$$

Эти же результаты дает и формула (2.1):

$$A_3^1 = \frac{3!}{2!} = \frac{6}{2} = 3;\quad A_3^2 = \frac{3!}{1!} = \frac{6}{1} = 6;\quad A_3^3 = \frac{3!}{0!} = \frac{6}{1} = 6 \qquad (2.12)$$

2. *Перестановки.* Перестановками из данной совокупности n элементов (a_1; a_2;…a_n) называются различным образом упорядоченные (по разному переставленные) комбинации *всех этих элементов*. Их общее количество обозначается символом P_n. Так как перестановки – это по сути размещения из n элементов по n элементов в каждом размещении, то

$$P_n = A_n^n = \frac{n!}{0!} = n! \qquad (2.13)$$

3. *Сочетания.* Сочетаниями из n элементов по m элементов в каждом сочетании называются (в отличие от размещений) всевозможные *неупорядоченные группы* по m элементов в каждой группе. Неупорядоченные - это значит, что важно, какие элементы содержатся в каждом сочетании, а в каком порядке они там находятся – это неважно. Различные размещения отличаются друг от друга хотя бы одним элементом. Общее их количество обозначается символом C_n^m, и находится оно по формуле:

$$C_n^m = \frac{n!}{m!(n-m)!} \qquad (2.14)$$

Доказательство формулы (2.14)

Очевидно, что если взять все возможные сочетания из n элементов по m элементов и сделать в каждом из них все возможные перестановки, то в итоге получим все возможные размещения из n элементов по m элементов в каждом размещении. Отсюда следует:

$$C_n^m \cdot P_m = A_n^m, \text{ и значит } \quad C_n^m = \frac{A_n^m}{P_m} = \frac{n!}{m!(n-m)!} \qquad (2.15)$$

Для примера подсчитаем общее количество всех возможных сочетаний из трех элементов ($a_1; a_2; a_3$) по одному, по два и по три элемента. Причем сделаем это и непосредственно, составив все эти сочетания и пересчитав их, и по формуле (2.14):

$$\left.\begin{matrix} a_1 \\ a_2 \\ a_3 \end{matrix}\right] C_3^1 = 3 \qquad \left.\begin{matrix} a_1a_2 \\ a_1a_3 \\ a_2a_3 \end{matrix}\right] C_3^2 = 3 \qquad a_1a_2a_3\,] \, C_3^3 = 1 \qquad (2.16)$$

Эти же результаты дает и формула (2.14):

$$C_3^1 = \frac{3!}{1!\cdot 2!} = \frac{6}{1\cdot 2} = 3; \quad C_3^2 = \frac{3!}{2!\cdot 1!} = \frac{6}{2\cdot 1} = 3; \quad C_3^3 = \frac{3!}{3!\cdot 0!} = \frac{6}{1\cdot 6} = 1; \qquad (2.17)$$

Кстати, комбинации в (2.11) и в (2.16) наглядно демонстрируют разницу между размещениями и сочетаниями.

Выводя формулы (2.1), (2.13) и (2.14) для общего числа размещений, перестановок и сочетаний, мы полагали, что в каждом из указанных соединений любой из элементов совокупности ($a_1; a_2; \ldots a_n$) может встретиться только один раз. То есть повторять в них элементы нельзя. Если же повторять их всё же можно, то мы придём к *размещениям, перестановкам и сочетаниям с повторениями*.

4. *Размещения с повторениями.* Начнём с рассмотрения частного случая. Пусть исходная совокупность элементов составляет всего три элемента ($a_1; a_2; a_3$). Составим из них все возможные размещения с повторениями по два элемента в каждом размещении и пересчитаем их. Очевидно, их будет 9 – те 6 , что представлены в (2.11) и в которых оба элемента разные, и ещё три размещения $a_1a_1; a_2a_2; a_3a_3$ с повторяющимися элементами. Если обозначить общее число всех возможных размещений из трёх элементов по два в каждом размещении символом $\tilde{A}_3^2$, то получим: $\tilde{A}_3^2 = 9$.

Впрочем, мы могли подсчитать это число и иначе. Составляя любое размещение из двух элементов, на первое место в таком размещении можно поставить любой из данных трёх элементов (три варианта). На второе место – тоже любой из трёх элементов (тоже три варианта). Комбинируя каждый элемент, стоящий на первом месте, с каждым элементом, стоящим на втором месте, получим $3^2=9$ всех возможных комбинаций. То есть $\tilde{A}_3^2=3^2=9$.

А теперь легко понять, что если всех элементов не три, а n, и из них составляются все возможные размещения с повторениями по m элементов в каждом размещении, то их общее число $\tilde{A}_n^m$ найдётся по формуле:

$$\tilde{A}_n^m = n^m \qquad (2.18)$$

5. *Сочетания с повторениями.* Опять начнём с частного случая. А именно, подсчитаем $\tilde{C}_3^2$ – общее число всех возможных сочетаний с повторениями из трёх элементов ($a_1; a_2; a_3$) по два элемента в каждом сочетании. Этих сочетаний, очевидно, будет 6 – те 3, которые представлены в (2.16) и в которых оба элемента разные, и ещё три сочетания $a_1a_1; a_2a_2; a_3a_3$ с повторяющимися элементами. То есть $\tilde{C}_3^2=6=C_4^2$. И вообще, можно доказать, что

$$\tilde{C}_n^m = C_{n+m-1}^m \qquad (2.19)$$

6. *Перестановки с повторениями.* Пусть среди элементов ($a_1; a_2; \ldots a_n$) содержится лишь k различных элементов ($k<n$), причём первый из них повторяется n_1 раз, второй n_2 раз, …k-ый n_k раз. Очевидно, что $n_1+n_2+\ldots n_k=n$. Тогда число всех возможных перестановок из таких n элементов обозначается символом $P_n(n_1; n_2; \ldots n_k)$ и находится по формуле:

$$P_n(n_1; n_2; \ldots n_k) = \frac{n!}{n_1!\,n_2!\cdots n_k!} \qquad (2.20)$$

В самом деле, если бы все n элементов были разными, то число всех возможных перестановок из них, согласно (2.13), было бы равно $P_n = n!$. Но среди них разных элементов лишь k, остальные $n-k$ элементов повторяют эти k элементов. В частности, первый из этих k элементов повторяется n_1 раз. Сделав любую перестановку из этих n_1 одинаковых элементов (не трогая остальных!) мы ничего не нарушим ни в какой перестановке из n элементов. А вот если бы все эти n_1 повторяющихся элементов были разными, то сделав любую их перестановку, мы получили бы другую перестановку из n элементов. Количество всех возможных перестановок из n_1 элементов равно $n_1!$. Значит, наличие этих n_1 одинаковых элементов уменьшает в $n_1!$ раз общее количество всех возможных перестановок из n элементов по сравнению с тем случаем, когда все n элементов были бы разными. Тот же эффект производит наличие второго повторяющегося n_2 раз элемента, третьего, … k-ого. В итоге и приходим к формуле (2.20).

А теперь рассмотрим примеры на применение полученных выше формул.

Пример 1. В посёлке устанавливается телефонная сеть с трёхзначными телефонными номерами. Сколько всего можно установить телефонных номеров?

Сколько из них будет тех, которые содержат: 1) три разные цифры? 2) две одинаковые цифры? 3) три одинаковые цифры?

Решение. Всех возможных трёхзначных телефонных номеров будет, очевидно, 1000: это номера 000, 001, 002, … 999.

1) Номеров с тремя разными цифрами будет, очевидно, столько, сколько существует всех возможных соединений (комбинаций) из 10 цифр 0,1,2,…9 по три цифры в каждом соединении. Так как в этих соединениях важен порядок следования цифр (например, 137 и 173 – это разные номера), то этими соединениями будут размещения. А значит, их общее количество N_1 можно найти по формуле (2.1):

$$N_1 = A_{10}^3 = \frac{10!}{7!} = \frac{8 \cdot 9 \cdot 10}{1} = 720$$

3) Пропуская вопрос (2), ответим на вопрос (3). Номеров с тремя одинаковыми цифрами будет, очевидно, 10. Это номера 000, 111, 222, … 999. То есть $N_3 = 10$.

2) Все остальные номера – с двумя одинаковыми цифрами. Следовательно, их общее количество

$$N_2 = 1000 - (N_1 + N_3) = 1000 - (720 + 10) = 270.$$

Пример 2. Сколько всего диагоналей у выпуклого n – угольника?

Решение: Соединяя каждую пару вершин треугольника, получим либо диагональ, либо сторону многоугольника. Число всех различных пар вершин n-угольника равно числу всех возможных соединений из n элементов (вершин многоугольника) по два элемента (по две вершины) в каждом соединении. Так как порядок следования элементов в этих парах, очевидно, не важен (диагональ, соединяющая, например, 2-ю и 5-ю вершину – это та же диагональ, которая соединяет 5-ю и 2-ю вершину), то такими парными соединениями будут сочетания из n элементов по два элемента в каждом сочетании. Следовательно, их общее число равно C_n^2. В это число входят и сами n сторон многоугольника. Поэтому искомое число N диагоналей n-угольника найдётся по формуле:

$$N = C_n^2 - n = \frac{n(n-1)}{2} - n = \frac{n(n-3)}{2}$$

Пример 3. В чемпионате области по футболу участвуют 20 команд, причём каждые две из них встречаются между собой два раза (игры идут в два круга). Сколько всего матчей играется в течение сезона?

Решение. В первом круге состоится столько матчей, сколько существует сочетаний из 20 команд по две в каждом сочетании. То есть их число равно:

$$C^2{}_{20} = \frac{20!}{2! \cdot 18!} = \frac{19 \cdot 20}{2} = 190.$$

Во втором круге играется столько же матчей, поэтому в течение сезона их состоится $190 \cdot 2 = 380$.

Пример 4. Сколькими различными способами можно рассадить n человек за круглым столом?

Решение. Общее число всех способов, с помощью которых n человек может занять n мест за любым (не только круглым) столом равно, очевидно, числу всех перестановок из n элементов (из n человек), то есть равно $P_n = n!$. Но если

стол круглый, то любая циклическая перестановка (одновременная пересадка всех вправо или влево на одно или несколько мест) не нарушает порядка рассаживания людей за столом. А таких циклических пересаживаний всего n. Поэтому искомое число N_n рассаживания n человек за круглым столом равно:

$$N_n = \frac{P_n}{n} = \frac{n!}{n} = (n-1)!$$

В частности, N_2=1; N_3 = 2; N_4 = 6; N_5 = 24;…..

Пример 5. Автомобильные номера состоят из двух букв (всего используется 30 букв) и трёх цифр (используются все 10 цифр). Сколько всего автомобилей можно занумеровать таким образом, чтобы никакие два автомобиля не имели одинакового номера?

Решение: Различных пар букв будет столько, сколько можно составить размещений с повторениями из 30 букв по две в каждом размещении. То есть, согласно формуле (2.18), их будет:

$$\tilde{A}_{30}^{2} = 30^2 = 900.$$

Аналогично различных троек цифр будет:

$$\tilde{A}_{10}^{3} = 10^3 = 1000.$$

(впрочем, это и так очевидно). Комбинируя (соединяя) теперь каждую пару букв с каждой тройкой цифр, получим искомое общее число N различных автомобильных номеров:

$$N = 900 \cdot 1000 = 900000.$$

Пример 6. Имеются две колоды по 36 карт. Из каждой колоды вынимаются по одной карте. Сколько различных пар карт может быть при этом образовано?

Решение. В образовываемых парах карт порядок их следования, очевидно, не важен. Поэтому эти пары – различные сочетания из 36 карт. По условию, среди этих пар могут оказаться и пары с одинаковыми картами. То есть образованные пары карт – это сочетания с повторениями. А значит, их общее число:

$$N = \tilde{C}_{36}^{2} = C_{37}^{2} = \frac{37!}{2!\cdot 35!} = \frac{36\cdot 37}{2} = 666$$

Пример 7. Сколько всех возможных перестановок букв можно сделать в слове «математика»?

Решение. В слове «математика» всего 10 букв, из которых буква *а* повторяется 3 раза, буква *м* – два раза, буква т – два раза, остальные буквы (*е, и, к)* по разу. Поэтому искомое число N перестановок в слове «математика» - это число перестановок с повторениями. Согласно формуле (2.20),

$$N = P_{10}(3;2;2) = \frac{10!}{3!\cdot 2!\cdot 2!} = 151200$$

Пример 8. Сколькими различными способами могут занять места в президиуме 5 человек, если в президиуме 8 мест?

Решение. C_8^5 – число всех возможных способов выбора пяти различных мест из имеющихся восьми. На этих местах 5 человек можно рассадить числом способов $P_5 = 5!$ В итоге искомое число способов

$$N = C_8^5 \cdot P_5 = 6720$$

Пример 9. Из колоды в 36 карт наудачу вынимаются две карты. Какова вероятность того, что ими окажется два туза (любых)?

Решение. В данной задаче испытание – это вынимание из колоды наудачу двух карт, а событие A – вынимание двух тузов. Возможных исходов в испытании столько, сколько всего пар карт можно составить. Таких пар будет, очевидно, $n = C_{36}^2 = 630$. Все эти 630 возможных исходов испытания, очевидно, равновозможны. А число m исходов, благоприятствующих событию A – это, очевидно, число всех пар из четырёх тузов, то есть $m = C_4^2 = 6$. Поэтому по классической формуле (1.3) получаем:

$$p(A) = \frac{m}{n} = \frac{6}{630} = \frac{1}{105}$$

Пример 10. Ребёнок играет двумя карточками с буквой «М» и двумя карточками с буквой «А», выкладывая их в линию. Какова вероятность того, что у него получится слово «МАМА»?

Решение. В данной задаче испытание - это выкладывание ребёнком наудачу в линию четырёх карточек, а событие A – выкладывание слова «МАМА». Возможные исходы испытания – это различные перестановки четырёх карточек, причём перестановки с повторениями, в которых и буква М, и буква А повторяются два раза. Число таких перестановок, согласно формуле (2.20), равно

$$N = \frac{4!}{2!\cdot 2!} = \frac{24}{4} = 6$$

Это – число n всех возможных исходов испытания, причём исходов равновозможных. А число m исходов, благоприятствующих событию A, равно 1: m=1 (перестановка «МАМА»). В итоге по классической формуле (1.3) получаем:

$$p(A) = \frac{m}{n} = \frac{1}{6}$$

Пример 11. Из десяти билетов выигрышными являются два. Определить вероятность того, что среди наудачу взятых пяти билетов: а) только один выигрышный; б) оба выигрышные.

Решение. Здесь испытание – выбор пяти билетов из десяти. Всего существует C_{10}^5 способов выбрать 5 билетов из 10. Следовательно, число n всех возможных исходов испытания и в задаче а), и в задаче б) одинаково и равно C_{10}^5. Причём все эти исходы равновозможны.

В задаче а) благоприятный исход испытания состоит в том, что из двух выигрышных билетов будет выбран один (это можно сделать двумя способами), а из восьми невыигрышных будут выбраны четыре (это можно сделать C_8^4 способами). Таким образом, общее число всех благоприятных исходов равно $2\cdot C_8^4$, а значит, вероятность события A, состоящего в том, что среди пяти отобранных билетов окажется лишь один выигрышный, по классической формуле (1.3) равна:

$$p(A) = \frac{m}{n} = \frac{2\cdot C_8^4}{C_{10}^5} = \frac{5}{9}$$

В задаче б) благоприятный исход испытания состоит в том, что из двух выигрышных билетов будут взяты оба (их можно взять одним способом) и ещё

три билета будут взяты невыигрышных (их можно взять C_8^3 способами). Число благоприятных исходов, таким образом равно C_8^3, а вероятность события *B*, состоящего в том, что среди пяти отобранных билетов окажется два выигрышных, по классической формуле равна

$$p(B) = \frac{m}{n} = \frac{C_8^3}{C_{10}^5} = \frac{2}{9}$$

Пример 12. На книжную полку в произвольном порядке выставлены 5 книг. Какова вероятность того, что некоторые две из них, составляющие двухтомник, окажутся на полке рядом?

Решение. В данной задаче испытание – это установка на полку в произвольном порядке пяти книг. А событие *A* – то, что книги двухтомника окажутся рядом.

Всех возможных исходов испытания, очевидно, столько, сколько существует перестановок из пяти книг. То есть их $P_5 = 5! = 120$. Благоприятствующими событию *A* будут те из них, когда книги двухтомника стоят рядом.

Для начала подсчитаем число тех благоприятствующих исходов, когда книги двухтомника стоят на первых двух местах, причём первый том стоит на первом месте, а второй на втором. Так как последние три книги могут быть установлены произвольно, то таких исходов будет $P_3 = 3! = 6$. Меняя местами книги двухтомника, получим ещё 6 благоприятствующих исходов. Таким образом, если книги двухтомника занимают на полке первые два места, то всего соответствующих благоприятствующих исходов (перестановок книг) оказывается $2 \cdot P_3 = 12$. Но благоприятствующими событию *A* исходами будут и те, при которых книги двухтомника стоят на 2-3, 3-4, 4-5 местах. Итого всех таких исходов оказывается $12 \cdot 4 = 48$. А тогда по классической формуле (1.3) получаем:

$$p(A) = \frac{m}{n} = \frac{48}{120} = 0{,}4$$

Упражнения.

1. В пятом классе изучается 12 предметов. Сколькими способами можно составить расписание занятий на понедельник, если в этот день должно быть 4 урока? Решить задачу в предположении, что:

а) порядок уроков важен;

б) порядок уроков не важен.

Ответ: а) 11880; б) 495.

2. Сколько различных символов можно закодировать с помощью 8-значных двоичных чисел (чисел, содержащих в своей записи лишь нули и единицы общим числом 8)?

Ответ: 256.

3. Пять девушек и пять юношей разыграли 10 мест выделенного им на концерт зрительного ряда (места с 31 по 40). Какова вероятность того, что они

будут сидеть строго вперемешку (никакие две девушки и два юноши не будут сидеть рядом)?

Ответ: 1/126.

4. Найти вероятность того, что в лотерее «Спортлото 5 из 36» можно угадать:

а) все 5 номеров; б) 4 номера; в) 3 номера.

Ответ: а) 1/376992; б) 155/376992; в) 4650/376992.

5. Доказать, что вероятность того, что у двенадцати случайно выбранных человек дни рождения приходятся на разные месяцы, меньше 0,0001.

6. Сколькими различными способами можно распределить 6 различных учебников между тремя студентами по два учебника каждому?

Ответ: 90 способами.

§3. Классификация событий. Сумма и произведение событий.

Определение 1. *Два случайных события A и B называются несовместными, если появление одного из них исключает появление другого. В противном случае событие A и B называются совместными.*

Пример 1. Попадание A и непопадание B стрелка в мишень при одном выстреле – события несовместные.

Пример 2. Попадание A стрелка в мишень в первом выстреле и промах B во втором – события совместныые.

Определение 2. *Два случайных события A и B называются независимыми, если возможность (вероятность) появления каждого из них не зависит от того, появится или не появится другое событие. В противном случае события A и B называются зависимыми.*

Отметим сразу, что *несовместные события всегда зависимы*. Действительно, пусть A и B – несовместные события. Тогда при появлении события A появление события B невозможно, а значит $p(B) = 0$. А при непоявлении события A появление события B возможно, и тогда $p(B) \neq 0$. Таким образом, вероятность (возможность) появления события B зависит от того, появится или не появится событие A. Следовательно, несовместные события A и B действительно являются зависимыми.

А вот если события A и B совместные, то они могут быть *как зависимыми, так и независимыми*. Подтвердим это на примере.

Пример 3. Пусть из колоды игральных карт (36 карт) вынимают наудачу одну за одной две карты. И пусть событие A состоит в том, что первая вынутая карта окажется тузом, а событие B – что и вторая карта окажется тузом. Рассмотрим два варианта испытания:

а) первая вынутая карта возвращается в колоду;

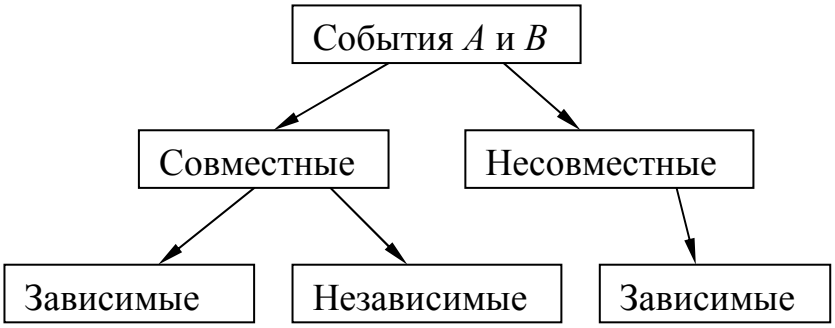
б) первая вынутая карта не возвращается в колоду.

Очевидно, что в обоих вариантах события A и B совместны. Исследуем их на зависимость –независимость.

а) В этом варианте вероятность появления события B (вынимания второго туза) не зависит, очевидно, от того, произошло или не произошло перед этим событие A (вынимание первого туза) и равна $p(B)=4/36=1/9$. Следовательно, в варианте а) имеем *совместные и независимые события A и B.*

б) В этом варианте, если событие A произошло, то $p(B)=3/35$ (останется 35 карт, и из них три туза). А если событие A не произошло (первая вынутая карта тузом не оказалась), то $p(B)=4/35$ (останется 35 карт, и из них четыре туза) Следовательно, в варианте б) имеем *совместные и зависимые события A и B.*

Связь между совместностью-несовместностью и зависимостью - независимостью двух событий можно изобразить графически в виде схемы (3.1):



(3.1)

Понятия совместности-несовместности и зависимости-независимости обобщаются и на случай группы из нескольких случайных событий ($A_1; A_2; \ldots A_n$)

Определение 3. *События ($A_1; A_2; \ldots A_n$) называются попарно несовместными, если появление каждого из них исключает появление любого другого (то есть если любая пара из группы событий ($A_1; A_2; \ldots A_n$) несовместна). В противном случае события ($A_1; A_2; \ldots A_n$) называются совместными.*

Определение 4. *События ($A_1; A_2; \ldots A_n$) называются независимыми в совокупности, если каждое из этих событий и любое другое событие, состоящее в появлении и непоявлении остальных событий, являются событиями независимыми. В противном случае события ($A_1; A_2; \ldots A_n$) называются зависимыми.*

Пример 4. Пусть из коробки, содержащей три шара (синий, красный, зелёный) наудачу по одному вынимают три шара, каждый раз опуская вынутый шар в коробку. И пусть ($A; B; C$) – события, состоящие в вынимании последовательно синего, красного и зелёного шара. Очевидно, что эти три события – *совместные и независимые в совокупности.*

Если же вынутые шары в коробку не возвращать, то события ($A; B; C$) – *совместные и зависимые.*

Сумма и произведение событий

Определение 5. *Суммой случайных событий ($A_1; A_2; \ldots A_n$) называется событие B*

$$A_1 + A_2 + \ldots A_n = B, \qquad (3.2)$$

состоящее в появлении хотя бы одного из складываемых событий (или A_1, или A_2, … или A_n).

Таким образом, в теории вероятности знак сложения (+) означает союз «или».

Определение 6. *Произведением случайных событий (A_1; A_2; …A_n) называется событие C*

$$A_1 \cdot A_2 \cdots A_n = C, \qquad (3.3)$$

состоящее в появлении всех перемножаемых событий (и A_1, и A_2, … и A_n).

Таким образом, в теории вероятности знак умножения ($\cdot$) означает союз «и».

Пример 5. Пусть (A_1; A_2; A_3) – события, состоящее в попадании в мишень первого, второго и третьего стрелков соответственно. Тогда $B = A_1 + A_2 + A_3$ – событие, состоящее в попадании в мишень *хотя бы одного* из стрелков, а $C = A_1 \cdot A_2 \cdot A_3$ – событие, состоящее в попадании в мишень *всех трёх* стрелков.

Суммы и произведения событий обладают следующими очевидными *свойствами:*

1. $A+B=B+A$ – переместительный закон сложения.
2. $(A+B)+C=A+(B+C)$ – сочетательный закон сложения.
3. $A \cdot B=B \cdot A$ – переместительный закон умножения. (3.4)
4. $(A \cdot B) \cdot C=A \cdot (B \cdot C)$ – сочетательный закон умножения.
5. $(A+B) \cdot C=A \cdot C+B \cdot C$ – распределительный закон сложения и умножения.

Свойства (3.4) для случайных событий полностью аналогичны соответствующим свойствам для чисел. Но есть и отличные свойства:

1) $A+A=A$; 2) $A \cdot A=A$; 3) $(A+B) \cdot A=A$; 4) $(AB) \cdot A=AB$. (3.5)

(продумайте эти свойства самостоятельно).

Определение 7. *Символом $\bar{A}$ обозначается событие, противоположное событию A. То есть $\bar{A}$ – это непоявление события A.*

Пример 6. Если событие A – попадание стрелка в мишень, то событие $\bar{A}$ – его промах по мишени. Если A – выпадение орла при бросании монеты, то $\bar{A}$ – выпадение решки. Если A– выпадение пятёрки при бросании игральной кости, то $\bar{A}$ – выпадение любой другой цифры. Если событие A – попадание брошенной точки на одноимённую область A (рис.1.1), то событие $\bar{A}$ – непопадание в эту область. И так далее. Очевидно, что

$$p(\bar{A}) = 1 - p(A), \text{ откуда } p(A) = 1 - p(\bar{A}) \qquad (3.6)$$

Последнее равенство (3.6) является ещё одной формулой (в дополнение к формулам (1.1)-(1.4)) для подсчета вероятностей случайных событий. Ею удобно пользоваться в тех случаях, когда вероятность $p(\bar{A})$ противоположного события $\bar{A}$ найти проще, чем вероятность $p(A)$ самого события A.

Пример 7. Бросаются две игральные кости. Найти вероятность того, что хотя бы на одной из них не выпадет шестёрка.

Решение. В данной задаче испытание – это бросание двух костей, а событие A – это невыпадение шестёрки хотя бы на одной из них. Событию A благоприятствуют, очевидно, много различных исходов испытания (например, 2-5; 3-6; 6-1, и т.д). И лишь один исход (6-6) благоприятствует событию $\bar{A}$, противоположному событию A (событие $\bar{A}$ – это выпадение двух шестёрок). Так как всех возможных исходов испытания, очевидно, 36, то по классической формуле (1.3)

$$p(\bar{A}) = 1/36.$$

Значит, по формуле (3.6) получаем:

$$p(A) = 1 - p(\bar{A}) = 1 - 1/36 = 35/36.$$

Событие A (любое) и ему противоположное событие $\bar{A}$ обладают следующими очевидными свойствами:

1) $A+\bar{A}$ – достоверное событие; $p(A+\bar{A}) = 1$.

2) $A\cdot\bar{A}$ – невозможное событие; $p(A\cdot\bar{A}) = 0$. (3.7)

Упражнения

1. Бросаются две игральные кости. Какова вероятность того, что:
а) Сумма выпавших очков будет больше 10?
б) Сумма выпавших очков будет не больше 10?
Ответ: а) 1/12; б) 11/12.

2. Бросаются две игральные кости. Какова вероятность того, что
а) Произведение выпавших очков будет чётным?
б) Произведение выпавших очков будет нечётным?
Ответ: а) 3/4; б) 1/4.

3. Какова вероятность того, что первые три встречные прохожие:
а) Родились в один месяц?
б) Родились в разные месяцы?
в) Родились летом?
г) Хотя бы один из них родился летом?
д) Никто из них летом не родился?
Ответ: а) 1/144; б) 55/72; в) 1/64; г) 37/64; д) 27/64.

§4. Формулы сложения и умножения вероятностей.

Формулы сложения и умножения вероятностей – это формулы для нахождения вероятностей сумм и произведений событий. Наиболее просто и наглядно устанавливаются эти формулы с помощью геометрической интерпретации вероятности случайного события, выраженной формулой (1.4) и рисунком (1.1).

1. Формула сложения вероятностей двух несовместных событий.

Пусть A и B – любые несовместные события. Геометрически их можно интерпретировать (изобразить) как попадание брошенной наудачу точки на непересекающиеся одноимённые области A и B соответственно (рис. 1.2):

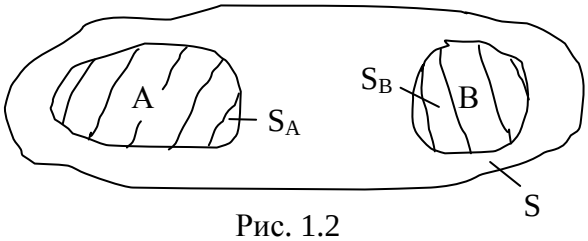


Рис. 1.2

Пусть S – площадь всей пластинки, на которую наудачу бросается точка, а S_A и S_B – площади областей A и B соответственно. Рассмотрим сумму $C = A+B$ событий A и B. По определению суммы событий, событие C состоится в появлении *хотя бы одного* из складываемых событий – или A, или B. То

есть событие $C = A+B$ состоит в попадании брошенной точки или в область A, или в область B. Следовательно, в соответствии с формулой (1.4) получаем:

$$p(C) = p(A+B) = (S_A + S_B)/S$$

Но

$$(S_A + S_B)/S = (S_A/S)+(S_B/S) = p(A)+p(B).$$

Таким образом, если A и B – несовместные события, то

$$p(A+B) = p(A) + p(B) \qquad (4.1)$$

2. Формула сложения вероятностей двух совместных событий.

Пусть A и B – любые два совместные события. Геометрически их можно интерпретировать в виде пересекающихся областей A и B соответственно (рис. 1.3). Заметим, что попадание бросаемой точки в общую часть AB областей A и B означает появление и события A, и события B, то есть представляет собой произведение AB событий A и B. Именно поэтому общую часть областей A и B мы обозначили символом AB.

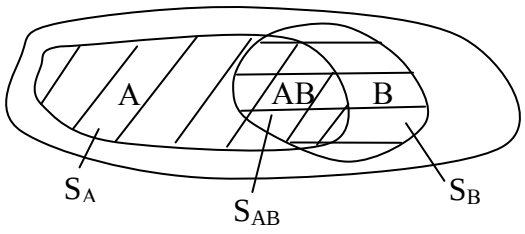


Рис. 1.3

Рассмотрим и здесь сумму $C = A+B$ событий A и B. Согласно рис. 1.3 вероятность появления события C равна отношению суммарной площади, занимаемой областями A и B, к площади S всей пластинки. Но указанная суммарная площадь равна, очевидно, $S_A + S_B - S_{AB}$ (продумайте это!). Поэтому

$$p(C) = p(A+B) = (S_A + S_B - S_{AB})/S = S_A/S + S_B/S - S_{AB}/S = p(A)+p(B)-p(AB).$$

Таким образом, если A и B – совместные события, то

$$p(A+B) = p(A)+p(B)-p(AB) \qquad (4.2)$$

Итак, получаем следующие формулы сложения вероятностей для двух произвольных случайных событий A и B:

$$\begin{aligned} &p(A+B) = p(A) + p(B) && \text{- если } A \text{ и } B \text{ несовместны} \\ &p(A+B) = p(A)+p(B)-p(AB) && \text{- если } A \text{ и } B \text{ совместны} \end{aligned} \qquad (4.3)$$

3. Формула умножения вероятностей двух зависимых событий.

Рассмотрим опять рис. 1.3. События A и B, изображённые на нём – совместные и при этом зависимые. Действительно, совершенно очевидно, что вероятность появления события B зависит от того, произошло или не произошло событие A (попала или не попала брошенная точка в область A).

Подтвердим это. Для этого введём следующие обозначения:

$$\begin{aligned} &p(B/A) \text{ – вероятность появления события } B \text{ при условии,} \\ &\quad \text{что } \textit{произошло событие } A. \\ &p(B/\overline{A}) \text{ – вероятность появления события } B \text{ при условии,} \\ &\quad \text{что } \textit{не произошло событие } A. \end{aligned} \qquad (4.4)$$

Введенные вероятности $p(B/A)$ и $p(B/\overline{A})$ называются *условными вероятностями события B*. Эти вероятности будут одинаковыми, если события A и B

независимы (друг от друга), и неодинаковыми, если события A и B друг от друга зависимы. Кстати совершенно аналогично понимаются условные вероятности $p(A/B)$ и $p(A/\overline{B})$ события A.

Согласно рис. 1.3,

$$p(B/A)=\frac{S_{AB}}{S_A}\ ;\qquad p(B/\overline{A})=\frac{S_B-S_{AB}}{S-S_A}.$$

Таким образом, $p(B/A)\neq p(B/\overline{A})$. А значит, события A и B зависимые.

А теперь рассмотрим событие $D=AB$. Согласно рис. 1.3

$$p(D)=p(AB)=\frac{S_{AB}}{S}.$$

Но

$$\frac{S_{AB}}{S}=\frac{S_{AB}}{S_A}\cdot\frac{S_A}{S}=p(B/A)\cdot p(A).$$

Таким образом, если события A и B зависимые, то

$$p(AB)=p(A)\cdot p(B/A) \qquad (4.5)$$

Кстати, так как $p(AB)=p(BA)$, то меняя в равенстве (4.5) A и B местами, получим еще одну формулу:

$$p(AB)=p(B)\cdot p(A/B) \qquad (4.6)$$

4. Формула умножения вероятностей двух независимых событий.

Если события A и B независимы, то

$$p(B/A)=p(B/\overline{A})=p(B);\qquad p(A/B)=p(A/\overline{B})=p(A).$$

И тогда на основании и формулы (4.5), и формулы (4.6) для независимых событий A и B получим:

$$p(AB)=p(A)\cdot p(B) \qquad (4.7)$$

Таким образом, получаем следующие формулы умножения вероятностей для произвольных случайных событий A и B:

$p(AB)=p(A)\cdot p(B)$ - если события A и B независимые (4.8)

$p(AB)=p(A)\cdot p(B/A)=p(B)\cdot p(A/B)$ - если события A и B зависимые

Формулы (4.3) и (4.8) полностью исчерпывают вопрос о формулах сложения и умножения вероятностей для двух случайных событий. Как следует из этих формул, при нахождении вероятности $p(A+B)$ суммы событий важно знать, совместны или несовместны складываемые события A и B. А при нахождении вероятности $p(AB)$ произведения событий важно знать, зависимы или независимы перемножаемые события A и B.

5. Формулы сложения вероятностей для произвольного числа любых событий

Полученные выше формулы сложения вероятностей для двух случайных событий можно обобщить и на совокупность из нескольких событий.

Пусть, например, (A; B; C) – любые три попарно несовместные события. Тогда

$$p(A+B+C)=p[(A+B)+C]=$$

=|учтем, что события $A+B$ и C несовместные|=

$= p(A+B)+p(C)$=|учтем, что события A и B несовместные|=

$= p(A)+p(B)+p(C).$ (4.9)

И вообще, если имеется n попарно несовместных событий $(A_1; A_2; ....A_n)$, то

$$p(A_1 + A_2 + ... + A_n) = p(A_1) + p(A_2) + ... + p(A_n) \quad (4.10)$$

Если же складываемые события совместны, то подсчет вероятности их суммы сильно усложняется. Например, для трех совместных событий (*A; B; C*) получаем:

$p(A+B+C) = p[(A+B)+C] =$

=|учтем, что события $A+B$ и C совместные|=

$= p(A+B)+p(C)-p[(A+B)\cdot C] = p(A)+p(B)-p(AB)+p(C)-p(AC+BC) =$

=|учтем, что события AC и BC совместные|=

$= p(A)+p(B)-p(AB)+p(C)-[p(AC)+p(BC)-p(AC\cdot BC)] =$

=|учтем, что $AC \cdot BC = ABC$|=

$= p(A)+p(B)+p(C)-p(AB)-p(AC)-p(BC)+p(ABC)$ (4.11)

Для четырех и более совместных событий вероятность их суммы будет выглядеть еще сложнее. В связи с этим если

$$C = A_1 + A_2 + ... + A_n$$

- сумма n совместных событий, то вероятность $p(C) = p(A_1 + A_2 + ... + A_n)$ выгоднее искать через вероятность $p(\overline{C})$ противоположного события $\overline{C}$ по формуле $p(C) = 1 - p(\overline{C})$. А так как событие $\overline{C}$ состоит в непоявлении ни одного из событий $(A_1; A_2; ....A_n)$, то $\overline{C} = \overline{A}_1 \cdot \overline{A}_2 \cdots \overline{A}_n$. И тогда для совместных событий $(A_1; A_2; ....A_n)$ получаем:

$$p(A_1 + A_2 + ... + A_n) = 1 - p(\overline{A}_1 \cdot \overline{A}_2 \cdots \overline{A}_n) \quad (4.12)$$

Таким образом, учитывая (4.10) и (4.12), получаем следующие итоговые формулы для *вероятности суммы произвольного числа любых случайных событий:*

$p(A_1 + A_2 + ... + A_n) = p(A_1) + p(A_2) + ... + p(A_n)$ - если события $(A_1; A_2; ....A_n)$ попарно несовместны (4.13)

$p(A_1 + A_2 + ... + A_n) = 1 - p(\overline{A}_1 \cdot \overline{A}_2 \cdots \overline{A}_n)$ - если события $(A_1; A_2; ....A_n)$ совместны

6. Формулы умножения вероятностей для произвольного числа любых событий

Пусть $(A_1; A_2; ....A_n)$. - произвольные случайные события. Если эти события независимы в совокупности, то опираясь на формулу (4.7) для двух событий, получаем:

$p(A_1 \cdot A_2 \cdot A_3 \cdots A_n) = p[A_1 \cdot (A_2 \cdot A_3 \cdots A_n)] =$

$= \left|$ учтем, что события A_1 и $(A_2 \cdot A_3 \cdots A_n)$ независимы $\right| =$ (4.14)

$$= p(A_1)\cdot p[A_2\cdot(A_3\cdot\cdots A_n)] = (.....) = p(A_1)\cdot p(A_2)\cdot\cdots p(A_n)$$

Итак, если события $(A_1; A_2; ....A_n)$. независимы в совокупности, то

$$p(A_1\cdot A_2\cdot A_3\cdot\cdots A_n) = p(A_1)\cdot p(A_2)\cdot\cdots p(A_n) \qquad (4.15)$$

Если же эти события зависимы, то опять реализуя схему (4.14) по последовательному "отщеплению" отдельных событий, но используя уже формулу (4.6), получим:

$$p(A_1\cdot A_2\cdot A_3\cdot\cdots A_n) = p(A_1)\cdot p(A_2/A)\cdot p(A_3/A_1A_2)\cdot\cdots p(A_n/A_1A_2A_3...A_{n-1}) \qquad (4.16)$$

Таким образом, учитывая (4.15) и (4.16), получаем следующие итоговые формулы для *вероятности произведения произвольного числа любых случайных событий:*

$p(A_1\cdot A_2\cdot A_3\cdot\cdots A_n) = p(A_1)\cdot p(A_2)\cdot\cdots p(A_n)$ - если события $(A_1; A_2; ....A_n)$ независимы в совокупности (4.17)

$p(A_1\cdot A_2\cdot A_3\cdot\cdots A_n) = p(A_1)\cdot p(A_2/A_1)\cdot p(A_3/A_1A_2)\cdot\cdots p(A_n/A_1A_2A_3...A_{n-1})$ - если события $(A_1; A_2; ....A_n)$ зависимы

А теперь рассмотрим примеры решения задач с применением формул сложения и умножения вероятностей.

Пример 1. Два стрелка по разу стреляют по мишени. Вероятность попадания в мишень для первого стрелка равна 0,7, а для второго 0,8. Найти вероятность того, что

а) в мишень попадут оба; б) оба промахнутся; в) попадет хотя бы один из них; г) попадет только один из них.

Решение. Введем обозначения:

событие A_1 – попадание первого стрелка в мишень;
событие A_2 – попадание второго стрелка в мишень;
событие A – попадание обоих;
событие B – промах обоих;
событие C – попадание хотя бы одного из них;
событие D – попадание только одного из них.

По условию задачи, $p(A_1) = 0{,}7;\ \ p(A_2) = 0{,}8.$

а) Найдем $p(A)$. Очевидно, что $A = A_1\cdot A_2$. Поэтому

$p(A) = p(A_1\cdot A_2) =$ | учтем, что события A_1 и A_2 независимы | =

$= p(A_1)\cdot p(A_2) = 0{,}7\cdot 0{,}8 = 0{,}56.$

б) Найдем $p(B)$. Учтем, что $B = \overline{A}_1\cdot\overline{A}_2$. Поэтому

$p(B) = p(\overline{A}_1\cdot\overline{A}_2) =$ |учтем, что события $\overline{A}_1$ и $\overline{A}_2$ независимы| =

$= p(\overline{A}_1)\cdot p(\overline{A}_2) = (1-p(A_1))\cdot(1-p(A_2)) = 0{,}3\cdot 0{,}2 = 0{,}06.$

в) Найдем $p(C)$. Учтем, что $C = A_1 + A_2$. Поэтому $p(C) = p(A_1 + A_2) =$

=|учтем, что события A_1 и A_2 совместны|=

$= p(A_1) + p(A_2) - p(A_1\cdot A_2) = 0{,}7 + 0{,}8 - 0{,}56 = 0{,}94$

Заметим, что можно было найти $p(C)$ и иначе, если учесть, что $\overline{C} = B$:

$$p(C) = 1 - p(\overline{C}) = 1 - p(B) = 1 - 0{,}06 = 0{,}94$$

г) Найдем $p(D)$. Учтем, что $D = A_1 \cdot \overline{A}_2 + \overline{A}_1 \cdot A_2$.. Поэтому

$$p(D) = p(A_1\overline{A}_2 + A_1\overline{A}_2) =$$

$$= |\text{учтем, что события } A_1 \cdot \overline{A}_2 \text{ и } \overline{A}_1 \cdot A_2 \text{ несовместны}| =$$

$$= p(A_1 \cdot \overline{A}_2) + p(\overline{A}_1 \cdot A_2) =$$

$$= \left| \text{учтем, что множители в обоих произведениях независимы} \right| =$$

$$= p(A_1) \cdot p(\overline{A}_2) + p(\overline{A}_1) \cdot p(A_2) = 0{,}7 \cdot 0{,}2 + 0{,}3 \cdot 0{,}8 = 0{,}38$$

Таким образом, из четырех рассмотренных событий (*A; B; C; D*) наиболее вероятным является событие *C* – попадание в мишень хотя бы одного из двух стрелков. Его вероятность равна 0,94. Это значит, что *в среднем* из каждых 100 парных выстрелов будет 94 таких, когда в мишень кто-либо из двух стрелков попадет. А наименее вероятным является событие *B* – промах обоих. Его вероятность равна 0,06. Это значит, что *в среднем* из каждых 100 парных выстрелов будет лишь 6 таких, когда оба стрелка промахнутся.

Пример 2. Из колоды в 36 карт наудачу последовательно вынимаются две карты. Какова вероятность того, что они обе окажутся тузами?

Решение. Введем обозначения:

событие A_1 – первая вынутая карта туз;

событие A_2 – вторая вынутая карта туз;

событие A – обе вынутые карты тузы.

Очевидно, что $A = A_1 \cdot A_2$. Поэтому

$$p(A) = p(A_1 \cdot A_2) = \left| \text{учтем, что события } A_1 \text{ и } A_2 \text{ зависимые} \right| =$$

$$p(A_1) \cdot p(A_2 / A_1) = \frac{4}{36} \cdot \frac{3}{35} = \frac{1}{105}$$

Примечание. В примере 9, §2 мы рассмотрели эту же задачу, но при условии, что обе карты вынимаются из колоды не по очереди, а сразу (парой). Вероятность того, что они обе окажутся тузами, оказалась той же: $\frac{1}{105}$. И это совершенно естественно, ибо выбирать какие-то объекты из какой-либо их совокупности по очереди или сразу – это одно и то же и с точки зрения теории вероятностей, и с точки зрения здравого смысла.

Пример 3. На связке 4 ключа. К замку подходит только один ключ. Найти вероятность того, что потребуется не менее трех попыток открыть замок, если опробованный ключ в дальнейших испытаниях не участвует.

Решение. Введем обозначения:

событие A_1 – подойдет первый опробованный ключ;

событие A_2 – подойдет второй ключ;

событие A_3 – подойдет третий ключ;

событие A_4 – подойдет четвертый ключ;

событие A – понадобится не менее трех попыток открыть замок.

Очевидно, что $A = A_3 + A_4$. Поэтому

$$p(A) = p(A_3 + A_4) = \left| \text{учтем, что события } A_3 \text{ и } A_4 \text{ несовместны} \right| =$$

$$= p(A_3) + p(A_4) = \left| \text{учтем, что } A_3 = \overline{A}_1 \cdot \overline{A}_2 \cdot A_3;\ \ A_4 = \overline{A}_1 \cdot \overline{A}_2 \cdot \overline{A}_3 \cdot A_4 \right| =$$

$$= p(\overline{A}_1 \cdot \overline{A}_2 \cdot A_3) + p(\overline{A}_1 \cdot \overline{A}_2 \cdot \overline{A}_3 \cdot A_4) =$$

$= \left|$ учтем, что множители – события зависимые $\right| =$

$$= p(\overline{A}_1)\cdot p(\overline{A}_2/\overline{A}_1)\cdot p(A_3/\overline{A}_1\overline{A}_2) + p(\overline{A}_1)\cdot p(\overline{A}_2/\overline{A}_1)\cdot p(\overline{A}_3/\overline{A}_1\overline{A}_2) + p(A_4/\overline{A}_1\overline{A}_2\overline{A}_3) =$$

$$= \frac{3}{4}\cdot\frac{2}{3}\cdot\frac{1}{2} + \frac{3}{4}\cdot\frac{2}{3}\cdot\frac{1}{2}\cdot 1 = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

Впрочем, эту задачу можно было бы решить и гораздо проще. Действительно, ни один из четырех ключей не имеет заведомых преимуществ по сравнению с другими ключами. Поэтому:

$$p(A_1) = p(A_2) = p(A_3) = p(A_4) = \frac{1}{4}.$$

А тогда

$$p(A) = p(A_3) + p(A_4) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

Упражнения

1. Бросаются три игральные кости. Найти вероятность того, что:
   а) На всех трех костях выпадет одинаковое число очков.
   б) Произведение выпавших очков будет четным.
   в) Сумма выпавших очков будет больше 4.

   Ответ: а) $\frac{1}{36}$; б) $\frac{7}{8}$; в) $\frac{53}{54}$.

2. Отрезок разделен на три равные части. На этот отрезок наудачу брошены три точки. Найти вероятность того, что на каждую часть отрезка попадает по одной точке.

   Ответ: $\frac{2}{9}$.

3. Студент знает 25 из 30 экзаменационных вопросов. Найти вероятность того, что из трех заданных вопросов студент знает: а) не менее двух вопросов; б) все три вопроса.

   Ответ: а) $\frac{190}{203}$; б) $\frac{115}{203}$.

4. Четыре человека делают по одной попытке при совершении некоторого действия. Средний процент удачных попыток для них соответственно равен: 30%; 40%; 50%; 60%. Найти вероятность того, что:
   а) Все четыре попытки окажутся удачными.
   б) Все четыре попытки окажутся неудачными.
   в) Хотя бы одна из попыток окажется удачной.

   Ответ: а) 0,036; б) 0,084; в) 0,916.

§5. Формула полной вероятности и формула Байеса.

Определение. *События* $(B_1; B_2; ....B_n)$ *образуют полную группу событий, если:*

а) они попарно несовместны;

б) одно из них при производстве испытания обязательно произойдет.

Пример 1. События $(B_1; B_2; ....B_6)$, представляющие собой выпадение единицы, двойки, ... шестерки при бросании игральной кости, составляют, очевидно, полную группу событий.

Пример 2. Любое случайное событие A и ему противоположное событие $\overline{A}$ составляют, очевидно, полную группу событий.

Если события $(B_1; B_2; ....B_n)$ составляют полную группу событий, то событие $B = B_1 + B_2 + ... + B_n$, состоящие в появлении хотя бы одного из событий $(B_1; B_2; ....B_n)$, представляет собой достоверное событие. Значит,

$$p(B) = p(B_1 + B_2 + ... + B_n) = 1 \qquad (5.1)$$

С другой стороны, в силу попарной несовместности событий $(B_1; B_2; ....B_n)$ по формуле (4.10) получаем:

$$p(B_1 + B_2 + ... + B_n) = p(B_1) + p(B_2) + ... + p(B_n) \qquad (5.2)$$

Сравнивая (5.1) и (5.2), получаем для полной группы событий $(B_1; B_2; ....B_n)$:

$$p(B_1) + p(B_2) + ... + p(B_n) = 1 \qquad (5.3)$$

В частности,

$$p(A) + p(\overline{A}) = 1, \text{ откуда } p(A) = 1 - p(\overline{A}) \qquad (5.4)$$

- формула, приводившаяся ранее (см. (3.6)).

Если $(B_1; B_2; ....B_n)$ - полная группа событий, то любое интересующее нас событие A может произойти лишь совместно с одним из событий этой группы. А значит,

$$A = B_1 A + B_2 A + ... + B_n A \qquad (5.5)$$

Пусть нам известны вероятности

$$p(B_1);\ p(B_2);....;\ p(B_n) \qquad (5.6)$$

всех событий, составляющих полную группу, а также известны условные вероятности

$$p(A/B_1);\ p(A/B_2);....;\ p(A/B_n) \qquad (5.7)$$

появления события A совместно с каждым из событий этой группы. Тогда из выражения (5.5) получаем:

$$p(A) = p(B_1 A + B_2 A + ... + B_n A) =$$

$= \left|\right.$ учтем, что все слагаемые – это попарно несовместные события $\left|\right. =$

$$= p(B_1 A) + p(B_2 A) + ... + p(B_n A) =$$

$= \left|\right.$ учтем, что множители во всех произведениях– зависимые события $\left|\right. =$

$$= p(B_1) \cdot p(A/B_1) + p(B_2) \cdot p(A/B_2) + ... + p(B_n) \cdot p(A/B_n) =$$

$$= \sum_{k=1}^{n} p(B_k) \cdot p(A/B_k) \qquad (5.8)$$

Полученная формула

$$p(A) = \sum_{k=1}^{n} p(B_k) \cdot p(A/B_k) \qquad (5.9)$$

называется *формулой полной вероятности.* В отличие от условных вероятностей (5.7), обеспечиваемых событию *A* отдельными событиями полной группы, формула полной вероятности определяет *итоговую (полную) вероятность* события *A*, которую обеспечивает ему вся полная группа событий *в целом*.

Примечание. Во многих практических задачах роль событий полной группы играют некоторые предположения (гипотезы). Поэтому события $(B_1; B_2; ....B_n)$ обычно называют *гипотезами*.

Пример 3. Имеются две урны (ящика) с шарами. В первой урне содержатся 2 белых и 1 черный шар, а во второй 1 белый и 2 черных шара. Из первой урны во вторую наудачу перенесли один шар, а затем из второй урны наудачу извлекли один шар. Какова вероятность того, что в итоге из второй урны извлечен белый шар?

Решение. Неизвестно, какой шар (белый или черный) перенесли из первой урны во вторую. Есть две гипотезы:

гипотеза B_1 – из первой урны во вторую перенесен белый шар;

гипотеза B_2 – из первой урны во вторую перенесен черный шар.

Гипотезы (B_1; B_2) образуют, очевидно, полную группу случайных событий. При этом

$$p(B_1) = \frac{2}{3}; \quad p\{B_2) = \frac{1}{3}.$$

Далее, пусть *A* – событие, состоящее в том, что из второй урны после перенесения в нее одного шара из первой урны извлечен белый шар. Очевидно, что

$$p(A/B_1) = \frac{2}{4}; \quad p(A/B_2) = \frac{1}{4}$$

А искомую вероятность $p(A)$ события *A* (полную вероятность) найдем по формуле полной вероятности:

$$p(A) = \sum_{k=1}^{2} p(B_k) \cdot p(A/B_k) = p(B_1) \cdot p(A/B_1) + p(B_2) \cdot p(A/B_2) = \frac{2}{3} \cdot \frac{2}{4} + \frac{1}{3} \cdot \frac{1}{4} = \frac{5}{12}$$

А теперь предположим, что испытание, в котором событие *A* могло произойти или не произойти, уже произведено, и событие *A* в этом испытании появилось. Встает естественный вопрос: какими теперь, в свете этого факта, будут вероятности справедливости гипотез $(B_1; B_2; ....B_n)$? То есть каковы вероятности

$$p(B_1/A); p(B_2/A); ....; p(B_n/A)? \qquad (5.10)$$

Эти вероятности находятся по *формуле Байеса*:

$$p(B_k/A) = \frac{p(B_k) \cdot p(A/B_k)}{p(A)} \qquad (k=1,2...n) \qquad (5.11)$$

Здесь $p(A)$ - полная вероятность события *A*, подсчитываемая по формуле (5.9).

Вывод формулы Байеса очень прост: используя формулу (4.8) для зависимых событий, получим:

$$p(A \cdot B_k) = p(A) \cdot p(B_k / A) = p(B_k) \cdot p(A / B_k) \ \ (k=1,2\ldots n)$$

Из последнего равенства и вытекает формула Байеса.

Пример 4. Пусть указанный в примере 3 эксперимент совершен, и в итоге из второй урны вынут белый шар (событие A произошло). Вопрос: каковы вероятности $p(B_1 / A)$ и $p(B_2 / A)$ того, что при этом из первой урны во вторую был перенесен белый шар и черный шар соответственно?

Решение. Искомые вероятности получим по формуле Байеса:

$$p(B_1 / A) = \frac{p(B_1) \cdot p(A / B_1)}{p(A)} = \frac{\frac{2}{3} \cdot \frac{2}{4}}{\frac{2}{12}} = \frac{4}{5}; \ p(B_2 / A) = \frac{p(B_2) \cdot p(A / B_2)}{p(A)} = \frac{\frac{1}{3} \cdot \frac{1}{4}}{\frac{2}{12}} = \frac{1}{5}$$

Сравнивая вероятности справедливости гипотез B_1 и B_2, найденные до опыта ($p(B_1) = \frac{2}{3}; \ \ p(B_2) = \frac{1}{3}$) с вероятностями справедливости этих же гипотез после опыта ($p(B_1 / A) = \frac{4}{5}; \ \ p(B_2 / A) = \frac{1}{5}$) видим, что вероятность справедливости первой гипотезы выросла, а вероятность справедливости второй гипотезы уменьшилась. И это естественно: результат испытания (извлечение из второй урны белого шара) в большей мере подтвердил справедливость гипотезы B_1 (во вторую урну перенесен белый шар), чем гипотезы B_2 (во вторую урну перенесен черный шар).

Заметим, что сумма вероятностей гипотез B_1 и B_2 как до опыта, так и после опыта равна 1. И это естественно, ибо и после опыта полная группа гипотез продолжает оставаться полной.

Упражнения:

1. Из коробки домино (в ней всего 28 костей) наудачу последовательно вынимаются две кости. Какова вероятность того, что вторую кость можно приставить к первой?

Ответ: $\frac{7}{18}$.

2. Из коробки домино наудачу последовательно извлечено две кости, причем вторая кость подошла (приставилась) к первой. Какова вероятность того, что первая кость: а) была дублем? б) не была дублем?

Ответ: а) $\frac{1}{7}$; б) $\frac{6}{7}$.

3. Из колоды в 36 карт наудачу последовательно вынимаются две карты. Какова вероятность того, что вторая из них окажется тузом?

Ответ: $\frac{1}{9}$.

4. Студент знает не все экзаменационные билеты. Когдаа вероятность вытащить на экзамене билет, который он знает, у студента больше: когда он идет брать билет первым? вторым? …последним?

Ответ: Все вероятности одинаковы.

5. Имеются три одинаковых запечатанных ящика с деталями. В первом ящике все детали стандартные; во втором половина стандартных и половина нестандартных; в третьем – все детали нестандартные. Наудачу вскрыт один ящик, и вынутая из него деталь оказалась стандартной. Каковы вероятности того, что был вскрыт: а) первый ящик; б) второй; в) третий.

Ответ: а) $\frac{2}{3}$; б) $\frac{1}{3}$; в) 0.

§6. Повторение испытаний. Формулы Бернулли, Лапласа, Пуассона.

Пусть A – случайное событие, вероятности появления и не появления которого для одного испытания известны:

$$p(A) = p\,;\quad p(\overline{A}) = 1 - p(A) = 1 - p = q\,;\ (p + q = 1) \tag{6.1}$$

И пусть производится не одно, а n повторных испытаний (или, что одно и то же, испытание повторяется n раз). Возникает естественный вопрос: какова вероятность того, что событие A в этих n повторных испытаниях появится k раз (целое число k можно задавать любым в пределах от 0 до n)? При этом не важно, в каком порядке событие A появится k раз в n испытаниях. Важно лишь общее число k появлений этого события. Эту вероятность обозначают символом $P_n(k)$ ($P_n(k)$ - вероятность того, что в n испытаниях событие A наступит k раз). И находится она по формуле Бернулли (Яков Бернулли – швейцарский математик 17-го века):

$$P_n(k) = C_n^k p^k q^{n-k} = \frac{n!}{k!(n-k)!} p^k q^{n-k} \tag{6.2}$$

Доказательство. Если в n повторных испытаниях событие A появится k раз, то соответственно оно не появится n-k раз. И тогда вероятность любой конкретной комбинации k появлений события A и n-k непоявлений этого события можно найти по формуле (4.15) произведения вероятностей независимых в совокупности событий. То есть она равна $p^k q^{n-k}$. Таких конкретных комбинаций будет, очевидно, столько, сколько существует сочетаний из n элементов (номеров испытаний) по k элементов в каждом сочетании. Эти сочетания образуются из k номеров тех испытаний, в которых будет появляться событие А. Каждому такому сочетанию k номеров будет соответствовать единственное сочетание тех n-k номеров испытания, в которых событие A не будет появляться. Так как все-

го таких сочетаний C_n^k, и каждое из них несовместно с любым другим сочетанием, то по формуле (4.10) сложения вероятностей попарно несовместных событий искомая вероятность равна величине $p^k q^{n-k}$, взятой C_n^k раз. В итоге и приходим к формуле Бернулли (6.2).

Пример 1. Монету подбрасывают пять раз подряд. Какова вероятность того, что при этом герб выпадет ровно три раза?

Решение. Будем считать испытанием однократное подбрасывание монеты. Тогда $n=5$ – число повторных испытаний. Далее, будем считать событием A в каждом испытании (при каждом бросании монеты) выпадение герба. Тогда

$$p(A)=p=\frac{1}{2};\ p(\overline{A})=q=1-p=\frac{1}{2};\ n=5;\ k=3;\ \ P_5(3)=?$$

На основании формулы Бернулли получаем:

$$P_5(3)=\frac{5!}{3!\cdot 2!}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^2=10\cdot\left(\frac{1}{2}\right)^5=\frac{5}{16}.$$

Формула Бернулли – точная формула. Однако при больших значениях n (большом числе испытаний) вычисления по ней становятся громоздкими из-за необходимости вычисления факториалов больших чисел и степеней с большими показателями. В процессе этих вычислений неизбежно придется производить округления, что приведет к погрешности при определении искомой вероятности $P_n(k)$. Причем к погрешности тем большей, чем больше будет значение n (числа испытаний). В связи с этим из формулы Бернулли выведены упрощенные приближенные формулы для $P_n(k)$, которые, кстати, тем точнее, чем больше число n.

Одна из этих приближенных формул – *локальная формула Лапласа:*

$$p_n(k)\approx\frac{1}{\sqrt{npq}}\cdot\varphi(x),\ \text{где}\ \ \varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}};\ \ x=\frac{k-np}{\sqrt{npq}} \qquad (6.3)$$

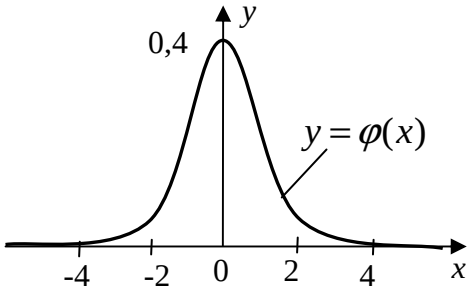


Рис. 1.4

Функция $\varphi(x)$ называется *функцией вероятностей* (или *функцией Гаусса*). График функции $y=\varphi(x)$ имеет вид, изображенный на рис. 1.4.

Для $0\le x\le 5$ значение функции $y=\varphi(x)$ можно взять в таблице, содержащейся во многих пособиях по теории вероятностей. Приведем небольшой фрагмент этой таблицы:

x	0	1	2	3	4	5
$\varphi(x)$	0,3989	0,2420	0,0540	0,0044	0,0001	≈ 0

Для отрицательных значений x используют четность функции $\varphi(x)$: $\varphi(-x)=\varphi(x)$. Для $x<-5$ и $x>5$ значение $\varphi(x)\approx 0$. На практике локальную формулу Лапласа применяют, если

$$npq > 10. \qquad (6.5)$$

Это условие обеспечивает приближенное нахождение вероятности $P_n(k)$ с точностью до процента. Доказательство локальной формулы Лапласа опускаем.

Пример 2. Монету подбрасывают 100 раз подряд. Какова вероятность того, что при этом герб выпадет ровно 40 раз?

Решение. Будем считать испытанием однократное подбрасывание монеты. Тогда n=100 – число повторных испытаний. Будем считать событием A в каждом испытании (при каждом бросании монеты) выпадение герба. Тогда

$$p(A) = p = \frac{1}{2}; \quad p(\bar{A}) = q = 1 - p = \frac{1}{2}; \quad n = 100; \quad k = 40; \quad P_{100}(40)\,?$$

Формулу Бернулли для подсчета искомой вероятности $P_{100}(40)$ применять не будем – слишком велико число испытаний n (n=100). А так как $npq = 25 \succ 10$, то вместо формулы Бернулли (6.2) применим локальную формулу Лапласа (6.3):

$$P_{100}(40) \approx \frac{1}{\sqrt{25}} \cdot \varphi(x) = \left| x = \frac{k - np}{\sqrt{npq}} = -2 \right| = \frac{1}{5} \cdot \varphi(-2) =$$

$$= \left| \text{ учтем, что } \varphi(-2) = \varphi(2) = 0{,}0540 \right| = 0{,}0108$$

Другой приближенной формулой для подсчета вероятностей $P_n(k)$, применяемой при больших n, является *формула Пуассона (формула редких событий):*

$$P_n(k) \approx \frac{\lambda^k}{k!} \cdot e^{-\lambda}, \text{ где } \lambda = np \qquad (6.6)$$

Она применяется, когда n велико (условно $n \geq 50$), а p мало ($0<p<0{,}1$), и когда $npq<10$. То есть когда не оправдано ни применение формулы Бернулли, ни применение локальной формулы Лапласа. При этих условиях приближенная формула Пуассона, как и локальная формула Лапласа, обеспечивает определение искомой вероятности $P_n(k)$ с погрешностью в пределах одного процента.

Кстати, так как вероятность $p(A) = p$ события A мала ($0<p<0{,}1$), то при повторении испытаний событие A наступает редко. Поэтому формула Пуассона и называется формулой редких событий. Вывод этой формулы, как и локальной формулы Лапласа, опустим.

Пример 3. Производится 50 повторных испытаний, причем вероятность появления некоторого события A в каждом из них равна 0,98. Определить вероятность того, что событие A наступит во всех 50 испытаниях.

Решение. В данной задаче

$p(A)=p=0{,}98$; $p(\bar{A})=q=0{,}02$; $n=50$; $k=50$; $P_{50}(50) = ?$

Если применить формулу Бернулли, то получим результат, который очевиден и без формулы Бернулли:

$$P_{50}(50) = (0{,}98)^{50}$$

Попробуем избежать громоздкой процедуры возведения числа 0,98 в 50-ую степень (её, впрочем, можно и избежать, если использовать логарифмы). То есть заменим формулу Бернулли на локальную формулу Лапласа или Пуассона.

Так как npq=50·0,98·0,02=0,98<10, то локальную формулу Лапласа применять нельзя - мы получим слишком грубый (неточный) результат. Но и формулу Пуассона (формулу редких событий) мы тоже применить не можем, так как вероятность p не мала, а наоборот, велика. Но зато мала вероятность q непоявления этого события. В связи с этим переформулируем задачу: найдем вероятность $P_{50}(0)$ того, что событие $\bar{A}$ появится 0 раз (ни разу). Эта вероятность, очевидно, совпадает с искомой вероятностью $P_{50}(50)$ того, что событие A появится во всех 50 испытаниях. Тогда в этой постановке получаем:

$$p(\bar{A})=p=0{,}02;\ p(A)=q=0{,}98;\ n=50;\ k=0;\ P_{50}(0)\ ?$$

Применяя формулу Пуассона (теперь ее применять можно), получим:

$$P_{50}(0)\approx\frac{\lambda^0}{0!}e^{-\lambda}=|\lambda=np=50\cdot0{,}02=1|=e^{-1}=\frac{1}{e}\approx\frac{1}{2{,}72}\approx 0{,}37.$$

Рассмотрим теперь следующую задачу: какова вероятность $P_n(k_1\le k\le k_2)$ того, что в n повторных испытаниях число k появлений события A окажется в заданных числовых пределах $[k_1; k_2]$?

Решение этой задачи очевидно:

$$P_n(k_1\le k\le k_2)=\sum_{k=k_1}^{k_2}P_n(k) \qquad (6.7)$$

Действительно, число k окажется в пределах $[k_1; k_2]$, если оно будет равно или k_1, или k_1+1, или k_1+2, … или k_2. События, приводящие к таким значениям k, друг с другом (попарно) несовместны. А тогда по формуле (4.10) сложения вероятностей попарно несовместных событий мы и приходим к формуле (6.7).

Пример 4. Монета бросается 5 раз. Какова вероятность того, что при этом герб выпадет не более четырех раз?

Решение. Будем считать испытанием однократное бросание монеты. Тогда n=5-число повторных испытаний. А событием A в каждом испытании будем считать выпадение герба. Тогда:

$$p(A)=p=\frac{1}{2};\ p(\bar{A})=q=\frac{1}{2};\ n=5;\ 0\le k\le 4;\quad P_5(0\le k\le 4)=?$$

Применяя формулу (6.7), получим:

$$P_5(0\le k\le 4)=\sum_{k=0}^{4}P_5(k)=P_5(0)+P_5(1)+P_5(2)+P_5(3)+P_5(4)$$

Для подсчета каждого из этих пяти слагаемых следует, очевидно, применить формулу Бернулли (6.2) и полученные числа сложить.

Однако эту задачу можно решить и гораздо проще - через противоположное событие:

$$P_5(0\le k\le 4)=1-P_5(5)=1-\frac{1}{32}=\frac{31}{32}$$

Если число n испытаний велико, границы $[k_1; k_2]$ широкие и если, кроме того, $npq>10$, то вместо точной формулы (6.7), использование которой становится громоздким, используют *приближенную интегральную формулу Лапласа:*

$$P_n(k_1 \le k \le k_2) \approx \int_{x_1}^{x_2} \varphi(x)dx = \Phi(x_2) - \Phi(x_1) \qquad (6.8)$$

Здесь:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}; \quad \Phi(x) = \int_0^x \varphi(t)dt = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt;$$

$$x_1 = \frac{k_1 - np}{\sqrt{npq}}; \quad x_2 = \frac{k_2 - np}{\sqrt{npq}}; \qquad (6.9)$$

Функция $\varphi(x)$ - это уже известная нам функция Гаусса (её график изображен на рис. 1.4), а функция $\Phi(x)$ называется *интегралом вероятностей*. График этой функции изображен на рисунке 1.5. В любом пособии по теории вероятностей имеются таблицы значений интеграла вероятностей $\Phi(x)$, который, как и функция Гаусса, принадлежит к числу важнейших функций теории вероятностей. Эти таблицы составлены для $0 \le x \le 5$. Приведем небольшой фрагмент этой таблицы:

x	0	1	2	3	4	5
$\Phi(x)$	0	0,3413	0,4772	0,49865	0,499968	0,499997

Для $x>5$ можно считать $\Phi(x)=0{,}5$. Для $x<0$ следует использовать нечетность функции $\Phi(x)$: $\Phi(-x) = -\Phi(x)$.

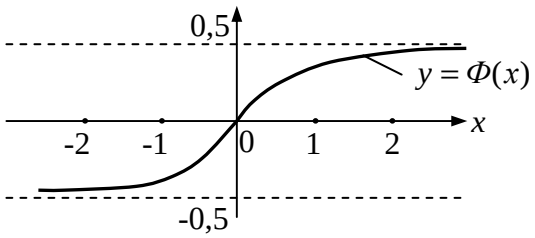


Рис. 1.5

Пример 5. Монету подбрасывают 100 раз подряд. Какова вероятность того, что при этом герб выпадет от 40 до 60 раз?

Решение. Будем считать испытанием однократное подбрасывание монеты. Тогда $n=100$ - число повторных испытаний. Событием A в каждом испытании будем считать выпадение герба. Тогда:

$$p(A)=p=\frac{1}{2};\ p(\bar{A})=q=\frac{1}{2};\ n=100;\ 40 \le k \le 60;\ P_{100}(40 \le k \le 60) = ?$$

Так как $npq=25>10$, то применим интегральную формулу Лапласа:

$$P_{100}(40 \le k \le 60) \approx \Phi(x_2) - \Phi(x_1);$$

$$x_1 = \frac{k_1 - np}{\sqrt{npq}} = \frac{40 - 100 \cdot 0{,}5}{\sqrt{25}} = -2; \quad x_2 = \frac{k_2 - np}{\sqrt{npq}} = \frac{60 - 100 \cdot 0.5}{\sqrt{25}} = 2;$$

$$P_{100}(40 \le k \le 60) \approx \Phi(2) - \Phi(-2) = \left| \text{учтем, что } \Phi(-2) = -\Phi(2) \right| = 2\Phi(2) =$$

$$= \left| \Phi(2) = 0{,}4772 \right| = 0{,}9544$$

Упражнения

1. Показать, что наиболее вероятным числом появления события A в n повторных испытаниях является число $k_{наив.}=np$ (или ближайшее к np целое число, если np не целое).

2. Два равносильных игрока играют в игру, ничьи в которой исключаются. Какова вероятность для первого игрока выиграть: а) одну партию из двух? б) две из четырех? в) три из шести?

Ответ: а) $\frac{1}{2}$; б) $\frac{3}{8}$; в) $\frac{5}{16}$

3. Отрезок АВ разделен точкой С в отношении 2:1. На этот отрезок наудачу брошены четыре точки. Найти вероятность того, что две из них окажутся левее точки С, а две - правее.

Ответ: $\frac{8}{27}$

4.Вероятность рождения мальчика равна 0,515. Найти вероятность того, что среди 100 новорожденных мальчиков и девочек окажется поровну.

Ответ: 0,07

5. Магазин получил 500 бутылок в стеклянной таре. Вероятность того, что при перевозке любая из бутылок окажется разбитой, равна 0,003. Найти вероятность того, что магазин получит разбитых бутылок: а) ровно две; б) менее двух; в) не менее двух; г) хотя бы одну.

Ответ: а) 0,25; б) 0,56; в) 0,44; г) 0,78

6. Автомобильный завод выпускает 80% автомобилей без существенных дефектов. Какова вероятность того, что среди 600 автомобилей, поступивших с завода на автомобильную биржу, окажется не менее 500 автомобилей без существенных дефектов?

Ответ: 0,02

7. Сколько раз нужно бросить монету, чтобы с вероятностью 0,95 можно было ожидать, что относительная частота $\frac{k}{n}$ появлений герба отклонится от вероятности p=0,5 появления герба при одном бросании монеты не более, чем на 0,02?

Ответ: $n \geq 2401$.

Глава 2
Случайные величины.

Вторым основным понятием теории вероятностей, наряду с понятием случайного события, является понятие *случайной величины* (случайного числа).

Определение 1. *Величина X называется случайной, если в результате испытания (измерения) она примет значение, которое заранее неизвестно и которое зависит от случайных причин, учесть которые заранее невозможно.*

Например, оценка, которую студент получит на предстоящем экзамене - случайная величина; сумма выигрыша в лотерею – случайная величина; ошибка измерения, допускаемая измерительным прибором – случайная величина, и т. д.

Не следует путать случайное событие со случайной величиной. Например, сдача экзамена – это случайное событие, а оценка на экзамене – случайная величина; выигрыш в лотерею – случайное событие, а сумма выигрыша – случайная величина, и т. д.

Случайные величины делятся на два основных класса - *дискретные* и *непрерывные.*

Определение 2. *Случайная величина называется дискретной (прерывистой), если её возможные значения, нанесенные на числовую ось, выглядят на ней в виде отдельных, изолированных друг от друга, точек (x_1; x_2; x_3; …), которые можно пронумеровать* (рисунок 2.1). Число этих возможных значений (x_1; x_2; x_3; …) дискретной случайной величины X может быть как конечным, так и бесконечным.

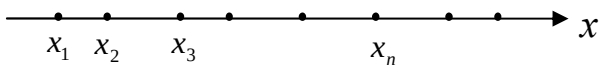


Рис. 2.1

Пример 1. Оценка X, которую получит студент на предстоящем экзамене – дискретная случайная величина. Действительно, её возможные значения (2, 3, 4, 5) на числовой оси представляются в виде отдельных, изолированных друг от друга, точек. Их число, кстати, конечно (равно четырем).

Пример 2. Число вызовов, поступающих за некоторое время (например, за сутки) на телефонную станцию – дискретная случайная величина X. Действительно, её возможные значения (0; 1; 2; 3; …) – это отдельные, изолированные друг от друга, точки числовой оси. Их число, заметим, теоретически бесконечно.

Определение 3. *Случайная величина X называется непрерывной, если она может принять любое значение x из некоторого числового промежутка [a; b] или интервала (a; b) числовой оси (конечного или бесконечного).*

Таким образом, возможные значения непрерывной случайной величины X заполняют *сплошь* (*непрерывно*) некоторый промежуток или интервал числовой оси (рисунок 2.2). При этом предполагается, что вероятность принятия величиной X своего значения на любом бесконечно малом участке этого промежутка или интервала *бесконечно мала.*

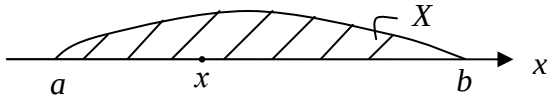


Рис. 2.2

Пример3. Время T работы некоторого механизма или устройства до очередного сбоя (поломки) – непрерывная случайная величина. Действительно, её возможные значения t заполняют сплошь положительную часть оси времени t (рисунок 2.3), ибо механизм может выйти из строя в любой момент времени t ($t > 0$). При этом вероятность того, что он выйдет из строя в течении любого бесконечного малого промежутка времени, очевидно, бесконечно мала.

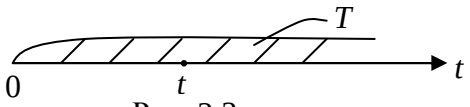


Рис. 2.3

§1. Дискретные случайные величины и их числовые характеристики.

Дискретная случайная величина X считается заданной, если известен список (x_1; x_2; x_3; … x_n) всех её возможных значений, а также известны вероятности (p_1; p_2; p_3; … p_n) принятия ею этих её возможных значений. Эти данные оформляются в виде таблицы,

X	x_1	x_2	x_3 …..x_n
p	p_1	p_2	p_3 …..p_n

(1.1)

которая называется *законом распределения* дискретной случайной величины X.

Отметим, что сумма вероятностей $p_1+p_2+\ldots+p_n$ представляет собой, согласно формуле (4.10) главы 1 для попарно несовместных событий, вероятность принятия случайной величиной X хотя бы одного из своих возможных значений (или x_1, или x_2, … или x_n). То есть представляет собой вероятность достоверного события – события, которое произойдет обязательно. Но вероятность достоверного события равна единице. Таким образом:

$$p_1+p_2+\ldots+p_n=1 \qquad (1.2)$$

Это равенство должно выполняться в любом законе распределения (1.1).

Пример1. В лотерее разыгрываются 200 лотерейных билетов. Из них 2 билета содержат выигрыш по 1000 рублей, 5 билетов по 500 рублей, 10 билетов по 300 рублей и 25 билетов по 100 рублей. Составить закон распределения дискретной случайной величины X – выигрыша для владельца одного лотерейного билета.

Решение. Искомый закон распределения очевиден:

X	0	100	300	500	1000
p	0,79	0,125	0,05	0,025	0,01

На практике обычно важно знать не столько сам закон распределения дискретной случайной величины (таблицу), сколько некоторые суммарные (обобщающие) числовые характеристики этой случайной величины, характеризующие её в целом. В частности, на практике наиболее важно знать:

а) среднее значение $x_{ср}$ случайной величины;

б) степень разброса её возможных значений (x_1; x_2; … x_n) вокруг её среднего значения $x_{ср}$.

Начнем с того, что найдем среднее значение $x_{ср}$ дискретной случайной величины X. Пусть таблица (1.1) – закон её распределения. Представим себе, что будет проведено N повторных испытаний. Так как (p_1; p_2; … p_n) – вероятности появления значений (x_1; x_2; … x_n) случайной величины X, то согласно формуле (1.1) главы 1 эти значения появятся в среднем (p_1N; p_2N; … p_nN) раз соответственно. Тогда среднее из всех этих N значений будет, очевидно, таким:

$$x_{ср} = \frac{x_1 \cdot p_1 N + x_2 \cdot p_2 N + ... + x_n \cdot p_n N}{N} = x_1 p_1 + x_2 p_2 + ... + x_n p_n = \sum_{k=1}^{n} x_k p_k \qquad (1.3)$$

Это среднее значение дискретной случайной величины X является *наиболее ожидаемым её значением* для каждого отдельного испытания, и потому называется *математическим ожиданием* случайной величины X. Оно обозначается символом $M(X)$. Итак,

$$M(X)=x_{ср}=\sum_{k=1}^{n} x_k p_k = x_1 p_1 + x_2 p_2 + ... + x_n p_n \qquad (1.4)$$

- математическое ожидание (среднее значение) дискретной случайной величины X. Вокруг математического ожидания группируются реальные средние значения случайной величины X в различных сериях испытаний.

Свойства математического ожидания дискретной случайной величины.

1. Математическое ожидание неизменной (постоянной) величины C равно самой этой постоянной:

$$M(C)=C \qquad (1.5)$$

Доказательство. Постоянная величина C принимает единственное значение (значение C) с вероятностью, равной единице. Следовательно, считая постоянную величину C частным случаем дискретной случайной величины X, будем иметь такой закон её распределения:

C	C
p	1

Отсюда $M(C)=C\cdot 1=C$. Доказательство закончено. Отметим, что этот вывод полностью согласуется и со смыслом математического ожидания случайной величины как её среднего значения: среднее значение неизменной величины C равно ей самой.

2. Постоянный множитель (константу) можно выносить за знак математического ожидания:

$$M(CX)=C\cdot M(X) \qquad (1.6)$$

Доказательство. Пусть X – дискретная случайная величина с законом распределения (1.1), и пусть C – любая константа. Тогда $C\cdot X$ также будет дискретной случайной величиной со следующим законом распределения:

CX	Cx_1	Cx_2	…	Cx_n
p	p_1	p_2	…	p_n

(1.7)

Действительно, умножая случайную величину X на множитель C, мы тем самым умножаем на C все её возможные значения (x_1; x_2; … x_n). А вероятности значений (Cx_1; Cx_2; … Cx_n) величины CX совпадают с вероятностями значений (x_1; x_2; … x_n) величины X, ибо величина CX примет значение C$x_к$ ($к$=1, 2, … n) тогда и только тогда, когда величина X примет значение $x_к$ ($к$=1, 2, … n). Но тогда из (1.7) следует:

$$M(CX) = \sum_{k=1}^{n} Cx_k \cdot p_k = Cx_1 p_1 + Cx_2 p_2 + ... + Cx_n p_n = C(x_1 p_1 + x_2 p_2 + ... + x_n p_n) = C \cdot M(X)$$

Доказательство закончено. И смысл доказанного свойства (1.6) понятен: если изменить (увеличить или уменьшить) в C раз все возможные значения случайной величины X, то в это же число раз изменится и её среднее значение.

3. Математическое ожидание суммы любых двух дискретных случайных величин X и Y равно сумме их математических ожиданий:

$$M(X+Y)=M(X)+M(Y) \qquad (1.8)$$

Доказательство. Пусть

X	x_1	x_2	…	x_n
p	p_1	p_2	…	p_n

Y	y_1	y_2	…	y_m
q	q_1	q_2	…	q_m

(1.9)

- законы распределения некоторых двух дискретных случайных величин X и Y. Составим закон распределения новой случайной величины $Z=X+Y$ – суммы величин X и Y. В качестве своих возможных значений величина $Z=X+Y$ будет принимать все возможные суммы вида x_i+y_j *(i=1, 2,… n; j=1, 2,… m)*, а вероятности этих значений, нам неизвестные, будем обозначать символами p_{ij}. Итак, закон распределения суммы $Z=X+Y$ примет вид:

$X+Y$	x_i+y_j
p	p_{ij}

$(i = 1,2,...n;\ \mathrm{j} = 1,2,...\mathrm{m})$ (1.10)

Тогда, следуя формуле (1.4), получаем:

$$M(X+Y) = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_i + y_j)p_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_i p_{ij} + y_j p_{ij}) =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} x_i p_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{m} y_j p_{ij} = \sum_{i=1}^{n} x_i \sum_{j=1}^{m} p_{ij} + \sum_{j=1}^{m} y_j \sum_{i=1}^{n} p_{ij} = \qquad (1.11)$$

=| учтем, что $\sum_{j=1}^{m} p_{ij} = p_{i1} + p_{i2} + ... + p_{im} = p_i$; $\sum_{i=1}^{n} p_{ij} = p_{1j} + p_{2j} + ... + p_{nj} = q_j$ -

- продумайте это, опираясь на формулу (4.10) главы 1, самостоятельно | =

$$= \sum_{i=1}^{n} x_i p_i + \sum_{j=1}^{m} y_j q_j = M(X) + M(Y)$$

Доказательство закончено. Из доказанного свойства (1.8) следует, что и математическое ожидание суммы нескольких дискретных случайных величин равно сумме их математических ожиданий. Например, для трех слагаемых получаем:

$$M(X+Y+Z) = M[(X+Y)+Z] = M(X+Y) + M(Z) = M(X) + M(Y) + M(Z) \qquad (1.12)$$

И вообще:

$$M(X_1 + X_2 + ... + X_p) = M(X_1) + M(X_2) + ... + M(X_p) \qquad (1.13)$$

- для любых дискретных случайных величин (X_1; X_2; … X_p).

4. Математическое ожидание разности любых двух дискретных случайных величин X и Y равно разности их математических ожиданий:

$$M(X - Y) = M(X) - M(Y) \qquad (1.14)$$

Доказательство. Оно непосредственно вытекает из уже доказанных свойств (1.6) и (1.8):

$$M(X - Y) = M[X + (-1)Y] = M(X) + M[(-1)Y] = M(X) + (-1)M(Y) = M(X) - M(Y)$$

И вообще, из доказанных выше свойств вытекает, что математическое ожидание любой линейной комбинации $C_1X_1 + C_2X_2 + \ldots + C_pX_p$ любых дискретных случайных величин (X_1; X_2; … X_p), где (C_1; C_2; … C_p) – любые числовые множители, находится по формуле:

$$M(C_1X_1 + C_2X_2 + ... + C_pX_p) = C_1M(X_1) + C_2M(X_2) + ... + C_pM(X_p) \qquad (1.15)$$

5. Математическое ожидание произведения двух дискретных случайных величин X и Y, *если эти величины независимы*, равно произведению их математических ожиданий:

$$M(XY) = M(X) \cdot M(Y) \qquad (1.16)$$

Доказательство. Пусть таблицы (1.9) представляют собой законы распределения двух независимых случайных величин X и Y. Их независимость означает, что в испытании эти величины принимают свои возможные значения вне всякой связи друг с другом (независимо друг от друга). Составим закон распределения дискретной случайной величины $Z=XY$. В качестве своих возможных значений величина $Z=XY$ будет принимать все возможные произведения вида x_iy_j $(i = 1,2,...n;\ \ j = 1,2,...m)$, а вероятности p_{ij} этих значений, по формуле умножения вероятностей независимых событий (формула (4.7) главы 1), будут равны $p_{ij} = p_i \cdot q_j$. Таким образом, закон распределения дискретной случайной величины $Z=XY$ будет выглядеть следующим образом:

XY	x_iy_j
p	p_iq_j

$(i = 1,2,...n;\ \ j = 1,2,...m)$ (1.17)

Отсюда:

$$M(XY) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_iy_jp_iq_j = \sum_{i=1}^{n} x_ip_i \cdot \sum_{j=1}^{m} y_jq_j = M(X) \cdot M(Y) \qquad (1.18)$$

Доказательство закончено. Из доказанного свойства (1.16) следует, что и математическое ожидание произведения нескольких *взаимно независимых* случайных величин (X_1; X_2; … X_p) равно произведению их математических ожиданий:

$$M(X_1 \cdot X_2 \cdot ... \cdot X_p) = M(X_1) \cdot M(X_2) \cdot ... \cdot M(X_p) \qquad (1.19)$$

Доказательство этого проводится аналогично доказательству равенства (1.13).

Перейдем теперь к вопросу о том, как охарактеризовать степень разброса возможных значений (x_1; x_2; … x_n) дискретной случайной величины X вокруг её математического ожидания (вокруг её среднего значения) $M(X) = x_{ср}$.

На первый взгляд может показаться, что для оценки указанного разброса следует вычислить все возможные отклонения значений случайной величины X от её $M(X)$, то есть вычислить разности

$$x_1 - M(X);\ \ x_2 - M(X);\ \ldots\ldots\ x_n - M(X)$$

и найти, с учетом вероятностей (p_1; p_2; … p_n) этих разностей, их среднее значение. То есть найти $M[X - M(X)]$. Однако такой путь ничего не дает, так как эта величина всегда равна нулю:

$$M[X - M(X)] = M(X) - M(M(X)) = M(X) - M(X) = 0 \qquad (1.20)$$

Этот результат объясняется тем, что одни возможные отклонения X от $M(X)$ положительны, другие отрицательны, так что их среднее значение в результате их взаимного погашения равно нулю. Это обстоятельство указывает на целесообразность замены отклонений $X - M(X)$ их абсолютными величинами $|X - M(X)|$ или их квадратами $(X - M(X))^2$. Первый из этих вариантов предполагает оперирование с абсолютными величинами (модулями), что не очень удобно. Поэтому общепринятым является второй путь. А именно, вычисляют $M[X - M(X)]^2$ - среднее значение квадрата отклонения величины X от её среднего значения $M(X)$, которое называют *дисперсией* $D(X)$ величины X. А затем, извлекая квадратный корень из дисперсии $D(X)$, находят среднее отклонение X от $M(X)$ уже без квадрата этого отклонения. В нем знак отклонения X от $M(X)$ уже не учитывается (этот знак всегда плюс).

Указанный $\sqrt{D(X)}$ называется *средним квадратическим отклонением* случайной величины X и обозначается символом $\sigma(X)$. Величина $\sigma(X)$ показывает, на сколько *в среднем* отклоняется X от $M(X)$, если не учитывать знак отклонения X от $M(X)$. Итак,

$$D(X) = M[X - M(X)]^2; \qquad \sigma(X) = \sqrt{D(X)} \qquad (1.21)$$

- *дисперсия* $D(X)$ *и среднее квадратическое отклонение* $\sigma(X)$ *величины* X.

Обе эти величины характеризуют степень разброса значений величины X вокруг её среднего значения $M(X)$.

Действительно, чем сильнее разбросаны значения X вокруг $M(X)$, тем больше числовые значения $X - M(X)$ отклонений X от $M(X)$. А значит, тем больше квадраты $[X - M(X)]^2$ этих отклонений. Но тогда тем больше среднее значение $M[X - M(X)]^2$ квадратов этих отклонений. То есть тем больше дисперсия $D(X)$ случайной величины X. А значит, тем больше и среднее квадратическое отклонение $\sigma(X)$ величины X. А чем меньше разброс значений случайной величины X вокруг её математического ожидания $M(X)$ (то есть чем кучнее они вокруг $M(X)$ расположены), тем меньше величины $D(X)$ и $\sigma(X)$.

Из этих двух числовых характеристик случайной величины X важнейшей является $\sigma(X)$ в силу её наиболее ясного смысла ($\sigma(X)$ – это среднее отклонение X от $M(X)$ без учета знака этого отклонения). А дисперсия $D(X)$ является вспо-

могательной величиной, по которой затем определяется среднее квадратическое отклонение *σ(X)*.

Если (1.1) – закон распределения дискретной случайной величины *X*, то случайная величина $Z=[X-M(X)]^2$ имеет, очевидно, следующий закон распределения:

$[X-M(X)]^2$	$[x_1-M(X)]^2$	$[x_2-M(X)]^2$	…	$[x_n-M(X)]^2$
p	p_1	p_2	…	p_n

(1.22)

И тогда, согласно определению (1.21) дисперсии $D(X)$ и формуле (1.4), получаем следующую формулу для практического подсчета дисперсии $D(X)$:

$$D(X)=\sum_{k=1}^{n}[x_k-M(X)]^2\cdot p_k \qquad (1.23)$$

Исходя из определения дисперсии (1.21), для её практического вычисления можно получить и другую, так называемую *упрощенную*, формулу. Используя свойства математического ожидания, получим:

$$\begin{aligned} D(X)&=M[X-M(X)]^2=M[X^2-2X\cdot M(X)+(M(X))^2]=\\ &=M(X^2)-2M(X)\cdot M(X)+(M(X))^2=M(X^2)-[M(X)]^2 \end{aligned} \qquad (1.24)$$

Итак,

$$D(X)=M(X^2)-[M(X)]^2 \qquad (1.25)$$

- *упрощенная формула* для дисперсии $D(X)$. Читается эта формула так: *дисперсия дискретной случайной величины X равна математическому ожиданию квадрата этой величины минус квадрат её математического ожидания*. При этом:

$$M(X)=\sum_{k=1}^{n}x_k\cdot p_k\ ;\quad M(X^2)=\sum_{k=1}^{n}x_k^2\cdot p_k \qquad (1.26)$$

После вычисления дисперсии $D(X)$ находится и $\sigma(X)=\sqrt{D(X)}$ - среднее квадратическое отклонение величины *X*.

Для сравнения *σ(X)* с *M(X)* вводится величина

$$V(X)\%=\frac{\sigma(X)}{M(X)}\cdot 100\ \%, \qquad (1.27)$$

которая называется *коэффициентом вариации* величины *X*. Она показывает, какую долю в процентах составляет для случайной величины *X* её среднее отклонение от среднего по отношению к самому среднему.

Пример 2. дискретная случайная величина *X* имеет следующий закон распределения:

X	0	3	6
p	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$

Найти её числовые характеристики *M(X); D(X); σ(X); V(X)%*.

Решение.

1) Найдем *M(X)*. Используя формулу (1.4), получим:

$$M(X)=\sum_{k=1}^{3}x_k p_k = 0\cdot\frac{1}{2}+3\cdot\frac{1}{3}+6\cdot\frac{1}{6}=2.$$

2) Найдем *D(X)*. Используя основную формулу (1.23), получаем:

$$D(X)=\sum_{k=1}^{3}(x_k-2)^2 p_k = (0-2)^2\cdot\frac{1}{2}+(3-2)^2\cdot\frac{1}{3}+(6-2)^2\cdot\frac{1}{6}=5;$$

Заметим, что то же самое значение *D(X)*=5 мы получим, если используем упрощенную формулу (1.25) и выражения (1.26):

$$M(X)=2;\quad M(X^2)=0^2\cdot\frac{1}{2}+3^2\cdot\frac{1}{3}+6^2\cdot\frac{1}{6}=9;\quad D(X)=9-2^2=5$$

3) Найдем *σ(X)*:

$$\sigma(X)=\sqrt{D(X)}=\sqrt{5}\approx 2{,}236.$$

4) Найдем *V(X)%:*

$$V(X)\% = \frac{\sigma(X)}{M(X)}\cdot 100\ \% \approx \frac{2{,}236}{2}\cdot 100\ \% \approx 112\ \%.$$

Таким образом, среднее значение $x_{ср}$ данной случайной величины *X* равно *M(X)*=2. То есть принимая три возможных значениях (0; 3; 6), случайная величина *X* в среднем принимает значение 2. Найденное значение не совпадает со средним арифметическим этих чисел (с 3) и оказалось ближе к 0, нежели к 6. Это произошло потому, что вероятность $\left(\frac{1}{2}\right)$ значения 0 гораздо больше вероятности $\left(\frac{1}{6}\right)$значения 6. А это значит, что при повторных испытаниях значение 0 будет встречаться гораздо чаще (втрое чаще), чем значение 6. Поэтому и среднее из значений, принимаемых случайной величиной *X*, будет сдвинуто от среднего арифметического (от числа 3) в сторону нуля. Точно так же средний балл аттестата зрелости выпускника школы ближе к трем, чем к пяти, если в этом аттестате больше троек, чем пятерок.

Далее, среднее отклонение *σ(X)* величины *X* от её среднего значения *M(X)*=2 оказалось равным ≈2,236. И это тоже хорошо объяснимо. Действительно, у случайной величины *X* три возможных значения (0; 3; 6) при $x_{ср}=M(X)=2$. Отклонения этих значений от их среднего значения $x_{ср}=2$ составляют (без учета знака отклонения) соответственно (2; 1; 4). Из этих трех отклонений средним оказалось *σ(X)*≈2,236. Оно наиболее близко к 2, то есть наиболее близко к отклонению значения 0 от $x_{ср}=2$. И это оправдано, так как значение 0 величины *X* при повторных испытаниях будет встречаться чаще других значений.

Наконец, величина коэффициента вариации величины *X* показывает, что *σ(X)*≈2,236 (среднее отклонение от среднего) по отношению к *M(X)*=2 (к самому среднему) составляет ≈112%.

Свойства дисперсии дискретной случайной величины.

Для нахождения *σ(X)* и *V(X)%* нужно предварительно найти *D(X)* – дисперсию величины *X*. При её нахождении полезно знать и использовать свойства дисперсии. Эти свойства таковы:

1. *Дисперсия постоянной величины C равна нулю*:

$$D(C)=0 \qquad (1.28)$$

Доказательство. Согласно определению дисперсии (1.21) и свойству (1.5) при $X=C$ получаем:

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

Свойство (1.28) имеет очевидный смысл: константа разброса не имеет.

2. *Постоянный множитель выносится из-под знака дисперсии в квадрате*:

$$D(CX)=C^2\cdot D(X) \qquad (1.29)$$

Доказательство. Согласно определению дисперсии (1.21) и свойству (1.6) математического ожидания получаем:

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

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

$$D(X+Y)=D(X)+D(Y) \qquad (1.30)$$

Доказательство. Используя упрощенную формулу (1.25) для дисперсии и доказанные выше свойства математического ожидания, получим:

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

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

$$D(X-Y)=D(X)+D(Y) \qquad (1.31)$$

Доказательство. Оно вытекает из уже доказанных свойств (1.29) и (1.30):

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

Дисперсия случайной величины, как мы знаем, характеризует степень разброса ее возможных значений вокруг ее среднего значения. Свойства 3 и 4, выражаемые равенствами (1.30) и (1.31), означают суммирование (увеличение) такого разброса как при сложении двух независимых случайных величин, так и при их вычитании, что вполне естественно.

Следствие 1 свойств 1– 4: *для любой дискретной случайной величины X и для любой константы C*

$$D(X+C)=D(X) \qquad (1.32)$$

Доказательство. Так как X и C очевидным образом независимы друг от друга, то

$$D(X+C)=D(X)+D(C)=D(X)+0=D(X).$$

Следствие 2 свойств 1– 4: *если (X_1; X_2; … X_P) – взаимно независимые случайные величины, то*

$$D(X_1\pm X_2\pm\ldots\pm X_P)=D(X_1)+D(X_2)+\ldots+D(X_P) \qquad (1.33)$$

Это следствие докажите самостоятельно (за образец возьмите доказательство аналогичного равенства (1.13)).

Упражнения

1. К какому типу случайных величин (дискретным или непрерывным) принадлежат следующие величины:

а) Количество очков, выбиваемых стрелком при стрельбе по мишени в заданной серии выстрелов.

б) Количество студентов студенческой группы, которые сдадут предстоящий экзамен.

в) Процент выполнения предприятием своего производственного задания.

г) Количество осадков, которые выпадут ближайшим летом в данной местности.

Ответ: (а) и (б) – к дискретным, (в) и (г) – к непрерывным.

2. Станок-автомат производит в среднем 10% нестандартных деталей. Наудачу из его продукции отобраны три детали. Написать закон распределения дискретной случайной величины X – числа нестандартных деталей среди трех отобранных. Найти $M(X)$, $D(X)$, $\sigma(X)$, $V(X)\%$.

Ответ:

X	0	1	2	3
p	0,729	0,243	0,027	0,001

$M(X)=0{,}3$; $D(X)=0{,}27$; $\sigma(X)\approx 0{,}52$; $V(X)\%\approx 173\%$.

3. После ответа студента на вопросы экзаменационного билета экзаменатор задает студенту дополнительные вопросы. Экзаменатор прекращает задавать дополнительные вопросы, как только студент обнаружит незнание заданного вопроса. Вероятность того, что студент ответит на любой заданный дополнительный вопрос, равна $0{,}9$. Составить закон распределения дискретной случайной величины X – числа дополнительных вопросов, которые задаст экзаменатор студенту. Найти $M(X)$, $D(X)$, $\sigma(X)$, $V(X)\%$.

Ответ:

X	1	2	3	4	…	n	…
p	0,1	$0{,}9\cdot 0{,}1$	$(0{,}9)^2\cdot 0{,}1$	$(0{,}9)^3\cdot 0{,}1$	…	$(0{,}9)^{n-1}\cdot 0{,}1$	…

$M(X)=10$; $D(X)=90$; $\sigma(X)\approx 9{,}5$; $V(X)\%\approx 95\%$.

4. Показать, что если X – произвольная дискретная случайная величина, а C_1 и C_2 – произвольные константы, то

$$M(C_1X+C_2)=C_1M(X)+C_2;\quad D(C_1X+C_2)=C_1^2D(X);\quad \sigma(C_1X+C_2)=|C_1|\sigma(X).$$

5. Брошены две игральные кости. Найти математическое ожидание и дисперсию суммы очков $X+Y$, выпадающих на обеих костях. Причем найти двумя способами:

а) Составив закон распределения суммы очков.

б) Воспользовавшись формулами (1.8) и (1.30).

Ответ:

$$M(X+Y)=7;\quad D(X+Y)\approx 2{,}9.$$

6. Пусть X, Y и Z – следующие случайные величины: X – выручка фирмы (доход от продаж); Y – ее затраты; $Z=X-Y$ – ее прибыль. Найти закон распределения прибыли Z, если выручка фирмы и её затраты – независимые случайные величины, заданные законами распределения:

X	3	4	5
p	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Y	1	2
p	$\frac{1}{2}$	$\frac{1}{2}$

Найти среднее значение прибыли (все суммы – в млн. рублей).

Ответ:

$Z=X-Y$	1	2	3	4
p	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{2}{6}$	$\frac{1}{6}$

$M(Z)=2{,}5$ (млн. руб.)

7. Дискретная случайная величина X имеет только два возможных значения x_1 и x_2, причем $x_1<x_2$. Вероятность того, что X примет значение x_1, равна $0{,}2$. Найти закон распределения величины X, если $M(X)=2{,}6$ и $\sigma(X)=0{,}8$.

Ответ:

X	1	3
p	0,2	0,8

§2. Некоторые важнейшие распределения дискретных случайных величин.

Мы ограничимся рассмотрением двух из таких распределений.

1. Биномиальное распределение.

Пусть X – число появлений некоторого события A в n повторных испытаниях, где $p(A)=p$ и $p(\overline{A})=1-p=q$ – вероятности появления и непоявления этого события в одном испытании. Очевидно, что X – дискретная случайная величина, у которой $(0;1;2;\dots n)$ – возможные значения, а вероятности $(p_0;p_1;p_2;\dots p_n)$ этих значений найдутся по формуле Бернулли:

$$p_k=P_n(k)=\frac{n!}{k!(n-k)!}p^k q^{n-k}\quad (k=0;1;2;\dots n) \tag{2.1}$$

Закон распределения

X	0	1	2	…	n
p	p_0	p_1	p_2	…	p_n

(2.2)

такой случайной величины называется *биномиальным*.

Докажем, что математическое ожидание $M(X)$ и дисперсия $D(X)$ случайной величины X, распределенной по биномиальному закону, могут быть найдены по следующим простым формулам:

$$M(X) = np; \qquad D(X) = npq. \tag{2.3}$$

Для этого представим величину X в виде суммы

$$X = X_1 + X_2 + \ldots + X_n \tag{2.4}$$

Здесь X_k – это число появлений события A в отдельно взятом k-ом испытании $(k = 1,2,\ldots n)$. Закон распределения случайной величины X_k при любом номере k будет, очевидно, иметь вид:

X_k	0	1
p	q	p

$(k = 1,2,\ldots n)$ (2.5)

Тогда

$$\begin{aligned} &M(X_k) = 0 \cdot q + 1 \cdot p = p; \\ &M(X_k^2) = 0^2 \cdot q + 1^2 \cdot p = p; \\ &D(X_k) = M(X_k^2) - [M(X_k)]^2 = p - p^2 = p(1-p) = pq. \end{aligned} \tag{2.6}$$

Случайные величины $(X_1; X_2; \ldots X_n)$ по смыслу своему независимы друг от друга. Поэтому, опираясь на формулы (1.13) и (1.33), получим:

$$\begin{aligned} &M(X) = M(X_1 + X_2 + \ldots + X_n) = M(X_1) + M(X_2) + \ldots + M(X_n) = p + p + \ldots + p = np; \\ &D(X) = D(X_1 + X_2 + \ldots + X_n) = D(X_1) + D(X_2) + \ldots + D(X_n) = pq + pq + \ldots + pq = npq \end{aligned} \tag{2.7}$$

Доказательство закончено.

Пример 1. Составить закон распределения случайной величины X – числа выпадений герба при трех бросаниях монеты. Найти $M(X)$; $D(X)$; $\sigma(X)$; $V(X)$%.

Решение. Так как случайная величина X представляет собой число появлений события A (выпадения герба) при трех повторных испытаниях (при трех бросаниях монеты), то она имеет биномиальный закон распределения:

X	0	1	2	3
p	p_0	p_1	p_2	p_3

Вероятности (p_0; p_1; p_2; p_3) найдем по формуле Бернулли (2.1) при $p(A) = p = \frac{1}{2}$ и $p(\overline{A}) = q = \frac{1}{2}$:

$$p_0 = \frac{1}{8}; \quad p_1 = \frac{3}{8}; \quad p_2 = \frac{3}{8}; \quad p_3 = \frac{1}{8}.$$

Таким образом, для данной биномиально распределенной случайной величины X получаем следующий закон распределения:

X	0	1	2	3
p	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Ее числовые характеристики $M(X)$; $D(X)$; $\sigma(X)$; $V(X)\%$ найдем с помощью доказанных выше формул (2.3):

$$M(X)=np=3\cdot\frac{1}{2}=1{,}5; \qquad D(X)=npq=3\cdot\frac{1}{2}\cdot\frac{1}{2}=0{,}75;$$

$$\sigma(X)=\sqrt{D(X)}\approx 0{,}866; \qquad V(X)\%=\frac{\sigma(X)}{M(X)}\cdot 100\%\approx 57{,}7\%.$$

2. Поток событий. Распределение Пуассона.

Распределение Пуассона используется в задачах, связанных с потоком событий. Под потоком событий понимают последовательность событий, появляющихся одно за другим в случайные моменты времени. Примерами потоков событий являются: поток звонков на телефонную станцию, в милицию или на станцию скорой помощи; поток заявок в системе массового обслуживания; поток автомобильных аварий на дорогах города и т.д. Поток событий называется простейшим, или пуассоновским, если он характеризуется следующими свойствами:

1) Вероятность появления k событий потока за промежуток времени длительностью t зависит лишь от величины k и длительности t времени, а не от начала отсчета времени.

2) Вероятность появления двух и более событий за малый промежуток времени пренебрежимо мала по сравнению с вероятностью появления за это время только одного события.

Можно доказать (см. например [6], § 2.4), что вероятность $P_t(k)$ появления k событий простейшего потока за время t определяется формулой:

$$P_t(k)=\frac{(\lambda t)^k\cdot e^{-\lambda t}}{k!} \qquad (2.8)$$

Здесь λ – *интенсивность пуассоновского потока*, под которой понимается *среднее число событий, появляющихся в единицу времени*.

Пусть X – число событий простейшего потока, происходящих за заданное время t. Очевидно, что X – дискретная случайная величина с возможными значениями $(0;1;2;\ldots n;\ldots)$ и вероятностями $(p_0;\ p_1;\ p_2;\ldots p_n;\ldots)$, находимыми, согласно (2.8), по формуле:

$$p_k=p(X=k)=P_t(k)=\frac{a^k}{k!}e^{-a}, \quad \text{где} \quad a=\lambda t \quad (k=0,1,2,\ldots) \qquad (2.9)$$

Распределение указанной дискретной случайной величины X

X	0	1	2	3	…	n	…
p	p_0	p_1	p_2	p_3	…	p_n	…

(2.10)

называется *распределением Пуассона*. Его еще *называют распределением редких событий*. Последнее название связано с тем, что по этой же формуле Пуассона (2.9), согласно формуле 6.6 главы 1, приближенно находятся вероятности $P_n(k)$ появления редкого события k раз в n испытаниях.

Если вычислить числовые характеристики величины X, имеющей распределение Пуассона, то получим (выкладки опускаем):

$$M(X)=a;\quad D(X)=a;\quad \sigma(X)=\sqrt{a};\quad V(X)\%=\frac{\sqrt{a}}{a}\cdot 100\%. \qquad (2.11)$$

Пример 2. Среднее число бракованных деталей, изготавливаемых станком-автоматом в течение одного часа, равно 6. Составить закон распределения дискретной случайной величины X – числа бракованных деталей, которые будут изготовлены станком-автоматом в ближайшие полчаса. Найти числовые характеристики $M(X)$; $D(X)$; $\sigma(X)$; $V(X)\%$ этой случайной величины.

Решение. Будем считать поток бракованных деталей простейшим (пуассоновским). Тогда закон распределения рассматриваемой случайной величины X будет иметь вид (2.10). При этом вероятности (p_0; p_1; p_2;…) находятся по формуле (2.9) при

$$a=\lambda t=\left|\lambda=6;\quad t=0{,}5\right|=3.$$

То есть

$$p_0=\frac{3^0}{0!}e^{-3}=e^{-3}\approx 0{,}0498;\quad p_1=\frac{3^1}{1!}e^{-3}=3e^{-3}\approx 0{,}1494;\quad p_2\approx 0{,}2241;\quad (\ldots).$$

И тогда закон распределения величины X примет вид:

X	0	1	2	3	4	…
p	0,0498	0,1494	0,2241	0,2241	0,1680	…

А числовые характеристики величины X найдутся по формулам (2.11):

$$M(X)=3;\quad D(X)=3;\quad \sigma(X)=\sqrt{3}\approx 1{,}73;\quad V(X)\%\approx 58\%.$$

Упражнения

1. Есть ли что-нибудь общее у биномиального распределения и распределения Пуассона. Что именно?

2. Два равносильных игрока играют в игру, ничьи в которой исключаются. Составить закон распределения случайной величины X – числа партий, которые выиграет первый игрок, если будут играться 4 партии. Найти числовые характеристики этой случайной величины.

Ответ:

X	0	1	2	3	4
p	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

$$M(X)=2;\ D(X)=1;\ \sigma(X)=1;\ V(X)\%=50\%.$$

3. В магазин привезли 300 стеклянных бутылок с минеральной водой. Известно, что в среднем при перевозке одна из 500 бутылок разбивается. Составить закон распределения случайной величины X – числа разбитых бутылок в

привезенной партии. Найти числовые характеристики этой случайной величины.

Ответ:

X	0	1	2	3	⋯	300
p	0,549	0,329	0,099	0,020	…	0,000

$M(X)=0{,}6;\ D(X)=0{,}6;\ \sigma(X)\approx 0{,}77;\ V(X)\%\approx 129\%.$

§3. Непрерывные случайные величины и их числовые характеристики.

Пусть X – непрерывная случайная величина, возможные значения которой заполняют сплошь (непрерывно) некоторый промежуток $[a;b]$ числовой оси ox (рис. 2.2). Для полной характеристики этой величины X недостаточно, очевидно, задать только этот промежуток ее возможных значений. Нужно еще как-то указать, насколько возможно (вероятно) каждое из этих значений. Сделать это так, как это делалось для дискретных случайных величин, то есть с помощью закона распределения (таблицы), в которой одно за другим перечислены все возможные значения случайной величины и указаны их вероятности, очевидно, невозможно. Для непрерывной случайной величины это делается иначе – с помощью так называемой *плотности вероятности*.

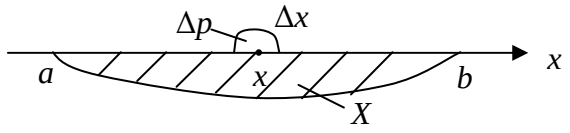


Рис. 2.4

Введем это важное понятие. Пусть x – одно из возможных значений непрерывной случайной величины X. Окружим это значение x некоторым малым промежутком длиной Δx (окрестностью Δx точки x). И пусть Δp – вероятность того, что в результате испытания величина X примет значение, принадлежащее этой окрестности (попадет в эту окрестность) – рис. 2.4. Если мы теперь найдем отношение $\frac{\Delta p}{\Delta x}$, то получим вероятность, приходящуюся *в среднем на единицу длины* участка Δx. То есть получим *среднюю линейную плотность вероятности* непрерывной случайной величины X на участке Δx. А теперь в отношении $\frac{\Delta p}{\Delta x}$ перейдем к пределу при $\Delta x \to 0$ (при этом и $\Delta p \to 0$). При таком предельном переходе отрезок Δx будет стягиваться в точку x, и в пределе получим истинную линейную плотность вероятности величины X *в самой точке* x. Ее обозначим символом $f(x)$

$$\lim_{\Delta x\to 0}\frac{\Delta p}{\Delta x}=f(x) \qquad \text{или} \qquad \frac{dp}{dx}=f(x) \tag{3.1}$$

и будем называть *плотностью вероятности непрерывной случайной величины* X *в точке* x. В последнем равенстве (3.1) dx – это длина бесконечно малого

промежутка, окружающего точку x, а dp – это бесконечно малая вероятность попадания значения величины X на этот бесконечно малый промежуток.

Ясно, что плотность вероятности $f(x)$ неотрицательна для любого $x \in [a;b]$, ибо сама вероятность по природе своей неотрицательна:

$$f(x) \geq 0 \qquad (a \leq x \leq b) \tag{3.2}$$

Если у непрерывной случайной величины X задан промежуток $[a;b]$ (или интервал $(a;b)$) ее возможных значений, а также задана ее плотность вероятности $f(x)$ для всех x, принадлежащих этому промежутку или интервалу, то непрерывная случайная величина X считается заданной.

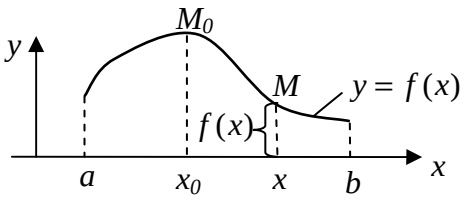


Рис. 2.5

Отметим, что если мы построим график плотности вероятности $y = f(x)$, то получим возможность наглядно сравнивать вероятности различных возможных значений x случайной величины X (рис. 2.5).

Действительно, согласно своему определению (3.1), плотность вероятности $f(x)$, вычисленная в точке x, показывает, какой была бы вероятность попадания значения случайной величины X *на единицу длины* отрезка $[a;b]$, если бы все значения этой величины были столь же вероятны, что и данное значение x. В этом смысле плотность вероятности совершенно аналогична, например, линейной плотности вещества материального отрезка $[a;b]$ (материальной нити), если вещество распределено вдоль этого отрезка. Указанную линейную плотность вещества $\rho = f(x)$ для различных значений $x \in [a;b]$ мы опять вычисляли бы по формуле (3.1), только Δp была бы не вероятность, а масса отрезка Δx. И ее числовое значение в выбранной точке x (например, $\rho = f(x) = 5 г/см$) означало бы, что если бы вещество было распределено вдоль отрезка $[a;b]$ равномерно и столь же плотно, что и в точке x (нить на всем своем протяжении имела бы ту же толщину, что и в точке x), то масса каждой единицы длины (каждого сантиметра) отрезка $[a;b]$ составляла бы $f(x)$ единиц массы ($5 г.$).

Итак, смысл плотности вероятности выяснен. Если при изменении x меняется и величина $f(x)$ (как на рис. 2.5), то более вероятны те значения x, для которых $f(x)$ больше. И во столько раз более вероятны, во сколько раз больше их плотность вероятности $f(x)$. В частности, на рис. 2.5 наиболее вероятным значением величины X на отрезке $[a;b]$ является значение x_0, а наименее вероятным – значение b. Можно даже зрительно и оценить, во сколько раз значение x_0 вероятнее значения b (раза в три).

Наиболее вероятное значение случайной величины (как непрерывной, так и дискретной) называется *модой* X (обозначается $MO(X)$). Для непрерывной случайной величины X, представленной на рис.2.5,

$$MO(X) = x_0.$$

Мода величины X, как наиболее вероятное ее значение, при повторных испытаниях встречается чаще других ее значений. Отсюда и название: мода.

А теперь получим в некотором смысле парадоксальный вывод: *у непрерывной случайной величины X все ее возможные значения x имеют нулевую вероятность!*

Действительно, из (3.1) следует:

$$dp = f(x)dx \qquad 3.4)$$

Здесь dp – это бесконечно малая вероятность того, что случайная величина X попадет в бесконечно малую окрестность dx точки x. В этой окрестности, несмотря на ее бесконечную малость, кроме точки x содержится еще бесконечно много других точек. Если сузить указанную окрестность до одной точки x, то в этом случае $dx = 0$, а значит, и $dp = 0$. И эта $dp = 0$ – вероятность того, что $X = x$. То есть действительно

$$p(X = x) = 0 \qquad (3.5)$$

То, что все возможные значения x непрерывной случайной величины X имеют нулевую вероятность – еще не значит, что эта величина X не может принять то или иное свое значение. Ведь в каждом испытании случайная величина X какое-то значение примет. Значит, в принципе она может принять любое из своих возможных значений. Просто этих возможных значений у нее бесконечно много, поэтому для каждого возможного значения x имеется лишь один шанс быть принятым против бесконечного числа шансов быть не принятым. Отсюда и следует нулевая вероятность для каждого возможного значения x непрерывной случайной величины X. Но шанс этот все-таки есть!

Далее: нулевые вероятности различных значений x непрерывной случайной величины X в сумме должны давать единицу! Действительно, указанная сумма вероятностей всех возможных значений x величины X – это вероятность того, что величина X примет одно из этих своих значений. А это – достоверное событие, вероятность которого, как известно, равна единице. Ну, а то, что из бесконечного числа нулей в сумме получается единица – так точно так же из бесконечного числа точек, не имеющих размера (имеющих нулевой размер) складываются реальные отрезки, материальные тела и т.д.

Несмотря на нулевые вероятности возможных значений x непрерывной случайной величины X, эти нулевые вероятности, как указывалось выше, можно сравнивать (с помощью плотности вероятности $f(x)$). Те значения x, для которых $f(x)$ больше, и вероятность имеют большую (одна нулевая вероятность больше другой!). То есть с помощью плотности вероятности $f(x)$ сравниваются весомости тех единственных шансов быть принятыми, которые есть у каждого возможного значения x величины X. В частности, на рис. 2.5 самым весомым этот шанс является у значения x_0 – моды величины X. Указанное сравнение нулевых вероятностей совершенно аналогично сравнению нулевых масс отдельных точек разных веществ. Например, совершенно естественно считать, что отдельно взятая точка свинца имеет нулевую массу, но большую, чем нулевая масса отдельно взятой точки алюминия, ибо при одном и том же количестве точек (при одном и том же объеме) масса свинца больше массы алюминия. И сравнивать количественно нулевые массы отдельных точек разных веществ можно по плотности этих веществ: во сколько раз плотность свинца больше

плотности алюминия, во столько же раз точка свинца массивнее точки алюминия.

При исследовании непрерывных случайных величин одной из основных задач является задача нахождения вероятности $p(x_1 \le X \le x_2)$ того, что в результате испытания заданная непрерывная случайная величина X примет значение, содержащееся в некоторой заданной части $[x_1; x_2]$ промежутка $[a; b]$. Эта вероятность находится по формуле:

$$p(x_1 \le X \le x_2) = \int_{x_1}^{x_2} f(x)dx \qquad (3.6)$$

Иначе говоря,

$$p(x_1 \le X \le x_2) = S, \qquad (3.7)$$

где S – площадь вертикальной полосы, изображенной на рис. 2.6:

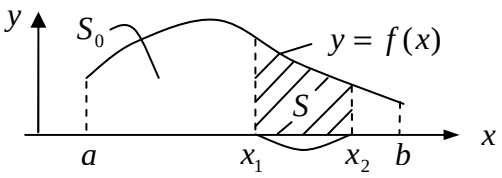


Рис. 2.6

Докажем формулу (3.7). Для этого мысленно разобьем отрезок $[x_1; x_2]$ на бесконечно малые участки с длинами dx и выберем внутри каждого такого участка некоторую точку x. Тогда вероятность dp попадания значения X на каждый из таких участков dx найдется по формуле (3.4). Попадание же значения X на отрезок $[x_1; x_2]$ равносильно попаданию этого значения хотя бы на один из указанных участков dx. Попадание значения величины X на различные участки dx – это попарно несовместные случайные события. Поэтому по формуле сложения вероятностей попарно несовместных событий (по формуле (4.10) главы 1) получаем:

$$p(x_1 \le X \le x_2) = \sum dp = \sum f(x)dx = \int_{x_1}^{x_2} f(x)dx .$$

Формула (3.6) доказана. А вспоминая геометрический смысл определенного интеграла, приходим и к формуле (3.7).

Так как $p(a \le X \le b) = 1$ как вероятность достоверного события, то из формул (3.6) и (3.7) как частный случай следует:

$$\int_a^b f(x)dx = S_0 = 1 \qquad (3.8)$$

Здесь площадь S_0 – площадь всей криволинейной трапеции, расположенной между осью ox и графиком плотности вероятности $y = f(x)$ для $a \le x \le b$.

Из равенства (3.8) следует, что плотность вероятности $f(x)$ любой непрерывной случайной величины X нельзя задавать произвольно. При ее задании обязательно должны быть выполнены два вытекающих из ее смысла условия: (3.2) и (3.8). В остальном она может быть какой угодно.

Числовые характеристики непрерывных случайных величин

Рассмотрим теперь вопрос об основных числовых характеристиках ($M(X)$; $D(X)$; $\sigma(X)$; $V(X)\%$) непрерывной случайной величины X. Этим числовым ха-

рактеристикам придадим тот же смысл, какой они имели в §1 для дискретных случайных величин:

$$M(X)=x_{ср};\quad D(X)=M[X-M(X)]^2;$$

$$\sigma(X)=\sqrt{D(X)};\quad V(X)\%=\frac{\sigma(X)}{M(X)}\cdot 100\% \qquad (3.9)$$

Начнем с $M(X)$ – с математического ожидания величины X. Найти $M(X)$ – это значит найти среднее значение $x_{ср}$ величины X. Для его нахождения разобьем мысленно отрезок $[a;b]$ всех возможных значений непрерывной случайной величины X на бесконечно большое число бесконечно малых участков с длинами dx, и на каждом из них выберем некоторую произвольную точку x. Попадание значения X на каждый из этих участков – это фактически, в силу их бесконечной малости, попадание в соответствующую точку x, осуществляемое с вероятностью $dp=f(x)dx$ (см. (3.4)). Поэтому, в соответствии с формулой (1.4), получаем:

$$M(X)=x_{ср}=\sum xdp=\sum xf(x)dx=\int_a^b xf(x)dx \qquad (3.10)$$

То есть, получаем окончательно:

$$M(X)=x_{ср}=\int_a^b xf(x)dx \qquad (3.11)$$

Опираясь на полученную формулу для математического ожидания $M(X)$ непрерывной случайной величины X и используя определение (3.9) дисперсии $D(X)$ величины X, получим формулу для вычисления $D(X)$:

$$D(X)=\int_a^b [x-M(X)]^2 f(x)dx \qquad (3.12)$$

После раскрытия квадрата разности, разбиения интеграла (3.12) на три интеграла и учете равенств (3.8) и (3.11) получим так называемую *упрощенную формулу* для дисперсии $D(X)$:

$$D(X)=M(X^2)-[M(X)]^2 \qquad (3.13)$$

Здесь

$$M(X)=\int_a^b xf(x)dx;\quad M(X^2)=\int_a^b x^2 f(x)dx \qquad (3.14)$$

– математические ожидания величин X и X^2 соответственно. Кстати, упрощенная формула (3.13) для дисперсии $D(X)$ непрерывной случайной величины полностью совпадает с аналогичной формулой (1.25) для дисперсии $D(X)$ дискретной случайной величины. Формулы (3.9) для среднего квадратического отклонения $\sigma(X)$ и коэффициента вариации $V(X)\%$ непрерывной случайной величины тоже полностью совпадают с аналогичными формулами (1.21) и (1.27) для дискретной случайной величины. И смысл всех этих числовых характеристик для обеих случайных величин полностью совпадает.

Примечание. Если интервал $(a;b)$ возможных значений непрерывной случайной величины X бесконечен (в одну или в обе стороны), то интегралы

(3.12) и (3.14) будут несобственными и могут оказаться расходящимися – то есть не дадут конечных значений (дадут $+\infty$ или $-\infty$). Тогда $M(X)$, $D(X)$, а вместе с ними $\sigma(X)$, $V(X)\%$ могут не иметь конечных значений.

Пример 1. Непрерывная случайная величина X имеет возможные значения, заполняющие отрезок $[0;2]$ оси ox, а ее плотность вероятности – это линейная функция вида

$$f(x)=Cx \qquad (0\le x\le 2) \tag{3.15}$$

а) Найти C; б) Построить график функции $y=f(x)$; в) Найти $p(0\le X\le 1)$ и $p(1\le X\le 2)$; г) Найти $M(X)$, $D(X)$, $\sigma(X)$, $V(X)\%$.

Решение. а) Неизвестную константу C найдем из условия (3.8):

$$\int_0^2 Cx\,dx=1\Leftrightarrow C\frac{x^2}{2}\Big|_0^2=1\Leftrightarrow C\left(\frac{2^2}{2}-\frac{0^2}{2}\right)=1\Leftrightarrow 2C=1\Leftrightarrow C=\frac{1}{2}.$$

Таким образом, плотность вероятности

$$f(x)=\frac{1}{2}x \quad (0\le x\le 2) \tag{3.16}$$

б) Построим график функции $y=\frac{1}{2}x \quad (0\le x\le 2)$ – график плотности вероятности рассматриваемой непрерывной случайной величины X (рис. 2.7). Судя по этому графику, наиболее вероятным значением величины X (модой X) является значение $x=2$.

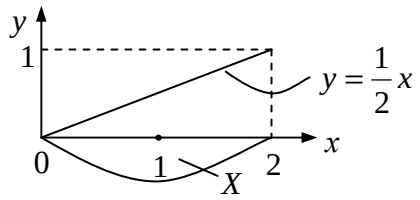


Рис. 2.7

в) Найдем вероятность $P(0\le X\le 1)$, то есть вероятность попадания значения X в левую половину отрезка $[0;2]$. Для этого используем формулу (3.6):

$$p(0\le X\le 1)=\int_0^1 \frac{1}{2}x\,dx=\frac{1}{2}\cdot\frac{x^2}{2}\Big|_0^1=\frac{1}{4}.$$

Аналогично найдем вероятность $p(1\le X\le 2)$ того, что X попадет в правую половину отрезка $[0;2]$:

$$p(1\le X\le 2)=\int_1^2 \frac{1}{2}x\,dx=\frac{1}{2}\cdot\frac{x^2}{2}\Big|_1^2=\frac{3}{4}.$$

Таким образом, вероятность попадания значения случайной величины X в правую половину $[1;2]$ отрезка $[0;2]$ втрое больше, чем вероятность ее попадания в левую половину $[0;1]$ этого же отрезка. Это произошло потому, что возможные значения x величины X правой половины отрезка $[0;2]$ вероятнее значений левой его половины (см. рис. 2.7). Отметим, что найденные вероятности $\frac{1}{4}$ и $\frac{3}{4}$ в сумме, как и положено, дают единицу.

г) Найдем числовые характеристики ($M(X)$; $D(X)$; $\sigma(X)$; $V(X)\%$) величины X:

$$M(X)=\left|см.(3.1)\right|=\int_0^2 x\cdot\frac{1}{2}x\,dx=\frac{1}{2}\cdot\frac{x^3}{3}\Big|_0^2=\frac{4}{3}=1\frac{1}{3}.$$

Таким образом, $M(X)=x_{ср}=1\frac{1}{3}$ – среднее значение величины X. Оно находится не в середине отрезка $[0;2]$, а правее, что совершенно естественно, ибо в правой половине отрезка $[0;2]$ содержатся более вероятные значения X, чем в его левой половине.

Теперь найдем дисперсию $D(X)$. Предварительно найдем $M(X^2)$:

$$M(X^2)=|см.(3.14)|=\int_0^2 x^2\cdot\frac{1}{2}xdx=\frac{1}{2}\cdot\frac{x^4}{4}\Big|_0^2=2.$$

Значит, согласно (3.13)

$$D(X)=2-\left(\frac{4}{3}\right)^2=2-\frac{16}{9}=\frac{2}{9}.$$

А теперь найдем оставшиеся две числовые характеристики величины X:

$$\sigma(X)=\sqrt{D(X)}=\sqrt{\frac{2}{9}}=\frac{\sqrt{2}}{3}\approx 0{,}47;$$

$$V(X)\%=\frac{\sigma(X)}{M(X)}\cdot 100\%\approx 35\%$$

Среднее квадратическое отклонение $\sigma(X)$ показывает, что среднее отклонение величины X от ее среднего значения $M(X)=x_{ср}=1\frac{1}{3}$ составляет (без учета знака отклонения) приблизительно 0,47. А коэффициент вариации $V(X)\%$ показывает, что по отношению к среднему значению это отклонение составляет приблизительно 35%.

Для непрерывных случайных величин, как и для дискретных, можно ввести понятия суммы и произведения.

Пусть X и Y – некоторые две непрерывные случайные величины, причем $[a;\ b]$ и $f_1(x)$ - промежуток возможных значений и плотность вероятности величины X, а $[c;\ d]$ и $f_2(y)$ - соответствующие характеристики величины Y. Тогда сумма $Z=X+Y$ – непрерывная случайная величина, которая в качестве своих возможных значений принимает всевозможные суммы $z=x+y$ $(a\le x\le b;\ c\le y\le d)$ значений величин X и Y. А произведение $U=XY$ - непрерывная случайная величина, которая в качестве своих возможных значений принимает всевозможные произведения $u=xy$ $(a\le x\le b;\ c\le y\le d)$ значений X и Y (рис. 2.8). Таким образом, возможные значения $z=x+y$ и $u=xy$ величин $Z=X+Y$ и $U=XY$ комбинируются из координат $(x;\ y)$ точек заштрихованного на рис.2.8 прямоугольника. А, следовательно, и плотности вероятности обеих этих величин комбинируются из плотностей вероятностей $f_1(x)$ и $f_2(y)$ величин X и Y. Комбинации эти выглядят сложно, особенно если случайные величины X и Y являются зависимыми друг от друга, поэтому приводить их не будем. Тем более, что для нахождения числовых характеристик случайных величин $Z=X\pm Y$ и $U=XY$ без них зачастую

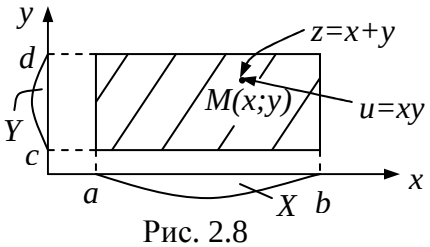


Рис. 2.8

можно и обойтись (это следует из приводимых ниже свойств сумм и произведений случайных величин).

Если случайные величины X и Y *независимы*, то каждая из них принимает свои возможные значения вне всякой связи с возможными значениями другой величины. Если же это не так, то случайные величины X и Y являются *зависимыми*. Кстати, смысл зависимости – независимости случайных величин полностью аналогичен смыслу зависимости – независимости случайных событий (см. §3 главы 1).

Для непрерывных случайных величин имеют место те же свойства, что и для дискретных случайных величин (§1). А именно:

1. Для любых непрерывных случайных величин X и Y
$$M(CX) = CM(X) \tag{3.17}$$
2. Для любых непрерывных случайных величин X и Y
$$M(X+Y) = M(X) + M(Y);\ \ M(X-Y) = M(X) - M(Y) \tag{3.18}$$

В частности, так как $M(C) = C$ для любой константы C (см.(1.5)), то
$$M(X+C) = M(X) + C \tag{3.19}$$
3. Если непрерывные случайные величины X и Y независимы, то
$$M(XY) = M(X) \cdot M(Y) \tag{3.20}$$
4. Для любой непрерывной случайной величины X и любой константы C
$$D(CX) = C^2 D(X) \tag{3.21}$$
5. Если непрерывные случайные величины X и Y независимы, то
$$D(X+Y) = D(X) + D(Y);\ \ D(X-Y) = D(X) + D(Y) \tag{3.22}$$

В частности, так как $D(C) = 0$ (см.(1.28)) и так как любая непрерывная случайная величина и любая константа независимы, то
$$D(X+C) = D(X) \tag{3.23}$$

Приведенные выше свойства для двух случайных величин переносятся и на несколько случайных величин:

1. Для любых непрерывных случайных величин (X_1; X_2; … X_p)
$$M(X_1 \pm X_2 \pm ... \pm X_p) = M(X_1) \pm M(X_2) \pm ... \pm M(X_p) \tag{3.24}$$
2. Если непрерывные случайные величины (X_1; X_2; … X_p) взаимно независимы, то
$$M(X_1 \cdot X_2 \cdots X_n) = M(X_1) \cdot M(X_2) \cdots M(X_p) \tag{3.25}$$
3. Если непрерывные случайные величины (X_1; X_2; … $X_\text{р}$) взаимно независимы, то
$$D(X_1 \pm X_2 \pm ... \pm X_p) = D(X_1) + D(X_2) + ... + D(X_p) \tag{3.26}$$

Упражнения

1. Когда непрерывная случайная величина считается заданной? Какие имеются ограничения при задании непрерывных случайных величин? Как называются основные числовые характеристики непрерывной случайной величины, как они находятся и каков их смысл?

2. Непрерывная случайная величина X может принять любое значение x от $-\infty$ до $+\infty$. Ее плотность вероятности $f(x) = Ce^{-|x|}$ $(-\infty < x < +\infty)$.

а) Найти C; б) Построить график функции $y = f(x)$; в) Найти числовые характеристики $(M(X),\ D(X),\ \sigma(X),\ V(X)\%)$ величины X.

Ответ: а) $C = \frac{1}{2}$; в) $M(X) = 0$; $D(X) = 2$; $\sigma(X) = \sqrt{2}$; $V(X)\%$ - не имеет смысла.

3. Муж и жена работают на сдельной работе и на разных предприятиях. То есть их месячные зарплаты X и Y – независимые случайные величины. Известно, что $M(X)$= 5000 руб.; $M(Y)$=4000 руб.; $\sigma(X)$=500 руб.; $\sigma(Y)$=200 руб. Найти для их совместной суммарной зарплаты $Z=X+Y$: а) $M(Z)$; б) $D(Z)$; в) $\sigma(Z)$; г) $V(Z)\%$.

Ответ: а) $M(Z)$=9000 руб.; б) $D(Z)$=290000 руб^2; в) $\sigma(Z)\approx$540 руб.; г) $V(Z)\% \approx 6\%$.

4. Плотность вероятности $y = f(x)$ непрерывной случайной величины X имеет следующий вид:

$$f(x) = \begin{cases} CSinx, & \text{если } x \in [0;\pi] \\ 0, & \text{если } x \notin [0,\pi] \end{cases}$$

а) Найти C; б) Построить график функции $y = f(x)$; в) Найти $p(0 \le X \le \frac{\pi}{2})$ и $p(\frac{\pi}{2} \le X \le \pi)$; г) Найти $MO(X)$; д) Найти $(M(X),\ D(X),\ \sigma(X),\ V(X)\%)$.

Ответ: а) $C = \frac{1}{2}$; в) $p(0 \le X \le \frac{\pi}{2}) = \frac{1}{2}$ и $P(\frac{\pi}{2} \le X \le \pi) = \frac{1}{2}$;

г) $MO(X) = \frac{\pi}{2}$; д) $M(X) = \frac{\pi}{2}$; $D(X) = \frac{\pi^2}{4} - 2$; $\sigma(X) \approx 0{,}68$; $V(X)\% \approx 43\%$.

5. Доказать, что математическое ожидание непрерывной случайной величины заключено между наименьшим и наибольшим ее возможными значениями.

6. Доказать, что если X и Y – любые две независимые случайные величины (дискретные или непрерывные), то

$$D(XY) = D(X)\cdot D(Y) + [M(X)]^2 \cdot D(X) + [M(Y)]^2 \cdot D(Y).$$

§4. Некоторые важнейшие распределения непрерывных случайных величин.

1. Равномерное распределение.

Определение. *Непрерывная случайная величина X называется распределенной равномерно, если все ее возможные значения равновероятны.*

Это значит, что если X - равномерно распределенная случайная величина, и $[a;b]$– промежуток ее возможных значений, то плотность вероятности $f(x)$ величины X для всех $x \in [a;b]$ одинакова:

$$f(x) = C \quad (a \le x \le b). \tag{4.1}$$

Константа C не может быть произвольной. Ее значение должно быть таким, чтобы обеспечивалось условие (3.8). Из него следует:

$$\int_a^b C dx = 1 \Rightarrow C = \frac{1}{b-a} \tag{4.2}$$

Таким образом, равномерно распределенная на отрезке $[a;b]$ непрерывная случайная величина X имеет на нем следующую плотность вероятности (рис. 2.9):

$$f(x) = \frac{1}{b-a} \ (a \le \mathrm{x} \le b) \tag{4.3}$$

А числовые характеристики этой величины таковы (получите их самостоятельно):

$$M(X) = \frac{a+b}{2}; \quad D(X) = \frac{(b-a)^2}{12}; \quad \sigma(X) = \frac{b-a}{2\sqrt{3}}; \quad V(X)\% = \frac{\sqrt{3}(b-a)}{3(b+a)} \cdot 100\% \tag{4.4}$$

Равномерно распределенными будут, например, следующие величины:

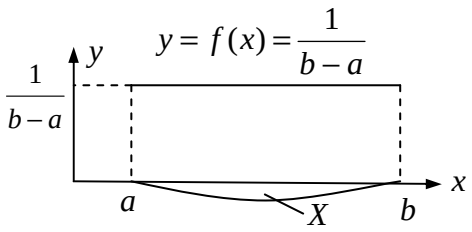


Рис. 2.9

1) Координата X точки, бросаемой наудачу на отрезок $[a;b]$;

2) Ошибка X, производимая при округлении приближенных чисел до нужного десятичного знака при производстве арифметических вычислений;

3) Время T ожидания пассажиром троллейбуса, маршрутного такси, поезда метро и прочих средств общественного транспорта при регулярном графике их движения. И т.д.

Пример 1. При производстве арифметических вычислений их результаты округляют до сотых. Какова вероятность того, что ошибка X, допущенная при округлении очередного числа, окажется не больше одной тысячной (без учета ее знака)?

Решение. Ошибка X округления чисел до сотых – равномерно распределенная случайная величина, принимающая свои возможные значения (если не учитывать их знак) на промежутке [0; 0,005]. Поэтому ее плотность вероятности, согласно (4.3), будет иметь вид:

$$f(x) = 200 \quad (0 \le x \le 0,005).$$

А тогда, используя формулу (3.6), получим и искомую вероятность:

$$p(0 \le X \le 0,001) = \int_0^{0,001} 200 dx = 200x \Big|_0^{0,001} = 0,2$$

2. Нормальное распределение.

Пусть случайная величина X может быть представлена в виде:

$$X = a + \sum X_k \tag{4.5}$$

Здесь a - некоторая константа, а $\sum X_k$ - сумма очень большого (бесконечно большого) числа независимых случайных величин $X_к$, влияние каждой из которых на всю эту сумму ничтожно мало.

Примерно такая ситуация, в частности, сложится, если X – результат измерения некоторым прибором некоторой величины a (например, X – результат взвешивания на весах некоторой массы a). Действительно, результат X измерения не обязательно совпадет с истинным значением a измеряемой величины, ибо к этому истинному значению добавляются ошибки X_k, вносимые:

а) прибором; б) измерителем; в) внешней средой.

Такого рода помех измерению в принципе бесконечно много, и каждая из них вносит свою ошибку X_k в результат X измерения.

Если бы ошибок X_k не было (а значит, не было бы помех измерению), то результат измерения X совпал бы с a - с истинным значением измеряемой величины. И, таким образом, он не был бы случайной величиной. Но от окружающей среды защититься полностью невозможно. А потому ошибки X_k измерения, вызываемые помехами, неизбежны при любом измерении. Все эти ошибки X_k, вообще говоря, мелкие. Но их очень много, ибо очень много внешних факторов, влияющих на процесс измерения. Действуют эти факторы, как правило, независимо друг от друга. Поэтому результат X измерения –случайная величина, которая в итоге и принимает вид (4.5).

Будем считать, что среди ошибок измерения нет ошибок систематических. То есть нет ошибок, приводящих к систематическому завышению или занижению результатов измерения (из-за неотрегулированности прибора, недобросовестности измерителя, и т.д.). То есть все ошибки $X_к$ связаны исключительно со случайными помехами измерению. Тогда величина X (результат измерения) с одинаковой вероятностью может оказаться как больше a, так и меньше a. При этом, в соответствии с формулой (4.5), она теоретически может принять значение x, как угодно далеко отстоящее от a в ту или другую сторону. Но чем дальше оно будет отстоять от a, тем менее вероятным оно будет. Действительно, большое отклонение x от a - это большая итоговая ошибка измерения. А большие ошибки измерения, естественно, менее вероятны, чем малые. Очень же большие ошибки практически невероятны.

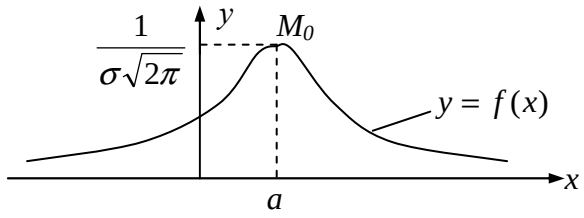


Рис. 2.10

Из всего сказанного выше вытекает, что случайная величина X, имеющая структуру (4.5) – это непрерывная случайная величина, возможные значения которой заполняют всю числовую ось ox. А её плотность вероятности $y = f(x)$ имеет в принципе графический вид, изображенный на рис. 2.10.

Российский математик Ляпунов в начале 20-го века вывел аналитический вид (формулу) для этой плотности вероятности:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}} \quad (-\infty \prec x \prec \infty) \qquad (4.6)$$

Здесь a - мода и одновременно математическое ожидание (среднее значение) величины X, а σ – ее среднее квадратическое отклонение (среднее отклонение X от a без учета знака отклонения). То есть

$$MO(X) = M(X) = a; \quad \sigma(X) = \sigma; \quad D(X) = \sigma^2 \tag{4.7}$$

Если X – результат измерения величины a, то в формулах (4.7) a - это истинное значение измеряемой величины (обычно неизвестное и подлежащее определению), а σ – среднее значение итоговой (суммарной) ошибки измерения, зависящее как от точности измерительного прибора, так и от условий измерения. Эта ошибка обычно определяется опытным путем.

Случайная величина X, представимая в виде (4.5), называется *распределенной нормально*. То есть такая величина X *имеет нормальное распределение*. Или, что то же самое, *величина X распределена по нормальному закону*.

Нормально распределенная случайная величина X может в принципе принять любое значение от $-\infty$ до $+\infty$, а ее плотность вероятности имеет вид (4.6).

Результаты различного рода измерений – это случайные величины, распределенные приблизительно по нормальному закону. Приблизительно нормально распределенными случайными величинами будут и итоговые (суммарные) ошибки этих измерений, ибо они имеют вид (4.5) при a=0. И вообще, приблизительно нормально распределенными будут все случайные величины X, которые имели бы в каждом испытании неизменное значение a, если бы не незначительные случайные добавки, которых много и которые “уводят” значение величины X от a в большую или в меньшую сторону.

Например, если станок-автомат настроен на изготовление деталей определенного размера a, то реальный размер X изготавливаемых деталей, в силу различного рода случайных факторов, будет иметь вид (4.5) и, следовательно, будет случайной величиной, распределенной приблизительно нормально.

Приблизительно нормально распределенными случайными величинами можно также считать суточный привес одного или группы животных при постоянном рационе их кормления; суточный надой одной или нескольких коров при одних и тех же условиях их содержания; расход семян на единицу засеваемой площади при неизменной технологии сева; ежедневный объем продаж магазина при неизменном в целом спросе на продукцию этого магазина, и т.д. С такого рода случайными величинами чаще всего приходится иметь дело на практике. Отсюда возник и термин для их названия – “нормально распределенные”.

Нормально распределенные случайные величины обладают важным свойством (мы докажем его в §5): если X – нормально распределенная величина с некоторыми параметрами $(a;\sigma)$, то любая линейная комбинация $Z = C_1X + C_2 \ \ (C_1 \neq 0)$ - тоже нормально распределенная случайная величина, имеющая параметры ($a_* = C_1a + C_2; \ \sigma_* = |C_1|\sigma$). В частности, нормально распределенная случайная величина

$$Z = \frac{X - a}{\sigma} \tag{4.8}$$

имеет параметры ($a_* = 0;\ \sigma_* = 1$) и называется *нормированной нормально распределенной случайной величиной*. Ее плотность вероятности $f(x)$, согласно (4.6), имеет вид

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (-\infty \prec x \prec \infty) \tag{4.9}$$

То есть она представляет собой функцию Гаусса (см. (6.3), глава 1). Ее график представлен на рис. 1.4 (стр.35).

Вид графика плотности вероятности $y = f(x)$ нормально распределенных случайных величин *X*, согласно рис. (2.10), зависит от величины их параметров ($a;\sigma$). А именно, значение a определяет вертикальную ось симметрии этого графика, а величина σ определяет высоту графика: чем σ меньше, тем вершина M_0 графика выше. А чем график выше, тем он уже. Сужение графика плотности вероятности $y = f(x)$ с увеличением его высоты связано с тем, что площадь между осью *ox* и графиком плотности вероятности $y = f(x)$ для любой непрерывной случайной величины, согласно (3.8), должна быть равна единице. Поэтому если этот график увеличивает свою высоту, то он автоматически становится уже (рис.2.11):

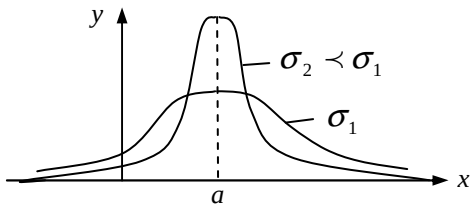


Рис. 2.11

Рассмотрим теперь следующий важный для практики вопрос: какова вероятность того, что в результате испытания нормально распределенная случайная величина *X* с параметрами (a; σ) примет значение в некотором заданном числовом промежутке $[x_1; x_2]$? Ответ на этот вопрос мы получим с помощью формулы (3.6), которая приводит к следующему результату:

$$p(x_1 \le X \le x_2) = \Phi\left(\frac{x_2 - a}{\sigma}\right) - \Phi\left(\frac{x_1 - a}{\sigma}\right) \tag{4.10}$$

Здесь $\Phi(x)$ – уже известный нам интеграл вероятностей (см. главу 1, формулу (6.9) и рис. 1.5). Действительно,

$$p(x_1 \le X \le x_2) = \int_{x_1}^{x_2} f(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{(x-a)^2}{2\sigma^2}} dx = \left| \begin{array}{l} \text{сделаем подстановку: } \dfrac{x-a}{\sigma} = t; \\ \text{тогда } x = \sigma t + a;\ \ dx = \sigma dt; \end{array} \right| =$$

$$= \frac{1}{\sqrt{2\pi}} \int_{\frac{x_1-a}{\sigma}}^{\frac{x_2-a}{\sigma}} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} \int_{\frac{x_1-a}{\sigma}}^{0} e^{-\frac{t^2}{2}} dt + \frac{1}{\sqrt{2\pi}} \int_{0}^{\frac{x_2-a}{\sigma}} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} \int_{0}^{\frac{x_2-a}{\sigma}} e^{-\frac{t^2}{2}} dt - \frac{1}{\sqrt{2\pi}} \int_{0}^{\frac{x_1-a}{\sigma}} e^{-\frac{t^2}{2}} dt =$$

$$= \left| \text{учтем, что } \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt \right| = \Phi\left(\frac{x_2 - a}{\sigma}\right) - \Phi\left(\frac{x_1 - a}{\sigma}\right)$$

На основании полученной формулы (4.10) следует (выведите это самостоятельно), что для любого $\delta \succ 0$

$$p(|X - a| \le \delta) = p(a - \delta \le X \le a + \delta) = 2\Phi\left(\frac{\delta}{\sigma}\right) \tag{4.11}$$

Геометрическая иллюстрация последнего равенства дана на рис.2.12.

В частности, используя формулу (4.11) и таблицу функции $\Phi(x)$ (фрагмент этой таблицы приведен в (6.10), глава 1), получим:

$$p(|X-a| \le \sigma) = p(a-\sigma \le X \le a+\sigma) = 2\Phi(1) = 0{,}6826$$
$$p(|X-a| \le 2\sigma) = p(a-2\sigma \le X \le a+2\sigma) = 2\Phi(2) = 0{,}9544 \qquad (4.12)$$
$$p(|X-a| \le 3\sigma) = p(a-3\sigma \le X \le a+3\sigma) = 2\Phi(3) = 0{,}9973 \approx 1$$

Последний из этих результатов означает, что *практически наверняка* экспериментальное значение нормально распределенной случайной величины X будет содержаться в пределах $[a-3\sigma; a+3\sigma]$. То есть событие, состоящее том, что экспериментальное значение нормально распределенной случайной величины X отклонится от её среднего значения a в ту или другую сторону больше, чем на 3σ (на три средних отклонения от среднего), практически невероятно. Этот вывод называют *правилом трех сигм*.

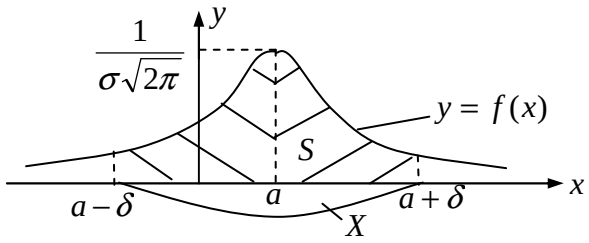


Рис. 2.12

Например, если X – результат взвешивания некоторой массы a, то $\sigma = \sigma(X)$ представляет собой среднюю ошибку взвешивания на данных весах. И если, например, σ = 5 г, то в соответствии с правилом трех сигм практически невероятно, что случайная ошибка взвешивания окажется больше 15 г (в ту или в другую сторону).

Пример 2. Станок-автомат изготавливает шарики. Шарик считается годным, если отклонение X – диаметра шарика от проектного размера по абсолютной величине не превосходит 0,7 мм. Опытным путем установлено, что $\sigma = \sigma(X)$ =0,4 мм. Определить, сколько процентов годных шариков изготовляет автомат.

Решение. Отклонение X диаметра шарика от проектного размера – это ошибка диаметра шарика. При условии, что станок настроен правильно (систематических ошибок в работе станка нет), эта ошибка X складывается из множества мелких случайных ошибок X_k, связанных с действием различных независимых друг от друга случайных факторов. То есть $X = \sum X_k$. Таким образом, ее выражение соответствует формуле (4.5) при a = 0. А значит, X – случайная величина, распределенная нормально с математическим ожиданием (средним значением) a = 0 и средним квадратическим отклонением σ = 0,4 мм (дано по условию задачи). Величина σ = 0,4 мм означает, что среднее значение отклонения ошибки X от ее среднего значения a = 0 без учета знака отклонения, то есть среднее значение самой этой ошибки, если не учитывать ее знак, составляет 0,4 мм. Иначе говоря, станок работает со средней ошибкой 0,4 мм (в ту или другую сторону).

По условию задачи, каждый изготовленный станком-автоматом шарик будет считаться годным, если ошибка X его диаметра будет удовлетворять нера-

венству $|X| \le 0{,}7$ мм, то есть она будет в пределах [-0,7 мм; 0,7 мм]. Найдем вероятность этого. Применяя формулу (4.11), получаем:

$$p(|X| \le 0{,}7) = p(-0{,}7 \le X \le 0{,}7) = 2\Phi\left(\frac{0{,}7}{0{,}4}\right) = 2\Phi(1{,}75) = 2 \cdot 0{,}4599 \approx 0{,}92$$

Итак, вероятность того, что каждый изготовленный станком-автоматом шарик окажется годным, составляет 0,92. Это значит, что этот станок изготовляет в среднем 92% годных шариков.

3. Распределение χ^2 (хи-квадрат).

Пусть $X_i\ (i = 1, 2, \ldots k)$ – независимые нормированные нормально распределенные случайные величины (то есть у каждой из них $a = 0$ и $\sigma = 1$). Тогда сумма квадратов этих величин

$$\chi^2 = \sum_{i=1}^{k} X_i^2 \qquad (4.13)$$

является случайной величиной, распределенной *по закону χ^2 (хи-квадрат) с k степенями свободы*. Понятие «число степеней свободы» заимствовано из физики, где под ним понимается число независимых координат, определяющих положение тел в процессе их движения. В частности, при движении по прямой точка обладает одной степенью свободы; при свободном движении по плоскости – двумя степенями свободы; при свободном движении в пространстве – тремя степенями свободы. Если же движение точки несвободно – например, она может двигаться по плоскости лишь вдоль какой-то линии L с уравнением $y = f(x)$, то ее положение на плоскости полностью определяется лишь одной координатой *x*, ибо другая координата *y* уже не является независимой, так как выражается через *x*. В этом случае у движущейся по плоскости точки будет не две, а одна степень свободы. И так далее. Возвращаясь к выражению (4.13) для случайной величины χ^2, видим, что в ее образовании участвуют k независимых величин $X_i\ (i = 1, 2, \ldots k)$. Поэтому у нее и k степеней свободы.

Очевидно, что величина χ^2 может принимать лишь неотрицательные значения $x \ge 0$. Доказано, что плотность вероятности случайной величины χ^2 с k степенями свободы имеет вид

$$f_k(x) = A_k e^{-\frac{x}{2}} x^{\frac{k}{2}-1}, \quad (x \ge 0) \qquad (4.14)$$

где A_k - быстро убывающие с номером k числовые коэффициенты, выражения которых ввиду их сложности не приводим. При этом график плотности вероятности $y = f_k(x)$ для k=1, k=2 и k>2 в принципе имеет вид:

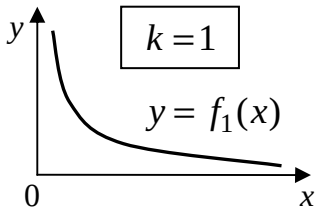


Рис. 2.13(а)

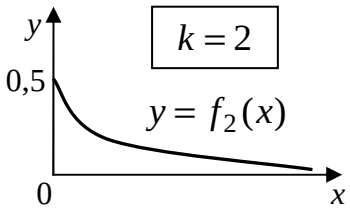


Рис. 2.13(б)

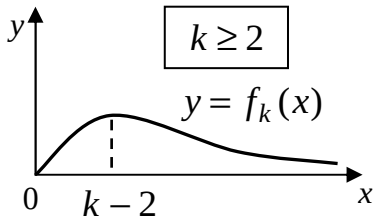


Рис. 2.13(в)

Также доказано, что

$$M(\chi^2) = k; \;\; D(\chi^2) = 2k \qquad (4.15)$$

С увеличением числа k степеней свободы распределение χ^2 медленно приближается к нормальному.

4. Распределение Стьюдента (T - распределение).

Пусть Z – нормированная нормально распределенная случайная величина ($M(Z) = 0;\ \sigma(Z) = 1$), а V – независимая от Z величина, распределенная по закону χ^2 с k степенями свободы. Тогда величина

$$T = \frac{Z}{\sqrt{\frac{V}{k}}} \qquad (4.16)$$

имеет распределение, которое называют T- распределением, или распределением Стьюдента (псевдоним английского статистика Госсета) с k степенями свободы. Очевидно, что случайная величина T может принимать любые значения t от $-\infty$ до $+\infty$. А ее плотность вероятности $f_k(t)$ имеет вид:

$$f_k(t) = B_k\left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}} \quad (-\infty \prec t \prec \infty) \qquad (4.17)$$

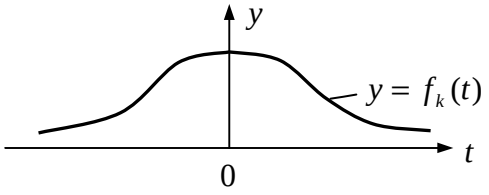


Рис. 2.14

График плотности вероятности $y = f_k(t)$ случайной величины T, имеющей распределение Стьюдента с k степенями свободы, для любого $k \geq 1$ в принципе имеет вид, изображенный на рис 2.14. Эта кривая, если сравнивать её с плотностью вероятности нормированного нормального распределения (с функцией Гаусса, рис. 1.4), ниже последней и убывает медленнее.

Очевидно, что $M(T) = t_{ср} = 0$. При этом (можно доказать) дисперсия $D(T) = \frac{k}{k-2}$. При $k \to \infty$

$$f_k(t) \to \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \quad (-\infty \prec t \prec \infty) \qquad (4.18)$$

То есть при $k \to \infty$ распределение Стьюдента стремится к нормированному нормальному распределению. Причем стремится быстро, так что при k>30 вместо распределения Стьюдента можно пользоваться нормированным нормальным распределением.

5. Распределение Фишера-Снедекора (F-распределение).

Если U и V – независимые случайные величины, распределенные по закону χ^2 со степенями свободы k_1 и k_2 соответственно, то случайная величина

$$F = \frac{\frac{U}{k_1}}{\frac{V}{k_2}} \tag{4.19}$$

имеет распределение, которое называют *F-распределением, или распределением Фишера-Снедекора с (k_1; k_2) степенями свободы*. Величина F принимает, очевидно, лишь неотрицательные значения $x \geq 0$ и имеет плотность вероятности вида

$$f_{k_1,k_2}(x) = C_{k_1,k_2} \cdot \frac{x^{\frac{k_1-2}{2}}}{(k_2 + k_1 x)^{\frac{k_1+k_2}{2}}} \tag{4.20}$$

Выражение для коэффициентов C_{k_1,k_2} ввиду его сложности не приводим. Форма графика функции $y = f_{k_1,k_2}(x)$ в принципе такова:

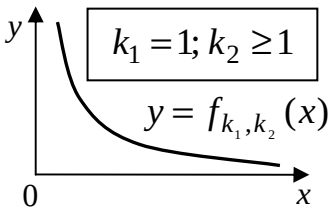


Рис. 2.15(а)

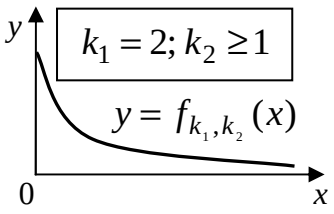


Рис. 2.15(б)

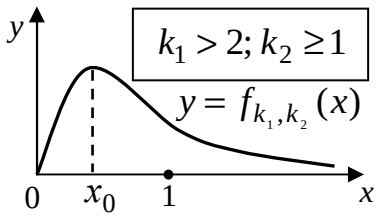


Рис. 2.15(в)

При этом мода x_0 случайной величины F для $k_1 \geq 2$ *и* $k_2 \geq 1$ определяется формулой:

$$x_0 = \frac{k_1 - 2}{k_1} \cdot \frac{k_2}{k_2 + 2} \tag{4.21}$$

Таким образом, x_0 всегда меньше 1 (см. рис.2.15(в)) и приближается к ней при $k_1 \to \infty$ и $k_2 \to \infty$ одновременно.

6. Показательное распределение. Функция надежности.

Пусть некоторый природный или искусственный объект начинает функционировать в момент времени t=0. В качестве такого объекта можно, например, рассматривать живое существо с момента его рождения или с любого другого этапного для него момента; работающий механизм или элемент этого механизма с момента его включения, и т.д.

Пусть T – время безаварийного функционирования этого объекта (для живого существа это может быть время до утраты пригодности его к тому или иному виду деятельности или до его смерти; для механизма это может быть время до первой его поломки или до окончательного выхода его из строя, и т.д.). Согласно своему смыслу, T – непрерывная случайная величина, возможные значения t которой могут быть, в принципе, любыми неотрицательными числами: $0 \leq t \leq \infty$.

Поставим теперь естественный вопрос: какова вероятность $p(T \geq t)$ того, что для данного объекта будет иметь место неравенство $T \geq t$, где t – некоторое заданное время? То есть поставим вопрос: какова вероятность того, что за время t функционирующий объект не выйдет из строя? Вероятность эту обозначим символом *R(t)* и назовем *функцией надежности:*

$$R(T) = p(T \geq t) \quad (t \geq 0) \tag{4.22}$$

Очевидно, что для любого функционирующего объекта указанная вероятность *R(t)* сохранения своей работоспособности в течение времени t будет убывать с возрастанием t, начиная с 1 при t=0, и стремиться к нулю при $t \to \infty$. Более того, как показали многочисленные практические исследования самых разнообразных объектов (технических устройств; природных образований; живых организмов и т.д.) это убывание функции надежности *R(t)* осуществляется приблизительно по показательному (экспоненциальному) закону

$$R(t) = e^{-\lambda t} \quad (t \geq 0;\ \lambda \succ 0) \tag{4.23}$$

Такое поведение функции надежности *R(t)* называют еще *показательным законом надежности* (рис.2.16).

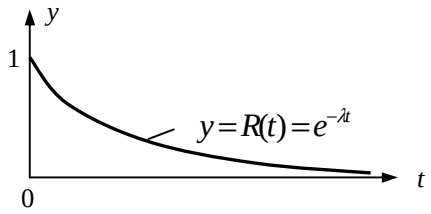


Рис. 2.16

Выясним смысл параметра λ в случае показательного закона надежности. Для этого найдем плотность вероятности $f(t)$ и основные числовые характеристики ($M(T)$; $D(T)$; $\sigma(T)$; $V(T)\%$) случайной величины T.

Начнем с нахождения $f(t)$. Ее будем искать, исходя из определения (3.1) плотности вероятности непрерывной случайной величины.

Пусть t – некоторое фиксированное значение величины T (t – момент выхода объекта из строя). Окружим это значение некоторым частичным промежутком $[t_1; t_2 = t_1 + \Delta t]$ длиной Δt и рассмотрим вероятность Δp того, что случайная величина T примет значение внутри этого промежутка. То есть рассмотрим вероятность Δp того, что объект выйдет из строя в какой-то момент времени, принадлежащий этому промежутку (рис.2.17).

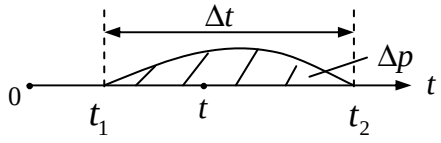


Рис. 2.17

Согласно смыслу функции надежности, *R(t₁)* и *R(t₂)* – это вероятности того, что объект выйдет из строя после моментов времени t_1 и t_2 соответственно (рис.2.18). Вероятность *R(t₁)* больше вероятности *R(t₂)* на Δp, откуда следует:

$$\Delta p = R(t_1) - R(t_2) \tag{4.24}$$

Тогда по формуле (3.1) получаем:

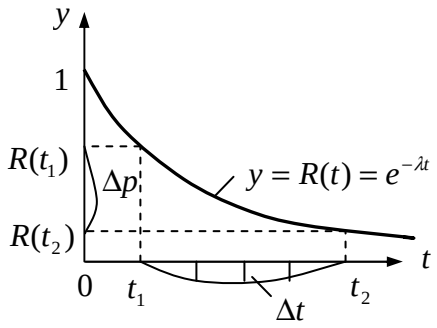


Рис. 2.18

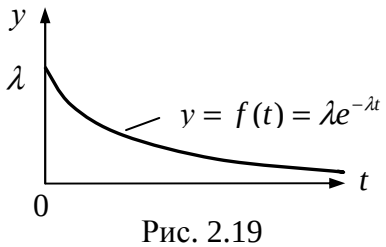


Рис. 2.19

$$f(t) = \lim_{\Delta t \to 0} \frac{\Delta p}{\Delta t} = \lim_{\substack{\Delta t \to 0 \\ (t_1 \to t; t_2 \to t)}} \frac{R(t_1) - R(t_2)}{\Delta t} = -\lim_{\substack{\Delta t \to 0 \\ (t_1 \to t; t_2 \to t)}} \frac{R(t_2) - R(t_1)}{\Delta t} = -\lim_{\substack{\Delta t \to 0 \\ (t_1 \to t)}} \frac{R(t_1 + \Delta t) - R(t_1)}{\Delta t} =$$

$$= -R'(t) = \lambda e^{-\lambda t} \quad (t \geq 0).$$

Итак,

$$f(t) = \lambda e^{-\lambda t} \tag{4.25}$$

– плотность вероятности случайной величины T, представляющей собой время безаварийного функционирования объекта, имеющего показательный закон надежности. График этой плотности вероятности изображен на рис.2.19. Кстати, распределение указанной случайной величины T носит название *показательного распределения*.

Зная плотность вероятности $f(t)$, можем теперь по формулам (3.9)-(3.14) найти и числовые характеристики ($M(T)$; $D(T)$; $\sigma(T)$; $V(T)\%$) случайной величины T (получите их самостоятельно):

$$M(T) = t_{cp} = \frac{1}{\lambda}, \text{ откуда } \lambda = \frac{1}{M(T)} = \frac{1}{t_{cp}};$$
$$D(T) = t_{cp}^2; \ \sigma(T) = t_{cp}; \ \mathrm{V(T)\%} = 100\% \tag{4.26}$$

Здесь t_{cp} – среднее время безаварийного функционирования объекта. Через t_{cp} можно выразить и функцию надежности (4.23) случайной величины T, и плотность ее вероятности (4.25):

$$R(t) = e^{-\frac{t}{t_{cp}}}; \quad f(t) = \frac{1}{t_{cp}} \cdot e^{-\frac{t}{t_{cp}}} \quad (t \geq 0) \tag{4.27}$$

В заключение отметим следующий важный факт: вероятность безаварийной работы любого объекта на интервале времени длительности t, если время T безаварийной работы этого объекта имеет показательное распределение, *не зависит от начала рассматриваемого интервала, а зависит только от его длительности t.*

Для доказательства этого утверждения введем следующие события (рис. 2.20):

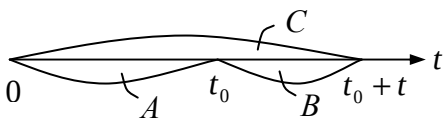


Рис. 2.20

A – безотказное функционирование объекта на интервале времени $(0; t_0)$;

B – безотказное функционирование объекта на интервале времени $(t_0; t_0 + t)$;

C – безотказное функционирование объекта на интервале времени $(0;t_0+t)$:
Очевидно, что $C=AB$, откуда следует:

$$p(C)=p(AB)=\left|\text{события } A \text{ и } B \text{ зависимые}\right|=p(A)\cdot p(B/A) \qquad (4.28)$$

Таким образом

$$p(B/A)=\frac{p(C)}{p(A)}=\frac{R(t_0+t)}{R(t_0)}=\frac{e^{-\frac{t_0+t}{t_{cp}}}}{e^{-\frac{t_0}{t_{cp}}}}=e^{-\frac{t}{t_{cp}}}=R(t), \qquad (4.29)$$

где $R(t)$ – вероятность безотказного функционирования объекта на интервале времени $(0;t)$, а $p(B/A)$ – вероятность безотказного функционирования объекта на интервале времени $(t_0;t_0+t)$. Эти вероятности равны, что и доказывает заявленный выше факт.

Отметим, что случайные величины, имеющие показательное распределение (показательный закон надежности), тесно связаны с событиями простейшего (пуассоновского) потока. Действительно, согласно формуле (2.8) главы 2 вероятность того, что за время t не появится ни одного из событий простейшего потока, найдется по формуле:

$$p_t(0)=e^{-\lambda t} \quad (t\geq 0) \qquad (4.30)$$

Здесь λ - интенсивность пуассоновского потока (среднее число событий потока, появляющихся за единицу времени). Тогда $\frac{1}{\lambda}=t_{cp}$ – среднее время, проходящее между появлениями отдельных событий потока. С учетом этого вероятность (4.30) примет вид:

$$p_t(0)=e^{-\frac{t}{t_{cp}}} \quad (t\geq 0) \qquad (4.31)$$

Но точно такой же вид, согласно (4.27), имеет показательная функция надежности $R(t)$, определяющая вероятность безаварийной работы объекта в течение времени t. Таким образом, время T безаварийной работы объекта при показательном законе надежности и время T, проходящее между соседними событиями простейшего потока, имеют одно и то же распределение. А именно, показательное распределение с плотностью вероятности $f(t)$, описываемой формулой (4.27).

Пример 3. Среднее время безотказной работы некоторого устройства, имеющего показательный закон надежности, равно 50 часам. Определить вероятность того, что устройство безотказно проработает 100 часов.

Решение. Пусть T – время безотказной работы устройства. Так как среднее значение этого времени t_{cp}=50 часам, то функция надежности $R(t)$ для рассматриваемого устройства имеет, согласно (4.27), вид:

$$R(t)=e^{-\frac{t}{50}} \quad (t-\text{время в часах}, \quad t\geq 0)$$

Тогда, согласно (4.22), получаем искомую вероятность:

$$p(T\geq 100)=R(100)=e^{-2}\approx 0{,}135$$

Упражнения.

1. Интервал движения троллейбусов составляет 5 минут. Какова вероятность того, что пришедшему на остановку пассажиру придется ожидать очередного троллейбуса не менее трех минут?

Ответ: 0,4.

2. Производится взвешивание некоторого вещества без систематических ошибок со средней случайной ошибкой 20 г (в ту или другую сторону). Найти вероятность того, что взвешивание будет произведено с ошибкой, не превосходящей по абсолютной величине 10 г.

Ответ: $2\Phi(0,5)\approx 0,383$.

3. Норма высева семян на 1 га равна 200 кг. Фактический расход семян на 1 га колеблется около этого значения со средним квадратическим отклонением 10 кг. Определить количество семян, обеспечивающих посев на площади 100 га с гарантией (вероятностью) 0,95.

Ответ: 21645 кг.

4. Станок-автомат производит цилиндрические болванки. Проектный размер диаметра болванок составляет 100 мм. Известно, что станок производит в среднем 2 % болванок диаметром более 101 мм. Болванка считается годной, если ее диаметр находится в пределах от 99 мм до 101 мм. Сколько процентов годных болванок производит станок-автомат?

Ответ: 96%.

5. Испытывают два независимо работающих устройства. Длительность безотказной работы обоих устройств имеет показательное распределение. Среднее время безотказной работы первого устройства составляет 40 часов, второго 20 часов. Найти вероятность того, что в течение 10 часов:

а) не откажет первое устройство;

б) не откажет второе устройство;

в) оба устройства не откажут;

г) оба устройства откажут;

д) хотя бы одно устройство не откажет.

Ответ: а) 0,78; б) 0,61; в) 0,47; г) 0,09; д) 0,91.

6. В городе рождается в среднем 5 детей в сутки. Считая рождения детей событиями, составляющими простейший поток событий, найти:

а) математическое ожидание;

б) среднее квадратическое отклонение;

в) коэффициент вариации случайной величины T – времени между последовательными рождениями детей.

Ответ: а) $M(T)$=4ч. 48мин.; б) $\sigma(T)$=4ч. 48мин.; в) $V(T)\%$=100%.

§5. Функция случайной величины.

Пусть Y – случайная величина, значение которой полностью определяется значением другой случайной величины X. То есть случайная величина Y является *функцией случайной величины* X:

$$Y = \varphi(X) \tag{5.1}$$

Например, если изготавливаются цилиндрические болванки, и X – диаметр болванок (X – случайная величина), то

$$Y = \varphi(X) = \frac{\pi X^2}{4} \tag{5.2}$$

- площадь поперечного сечения этих болванок, которая тоже является случайной величиной. И эта величина Y – функция величины X. Аналогично доход $Y=pX$ от ежедневных продаж товара – функция от объема X проданного товара, где p – цена единицы товара. И т.д.

Пусть известно распределение случайной величины X (аргумента). Спрашивается, каким будет распределение случайной величины $Y = \varphi(X)$ - функции величины X? Ответ на этот вопрос, очевидно, должен быть отдельным для дискретных и отдельным для непрерывных случайных величин.

1. Пусть X – дискретная случайная величина, а таблица (5.3) – закон ее распределения:

X	x_1	x_2	…	x_n
p	p_1	p_2	…	p_n

(5.3)

Тогда $Y = \varphi(X)$ - тоже дискретная случайная величина, а таблица (5.4) – закон её распределения:

$Y = \varphi(X)$	$\varphi(x_1)$	$\varphi(x_2)$	…	$\varphi(x_n)$
p	p_1	p_2	…	p_n

(5.4)

Действительно, значениями случайной величины $Y = \varphi(X)$ будут значения $\varphi(x_1); \varphi(x_2); \ldots \varphi(x_n)$, а вероятности $p_1; p_2; \ldots p_n$ у этих значений будут теми же, что и у значений $x_1; x_2; \ldots x_n$. Последние обстоятельство следует из того факта, что события $X = x_i$ и $Y = \varphi(X) = \varphi(x_i)$ ($i=1,2,\ldots n$) наступают или не наступают одновременно. Значит, они имеют одинаковые вероятности.

Пример 1. Случайная величина X имеет следующий закон распределения:

X	-1	0	1	2
p	0.5	0.2	0.2	0.1

Составить закон распределения случайной величины $Y=X^2$.

Решение. Согласно (5.4) имеем:

$Y = X^2$	1	0	1	4
p	0.5	0.2	0.2	0.1

Суммируя вероятности совпадающих значений, получим окончательно:

$Y = X^2$	0	1	4
p	0.2	0.7	0.1

После того, как закон распределения функции $Y = \varphi(X)$ составлен, можно, при желании, найти и её числовые характеристики *M(Y), D(Y),* σ*(Y) , V(Y)*%.

2. Пусть теперь *X* - непрерывная случайная величина, распределение которой задаётся интервалом (*a; b*) её возможных значений и плотностью вероятности *f (x)* для $x \in (a;b)$. Тогда $Y = \varphi(X)$ - тоже непрерывная случайная величина. Распределение величины *Y* будет известно, если мы найдём интервал (*c; d*) её возможных значений *y* и плотность вероятности *g(y)* для $y \in$ (*c; d*).

Сразу отметим, что интервал (*c ; d*), который содержит все возможные значения *y* случайной величины $Y = \varphi(X)$, представляет собой, очевидно, область значений функции $y = \varphi(x)$ $(a<x<b)$. А плотность вероятности *g(y)* величины $Y = \varphi(X)$ будем искать, исходя из следующих соображений.

Пусть *x* - некоторое возможное значение величины *X* ($a<x<b$). Тогда $y = \varphi(x)$ ($c<y<d$) – соответствующее ему значение величины *Y*. Окружим точку *x* некоторым бесконечно малым отрезком длиной *dx*. Тогда этому отрезку, окружающему точку *x*, будет соответствовать некоторой бесконечно малый отрезок длиной $dy = |\varphi'(x)dx| = |\varphi'(x)|dx$, окружающий точку *y* (рис. 2.21)

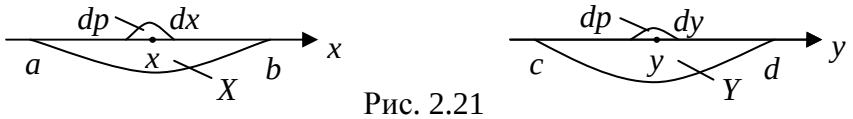


Рис. 2.21

В выражении $dy = d\varphi(x) = |\varphi'(x)|dx$ производная $\varphi'(x)$ взята по модулю с тем, чтобы выражение для длины *dy* было верным и для возрастающей функции $y = \varphi(x)$ (когда $\varphi'(x) \succ 0$), и для убывающей (когда $\varphi'(x) \prec 0$).

Попадание значения случайной величины *X* на отрезок *dx* будет, очевидно, автоматически означать попадание значения случайной величины *Y* на отрезок *dy*. Вероятности *dp* появления обоих этих событий, таким образом, одинаковы. Следовательно, одновременно имеем:

$$dp = f(x)dx; \quad dp = g(y)dy = g(y)|\varphi'(x)|dx \tag{5.5}$$

Сравнивая эти два значения *dp*, получаем:

$$g(y) = \frac{f(x)}{|\varphi'(x)|} \tag{5.6}$$

Чтобы получить окончательное выражение для *g (y)*, нужно правую часть равенства (5.6) выразить через *y*. Для этого из равенства $y = \varphi(x)$, связывающего *x* и *y*, нужно *x* выразить через *y*. То есть нужно найти функцию $x = \psi(y)$, обрат-

ную к функции $y = \varphi(x)$. Такая обратная функция заведомо существует, если исходная функция $y = \varphi(x)$ монотонно возрастает или монотонно убывает, что мы и будем предполагать. Найдя функцию $x = \psi(y)$, можем и производную $\varphi'(x)$ функции $y = \varphi(x)$ выразить через y:

$$\varphi'(x) = \frac{dy}{dx} = \frac{1}{\frac{dx}{dy}} = \frac{1}{\psi'(y)} \qquad (5.7)$$

С учетом этого функция $g(y)$ причем следующий окончательный вид:

$$g(y) = f(\psi(y))|\psi'(y)| \quad (c<y<d) \qquad (5.8)$$

Это и есть плотность вероятности непрерывной случайной величины Y, являющейся функцией $Y=\varphi(X)$ непрерывной случайной величины X. При этом $f(x)$ - плотность вероятности величины X, а $x = \psi(y)$ - функция, обратная функции $y = \varphi(x)$.

После того, как найдена плотность вероятности $g(y)$ случайной величины $Y = \varphi(X)$, можно, при желании, найти все числовые характеристики этой величины:

$$M(Y) = \int_c^d yg(y)dy; \quad D(Y) = M(Y^2) - [M(Y)]^2, \text{ где } M(Y^2) = \int_c^d y^2 g(y)dy; \qquad (5.9)$$

$$\sigma(Y) = \sqrt{D(Y)}; \quad V(Y)\% = \frac{\sigma(Y)}{M(Y)} \cdot 100\%$$

Впрочем, для нахождения этих числовых характеристик величины $Y = \varphi(X)$ не обязательно находить функцию $g(y)$. Подставляя в интегралы (5.9) выражение (5.8) для $g(y)$ и делая затем подстановку $x = \psi(y)$, получим (выкладки проделайте самостоятельно):

$$M(Y) = M(\varphi(X)) = \int_a^b \varphi(x) f(x)dx \;;$$

$$M(Y^2) = M(\varphi(X))^2 = \int_a^b \varphi^2(x) f(x)dx; \qquad (5.10)$$

$$D(Y) = M(Y^2) - [M(Y)]^2; \quad \sigma(Y) = \sqrt{D(Y)}; \quad V(Y)\% = \frac{\sigma(Y)}{M(Y)} \cdot 100\%$$

Пример 2. Непрерывная случайная величина X задана плотностью вероятности $f(x) = 3x^2\,(0 \le x \le 1)$. Найти плотность вероятности случайной величины $Y = \sqrt{X}$, а также вычислить её основные числовые характеристики.

Решение. Сначала найдём промежуток $[c;d]$ возможных значений y величины $Y=\sqrt{X}$. То есть найдем область значений функции $y = \sqrt{x}$ $(0 \le x \le 1)$. Она очевидна: $0 \le y \le 1$. То есть $[c;d] = [0;1]$.

Теперь найдём $g(y)$ $(0 \le y \le 1)$ – плотность вероятности случайной величины Y. Так как $y = \varphi(x) = \sqrt{x}$ $(0 \le x \le 1)$, то $x = \psi(y) = y^2$ $(0 \le y \le 1)$, а $\psi'(y) = 2y$. И тогда, согласно (5.8), получаем :

$$g(y) = 3\left(y^2\right)^2 \cdot |2y| = 6y^5 \;\; (0 \le y \le 1) \qquad (5.11)$$

А отсюда уже, согласно (5.9), находим:

$$M(Y)=\int_0^1 y\cdot 6y^5dy=\frac{6}{7};\quad M(Y^2)=\int_0^1 y^2\cdot 6y^5dy=\frac{3}{4};$$

$$D(Y)=\frac{3}{4}-\left(\frac{6}{7}\right)^2=\frac{3}{196};\ \ \sigma(Y)=\sqrt{\frac{3}{196}}\approx 0.124;\ \ V(Y)\%\approx 14.4\%.$$

Впрочем, для нахождения числовых характеристик случайной величины $Y=\sqrt{X}$ можно было бы применить и формулы (5.10):

$$M(Y)=M(\sqrt{X})=\int_0^1\sqrt{x}\cdot 3x^2dx=\frac{6}{7};$$

$$M(Y^2)=M\left(\sqrt{X}\right)^2=M(X)=\int_0^1 x\cdot 3x^2dx=\frac{3}{4};$$

$$D(Y)=\frac{3}{196};\ \ \sigma(Y)\approx 0.124;\ \ V(Y)\%=14{,}4\%$$

Пример 3. Непрерывная случайная величина X распределена нормально с параметрами $(a;\ \sigma)$. Доказать, что любая линейная функция $Y=C_1X+C_2$ $(C_1\neq 0)$ этой величины X тоже распределена нормально с параметрами $(a_*=C_1a+C_2;$ $\sigma_*=|C_1|\sigma)$.

Доказательство. Плотность вероятности $f(x)$ нормально распределенной величины X имеет, как известно, вид:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}\quad(-\infty<x<\infty)$$

Так как возможные значения x величины X - любые числа от - ∞ до ∞, то и возможные значения y величины $Y=\varphi(X)=C_1X+C_2$ - любые числа. То есть интервал $(c;\ d)$ возможных значений величины Y – это вся ось oy от - ∞ до ∞.

Теперь найдем $g(y)$ $(-\infty<x<\infty)$ - плотность вероятности случайной величины Y. Так как $y=\varphi(x)=C_1x+C_2$, то $x=\psi(y)=\dfrac{y-C_2}{C_1}$, а $\psi'(y)=\dfrac{1}{C_1}$. И тогда, согласно (5.8), получаем

$$g(y)=f\left(\frac{y-C_2}{C_1}\right)\cdot\frac{1}{|C_1|}=\frac{1}{|C_1|}\cdot\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{\left(\frac{y-C_2}{C_1}-a\right)^2}{2\sigma^2}}=\frac{1}{\sigma_*\sqrt{2\pi}}e^{-\frac{(y-a_*)^2}{2\sigma_*^2}}\ (-\infty<y<\infty),$$

где

$$a_*=C_1a+C_2;\ \ \sigma_*=|C_1|\sigma.$$

Доказательство закончено.

Для нормально распределенных случайных величин доказан и более общий факт: если $(X_1;\ X_2;\ldots;X_p)$ – независимые нормально распределенные случайные величины, то любая их линейная комбинация

$$Z=C_1X_1+C_2X_2+\ldots+C_pX_p \qquad (5.12)$$

тоже является случайной величиной, распределенной нормально. И если $(a_k;\sigma_k)$ – параметры величин X_k (k = 1, 2,..., p), то параметры $(a_*;\sigma_*)$ величины Z таковы:

$$a_*=C_1a_1+C_2a_2+.....+C_pa_p;\quad \sigma_*=\sqrt{C_1^2\sigma_1^2+C_2^2\sigma_2^2+.....+C_p^2\sigma_p^2} \qquad (5.13)$$

Упражнения

1. Дискретная случайная величина X задана законом распределения

X	-1	-2	1	2
p	0.3	0.1	0.2	0.4

Найти закон распределения величины $Y = |X|$ и ее числовые характеристики.

Ответ:

Y	1	2
p	0.5	0.5

$M(Y) = 1{,}5;\ D(Y) = 0{,}25;\ \sigma(Y) = 0{,}5;\ V(Y) \approx 33\%$

2. Случайная величина X распределена равномерно. Доказать, что любая линейная функция $Y = C_1X + C_2$ $(C_1 \neq 0)$ величины X тоже распределена равномерно.

3. Случайная величина X распределена по показательному закону. Доказать, что линейная функция $Y = C_1X + C_2$ величины X тоже будет распределена по показательному закону, но лишь при C_1>0 и C_2=0.

4. Случайная величина X распределена нормально с параметрами $(a;\sigma)$. Доказать, что случайная величина $Y = \dfrac{X - a}{\sigma}$ тоже будет распределена нормально с параметрами $a_* = 0; \sigma_* = 1,$ то есть будет нормированный нормально распределенной случайной величиной.

§6. Корреляционная зависимость случайных величин. Корреляционный момент (ковариация) и коэффициент линейной корреляции. Корреляционное отношение.

Пусть X и Y – любые две случайные величины (дискретные или непрерывные – неважно). Нас будет интересовать связь между ними. Относительно этой связи имеется, в принципе, три возможности.

1) Первая возможность: величины X и Y независимы друг от друга. Это значит, что каждая из этих величин принимает свои значения независимо от значений, принимаемых другой случайной величиной.

2) Вторая возможность - обратная первой: величины X и Y связаны жесткой (функциональной) зависимостью, т.е. зависимостью вида $Y=\varphi(X)$. В этом случае каждому возможному значению x величины X соответствует вполне определенное значение $y=\varphi(x)$ величины Y. То есть возможные значения величины Y жестко привязаны к возможным значениям величины X. Этому случаю был посвящен предыдущий параграф.

3) Третья возможность - промежуточная между первыми двумя: X и Y в принципе связаны между собой (независимыми они не являются), но эта связь не жёсткая (размытая). Это значит, что каждому возможному значению x величины X могут соответствовать различные значения (y_1; y_2; ...) величины Y, причём набор этих значений и (или) их вероятности меняются с изменением значения x. Такого рода связь между случайными величинами называются *статистической* (или *вероятностной*) связью. Статистическая связь между случайными величинами X и Y означает, что изменение значения одной из них ведет к изменению *внешних условий* для реализации другой величины. Например, меняющаяся среднесуточная температура статистически влияет на плотность сельскохозяйственных вредителей на засеянном поле; объем денежной массы у покупателей статистически влияет на объем закупаемых ими товаров, и т.д.

Если при статистической связи между случайными величинами X и Y при изменении значения x величины X еще и меняется *среднее значение* $\bar{y}_x$ величины Y, то говорят, что Y *корреляционно* (в среднем) зависит от X. Аналогично понимается корреляционная зависимость X от Y. В частности, очевидно, что между температурой X воздуха и количеством Y вредителей имеет место не просто статистическая, а корреляционная зависимость, ибо с изменением температуры изменяется и среднее количество сельскохозяйственных вредителей. Аналогично между количеством X денег у покупателей и их тратами Y на покупку товаров тоже имеется, очевидно, корреляционная зависимость, ибо чем больше денег у покупателей, тем больше в среднем они покупают. Корреляционно (в среднем) связаны также урожайность различных культур с количеством внесенных под них удобрений, производительность труда рабочих с их квалификацией, и т.д.

Рассмотрим корреляционную связь между случайными величинами X и Y подробнее. Пусть $\bar{y}_x$ - среднее значение тех значений y величины Y, которые соответствуют данному значению x величины X. Оно же - условное математическое ожидание величины Y при $X=x$:

$$\bar{y}_x = M(Y / X = x) \qquad (6.1)$$

Так как каждому возможному значению x величины X будет соответствовать единственное значение $\bar{y}_x$, то это значение $\bar{y}_x$ является функцией от x:

$$\bar{y}_x = f(x) \qquad (6.2)$$

Если $\bar{y}_x$ меняется с изменением x, то есть если $f(x) \neq Const$, то между X и Y имеется корреляционная связь – Y корреляционно (в среднем) зависит от X. А если $\bar{y}_x = Const$, то Y корреляционно от X не зависит. В последнем случае Y либо вообще не зависит от X, либо зависит, но лишь сугубо статистически.

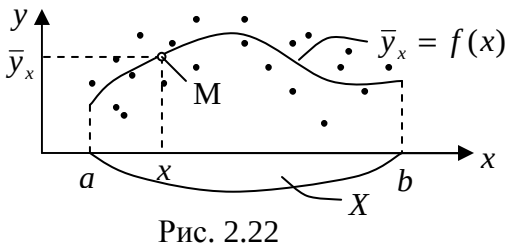


Рис. 2.22

Функциональная зависимость (6.2) называется *уравнением регрессии Y на X*,

а график этой зависимости – *линией регрессии Y на X* (рис 2.22):

Линия регрессии Y на X наглядно показывает, как *в среднем* меняется случайная величинаY при изменении случайной величины X. Точки вокруг линии регрессии символизируют разброс возможных значений y величины Y вокруг линии регрессии $\bar{y}_x = f(x)$. Именно из этих значений y для каждого x должно быть найдено их среднее значение $\bar{y}_x$.

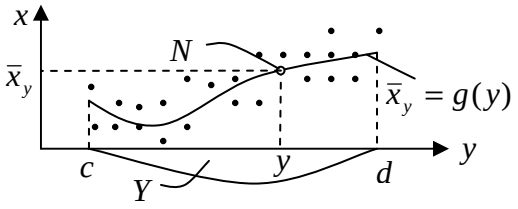


Рис. 2.23

Аналогично зависимость вида $\bar{x}_y = g(y)$ называется *уравнением регрессии X на Y*, а ее график – *линией регрессии X на Y* (рис 2.23).

Линия регрессии X на Y показывает, как *в среднем* меняется X при изменении Y.

Самой простой случай (и наиболее часто встречающийся на практике) – это когда функция $\bar{y}_x = f(x)$ или $\bar{x}_y = g(y)$ линейна, то есть когда её график – прямая линия. В этом случае корреляционная зависимость Y от X и соответственно корреляционная зависимость X от Y называется *линейной*, в противном случае – *нелинейной*.

В теории корреляции решаются *две основные задачи*:

Первая задача теории корреляции - нахождение *уравнения регрессии*, то есть нахождение зависимости между значениями одной случайной величины и соответствующими им средними значениями другой случайной величины.

Вторая задача теории корреляции – оценка *тесноты изучаемой корреляционной зависимости*. В частности, теснота корреляционной зависимости Y от X оценивается по степени рассеяния значений (y_1; y_2; ....) величины Y (рис. 2.22) вокруг линии регрессии $\bar{y}_x = f(x)$. Большое рассеяние свидетельствует о слабой корреляционной зависимости Y от X. Наоборот, малое рассеяние указывает на наличие достаточно сильной (тесной) корреляционной зависимости. Возможно даже, что Y зависит от X функционально, то есть жёстко, но из-за второстепенных случайных факторов или просто из-за погрешностей измерений эта зависимость оказалась несколько размытой.

Те же задачи, естественно, стоят, если исследуется корреляционная зависимость X от Y.

Наиболее просто решаются обе эти задачи при наличии линейной корреляционной зависимости одной случайной величины от другой. И здесь важную роль играет так называемый *корреляционный момент* $\mu(X;Y)$ или, что одно и то же, *ковариация* $Cov(X,Y)$ случайных величин X и Y, которые определяются как математическое ожидание произведения отклонений X и Y от их математических ожиданий:

$$\mu(X,Y) = Cov(X,Y) = M[(X - M(X)) \cdot (Y - M(Y))] \qquad (6.4)$$

Их можно преобразовать к виду (проделайте это самостоятельно):

$$\mu(X,Y) = Cov(X,Y) = M(XY) - M(X)M(Y) \qquad (6.5)$$

Как известно, у независимых случайных величин X и Y, как у дискретных, так и у непрерывных, $M(XY) = M(X)M(Y)$. А значит, для независимых случайных величин

$$\mu(X,Y) = Cov(X,Y) = 0 \qquad (6.6)$$

Поэтому если $\mu(X,Y) \neq 0$, то это автоматически указывает на зависимость случайных величин X и Y друг от друга.

Отметим, что обратное, вообще говоря, неверно: из того, что корреляционный момент $\mu(X,Y) = 0$, ещё нельзя сделать вывод, что X и Y независимы. Они могут быть зависимы, причём даже функционально. Например, если распределение величины X симметрично относительно точки x=0,так что автоматически $M(X) = 0$ и $M(X^3) = 0$, а $Y = X^2$ - функция от X, то на основании (6.5) получаем:

$$\mu(X,Y) = M(X^3) - M(X)M(X^2) = 0 - 0 = 0$$

И это несмотря на то, что X и Y связаны функциональной зависимостью $Y = X^2$.

Случайные величины, для которых $\mu(X,Y) = 0$, называются *линейно некоррелированными*. Независимые величины всегда линейно некоррелированы. Но линейно некоррелированные величины могут быть, как мы только что видели, как зависимыми, так и независимыми. Линейно коррелированные же величины (для них $\mu(X,Y) \neq 0$) всегда зависимы.

Кстати, если случайные величины X и Y распределены нормально, то можно доказать (на этом не останавливаемся), что *их линейная некоррелированность равнозначна их независимости*. Для других же величин X и Y это не обязательно одно и тоже.

Отметим, что корреляционный момент $\mu(X;Y)$ обладает одним существенным недостатком: он зависит от единиц измерения величин X и Y. Поэтому на практике вместо него часто используется безразмерная величина

$$\rho(X,Y) = \frac{\mu(X,Y)}{\sigma(X)\sigma(Y)}, \qquad (6.7)$$

которая называется *коэффициентом линейной корреляции*. Он играет, как мы увидим ниже, большую роль при решении обеих задач теории корреляции в случае линейной корреляционной зависимости между случайными величинами.

Корреляционный момент $\mu(X;Y)$ и коэффициент линейной корреляции $\rho(X,Y)$ равны или не равны нулю одновременно. Поэтому линейную коррелированность и линейную некоррелированность случайных величин X и Y можно устанавливать и по равенству или неравенству нулю коэффициента линейной корреляции $\rho(X,Y)$.

Так как, согласно (6.5), $\mu(Y,X) = \mu(X;Y)$, то и

$$\rho(Y,X) = \rho(X,Y) \qquad (6.8)$$

Коэффициент линейной корреляции обладает еще одним важным свойством: он не изменится, если от X и Y перейти к безразмерным нормированным случайным величинам

$$X_0 = \frac{X - M(X)}{\sigma(X)}; \quad Y_0 = \frac{Y - M(Y)}{\sigma(Y)} \qquad (6.9)$$

То есть

$$\rho(X,Y) = \rho(X_0,Y_0) \qquad (6.10)$$

Нормированными случайными величинами X_0 и Y_0 называются потому, что их математические ожидания равны нулю, а средние квадратические отклонения равны единице:

$$M(X_0)=0;\ \ M(Y_0)=0;\ \ \sigma(X_0)=\sqrt{D(X_0)}=1;\ \ \sigma(Y_0)=\sqrt{D(Y_0)}=1; \qquad (6.11)$$

Равенства (6.11) легко доказываются с помощью свойств (3.17) – (3.23) математического ожидания и дисперсии, которые справедливы как для непрерывных, так и для дискретных случайных величин (проделайте это самостоятельно). Ну, а то, что $\rho(X_0,Y_0)=\rho(X,Y)$, уже вытекает из (6.4), (6.5), (6.7), (6.9) и (6.11):

$$\rho(X_0,Y_0)=\frac{\mu(X_0,Y_0)}{1\cdot 1}=\mu\left[\frac{X-M(X)}{\sigma(X)},\frac{Y-M(Y)}{\sigma(Y)}\right]=M\left[\frac{X-M(X)}{\sigma(X)}\cdot\frac{Y-M(Y)}{\sigma(Y)}\right]-$$

$$-M\left[\frac{X-M(X)}{\sigma(X)}\right]\cdot M\left[\frac{Y-M(Y)}{\sigma(Y)}\right]=\frac{\mu(X,Y)}{\sigma(X)\sigma(Y)}-0\cdot 0=\rho(X,Y)$$

Для дальнейшего рассмотрения свойств коэффициента линейной корреляции $\rho(X,Y)$ случайных величин X и Y найдем дисперсию их суммы $X+Y$ и разности $X-Y$. Если величина X и Y независимы, то такая формула уже получена (см. (3.22)):

$$D(X\pm Y)=D(X)+D(Y) \qquad (6.12)$$

Причем эта формула верна как для дискретных, так и для непрерывных случайных величин. А если X и Y зависимы (функционально или статистически), то соответствующая формула имеет вид:

$$D(X\pm Y)=D(X)+D(Y)\pm 2\mu(X,Y) \qquad (6.13)$$

Действительно:

$$D(X\pm Y)=M(X\pm Y)^2-[M(X\pm Y)]^2=M(X^2\pm 2XY+Y^2)-[M(X)\pm M(Y)]^2=$$

$$=M(X^2)-[M(X)]^2+M(Y^2)-[M(Y)]^2\pm 2[M(XY)-M(X)M(Y)]=D(X)+D(Y)\pm 2\mu(X,Y)$$

В частности, для нормированных случайных величин (X_0,Y_0) формула (6.13) примет вид:

$$D(X_0\pm Y_0)=1+1\pm 2\rho(X_0,Y_0)=2[1\pm\rho(X_0,Y_0)] \qquad (6.14)$$

А так как, по смыслу дисперсии, $D(X_0\pm Y_0)\ge 0$, то из (6.14) получаем:

$$1\pm\rho(X_0,Y_0)\ge 0\Leftrightarrow -1\le\rho(X_0,Y_0)\le 1 \qquad (6.15)$$

И так как, согласно (6.10), $\rho(X_0,Y_0)=\rho(X,Y)$, то для любых случайных величин X и Y получаем следующий вывод:

$$-1\le\rho(X,Y)\le 1 \qquad (6,16)$$

Если коэффициент линейной корреляции $\rho(X,Y)\ne 0$, то он характеризует не только *наличие* зависимости (связи) между X и Y. Своей величиной, как мы это сейчас увидим, он характеризует *и тесноту* этой связи. Однако не любой, а *лишь линейной* корреляционной связи между X и Y. Отсюда и его название – коэффициент *линейной* корреляции. Максимальная теснота этой связи соответствует случаям, когда $\rho(X;Y)=\pm 1$. При этом между X и Y имеет место жёсткая функциональная связь, причём связь непременно линейная: $Y=aX+b$.

Действительно, при $\rho(X,Y)= \pm 1$ и $\rho(X_0,Y_0)=\pm 1$, а тогда из (6.14) вытекает, что имеет место одно из двух равенств: или $D(X_0+Y_0)=0$, или $D(X_0-Y_0)=0$. Но дисперсия случайной величины равна нулю, если только эта случайная величина является константой. То есть или $X_0+Y_0=C$, или $X_0-Y_0=C$. Заметим, что в обоих случаях константа $C=0$, ибо на основании (6.11) получаем:

$$M(X_0+Y_0)=M(C) \qquad M(X_0-Y_0)=M(C)$$
$$M(X_0)+M(Y_0)=C \qquad M(X_0)-M(Y_0)=C$$
$$0=C \qquad 0=C$$

Итак, при $\rho(X,Y)= \pm 1$ либо $X_0+Y_0=0$, либо $X_0-Y_0=0$. А отсюда уже, согласно связи (6.9) (X_0,Y_0) с (X,Y), следует подтверждение того, что в обоих случаях величины X и Y связаны линейной функциональной зависимостью вида $Y=aX+b$.

Верно и обратное: если случайные величины X и Y связаны линейной функциональной зависимостью $Y=aX+b$, то их коэффициент линейной корреляции $\rho(X,Y)$ равен либо 1, либо -1.

Докажем это. Действительно, если $Y=aX+b$, то согласно (6.9) и свойств математического ожидания и дисперсии получаем:

$$X_0=\frac{X-M(X)}{\sigma(X)};\quad Y_0=\frac{Y-M(Y)}{\sigma(Y)}=\frac{aX+b-M(aX+b)}{\sqrt{D(aX+b)}}=\frac{aX+b-aM(X)-b}{\sqrt{a^2D(X)+0}}=$$
$$=\frac{a(X-M(X))}{|a|\sigma(X)}=\frac{a}{|a|}X_0.$$

А тогда

$$\rho(X,Y)=\rho(X_0,Y_0)=\frac{\mu(X_0,Y_0)}{\sigma(X_0)\sigma(Y_0)}=\mu(X_0,\frac{a}{|a|}X_0)=$$
$$=\frac{a}{|a|}\mu(X_0,X_0)=\frac{a}{|a|}D(X_0)=\frac{a}{|a|}=\begin{cases}1, \text{ если } a\succ 0\\ -1, \text{ если } a\prec 0\end{cases}$$

Таким образом, коэффициент линейной корреляции $\rho(X,Y)$ есть показатель того, насколько зависимость между случайными величинами X и Y близка к строгой линейной зависимости $Y=aX+b$. Его малость (удаленность от ± 1) может означать одно из двух: или малую тесноту (большое рассеяние) линейной корреляционной связи между X и Y, или существенную нелинейность этой связи, которая, кстати, может быть весьма тесной.

Сформулируем это утверждение более определенно. Найдем такие числовые коэффициенты k и b, чтобы линейная функция $\kappa X+b$ случайной величины X наилучшим образом приближала случайную величину Y. Для этого представим Y в виде

$$Y=\kappa X+b+Z \qquad (6.17)$$

Случайную величину Z можно рассматривать как ошибку приближения величины Y линейной функцией $Y=\kappa X+b$. Эту ошибку естественно считать минимальной, если потребовать, чтобы математическое ожидание $M(Z)=0$ и дисперсия $D(Z)$ была минимальной. Первое из этих требований дает:

$$M(Z)=M(Y)-kM(X)-b=0, \text{ откуда } b=M(Y)-kM(X) \qquad (6.18)$$

С учетом найденного значения b и (6.17) ошибка Z примет вид:

$$Z = Y - M(Y) - k(X - M(X))$$

Теперь вычислим $D(Z)$ – дисперсию величины Z:

$$\begin{aligned} D(Z) &= M(Z^2) - [M(Z)]^2 = M(Z^2) = M[(Y - M(Y)) - k(X - M(X))]^2 = \\ &= M(Y - M(Y))^2 - 2kM[(Y - M(Y)(X - M(X)] + k^2 M(X - M(X))^2 = \\ &= D(Y) - 2k\mu(X,Y) + k^2 D(X) = \sigma^2(Y) - 2k\sigma(X)\sigma(Y)\rho(X,Y) + k^2\sigma^2(X) = \\ &= [\sigma^2(Y) - \rho^2(X,Y)\sigma^2(Y)] + [\rho^2(X,Y)\sigma^2(Y) - 2k\sigma(X)\sigma(Y)\rho(X,Y) + k^2\sigma^2(X)] = \\ &= \sigma^2(Y)(1 - \rho^2(X,Y)) + (\rho(X,Y)\sigma(Y) - k\sigma(X))^2 \end{aligned}$$

Первое из полученных слагаемых неотрицательно и не зависит от параметра k. Таким образом, дисперсия $D(Z)$ ошибки Z будет минимальной при том значении k, которое обеспечит обращение в нуль второго слагаемого. То есть при

$$k = \frac{\sigma(Y)}{\sigma(X)}\rho(X,Y) \qquad (6.19)$$

При этом дисперсия $D(Z)$ (её минимальное значение) примет вид:

$$D(Z) = \sigma^2(Y)(1 - \rho^2(X,Y)) \qquad (6.20)$$

Итак, вывод: наилучшее приближение случайной величины Y линейной функцией $\kappa X+b$ случайной величины X будет иметь место при значениях k и b, определяемых формулами (6.19) и (6.18). То есть такое приближение будет иметь вид:

$$Y \approx \frac{\sigma(Y)}{\sigma(X)}\rho(X,Y)(X - M(X)) + M(Y) \qquad (6.21)$$

Ошибка Z этого линейного приближения величины Y имеет математическое ожидание (среднее значение), равное нулю. А дисперсия этой ошибки определяется формулой (6.20).

Если $\rho(X,Y) = \pm 1$, то дисперсия ошибки $D(Z) = 0$. А это, с учетом равенства $M(Z) = 0$ означает, что $Z = 0$. То есть при $\rho(X,Y) = \pm 1$ в равенстве (6.21) ошибки нет и оно является точным. Но чем больше удален коэффициент линейной корреляции $\rho(X,Y)$ от ± 1, то есть чем ближе он к нулю, тем больше становится дисперсия $D(Z)$ ошибки Z, а вместе с ней тем больше становится и сама ошибка Z приближения (6.21). При $\rho(X,Y) = 0$ эта ошибка становится максимально возможной, а само приближение (6.21) принимает вид $Y \approx M(Y)$ и перестаёт, таким образом, зависеть от X. То есть при $\rho(X,Y)$=0 линейная зависимость Y от X отсутствует. Это значит, что или между случайными величинами X и Y вообще нет никакой связи, или они связаны, но какой-то нелинейной связью (функциональной или статистической).

Кстати, так как наилучшим приближением случайной величины Y при $X=x$ является, очевидно, условная средняя $\bar{y}_x$, то из (6.21) сразу вытекает *наилучшее линейное приближение* уравнения регрессии $\bar{y}_x = f(x)$ величины Y на величину X. Для его получения нужно в (6.21) заменить X на x и Y на $\bar{y}_x$. В итоге получим:

$$\bar{y}_x - \bar{y} \approx k(x - \bar{x}) \qquad (6.22)$$

Здесь

$$\bar{y} = M(Y);\ \bar{x} = M(X);\ k = \frac{\sigma(Y)}{\sigma(X)}\rho(X,Y) \qquad (6.23)$$

Полученное простое линейное уравнение (6.22) используют на практике для приближенной замены истинного уравнения регрессии $\bar{y}_x = f(x)$, если линия регрессии близка к прямой. Если же она сильно отличается от прямой (как на рис. 2.22), то его тоже можно использовать, только не на всем интервале (*a; b*) возможных значений величины *X*, а на коротких частях этого интервала, на которых линию регрессии можно приближенно считать прямой.

При $\rho(X,Y) = \pm 1$ приближенное линейное уравнение (6.22) становится точным. То есть становится истинным уравнением $\bar{y}_x = f(x)$ регрессии *Y* на *X*. Более того, при этом $\bar{y}_x$ превращается просто в *y* – в единственное значение *Y* при *X*=*x*. Это происходит потому, что при $\rho(X,Y) = \pm 1$ становится точным равенство (6.21). А это значит, что каждому значению *x* величины *X* будет соответствовать единственное значение *y* величины *Y*. И, таким образом, будет $\bar{y}_x = y$. Линия регрессии $\bar{y}_x = f(x)$ (см. рис. 2.22) станет прямой, и никакого разброса вокруг неё точек, изображающих возможные значения величины *Y*, не будет – все они окажутся на этой прямой.

Но если $\rho(X,Y) \neq \pm 1$, то по мере удаления его значения от ± 1 истинная линия регрессии или искривляется, или остается прямой, но вокруг нее появляется облако точек, причем тем более широкое, чем ближе $\rho(X,Y)$ к нулю. Или одновременно и линия регрессии искривляется, и облако точек вокруг нее расширяется. При $\rho(X,Y)$ близком к нулю или тем более равном нулю нельзя даже приближено считать величины *X* и *Y* связанными линейной корреляционной зависимостью. Связь между этими линейно некоррелированными (или слабо линейно коррелированными) случайными величинами будет или отсутствовать вообще, или будет существенно нелинейной. То есть в этом случае полученные выше формулы (6.21) и (6.22) приближенного линейного выражения одной величины (*Y*) через другую величину (*X*) применять нельзя - они могут давать слишком грубое приближение. Тут требуется дополнительное исследование характера связи между такого рода слабо линейно коррелированными случайными величинами *X* и *Y*, которое мы проведем ниже.

Перейдем к этому исследованию. То есть поставим вопрос об оценке тесноты *любой*, а не только линейной, корреляционной связи между случайными величинами *X* и *Y*.

Итак, допустим, что корреляционная связь между случайными величинами *X* и *Y* есть, и эта связь заведомо нелинейная (квадратичная, экспоненциальная, логарифмическая, и т. д.). Это значит, что уравнение $\bar{y}_x = f(x)$ регрессии *Y* на *X* таково, что $\bar{y}_x \neq Const$ и при этом $\bar{y}_x \neq kX + b$. То есть линия регрессии *Y* на *X* – кривая линия (рис. 2.22). Для оценки тесноты такой криволинейной корреляционной связи между *X* и *Y* коэффициент линейной корреляции $\rho(X,Y)$, который будет близок к нулю, не годится. В этом случае указанную тесноту оценивают с помощью так называемого *корреляционного отношения.*

Чтобы ввести это понятие, рассмотрим случайною величину $\overline{Y} = f(X)$, которая является функцией величины X и которая при $X = x$ принимает среднее значение $\bar{y}_x = f(x)$ величины Y. Математическое ожидание $M(\overline{Y}) = M(f(X))$ величины $\overline{Y}$ совпадает с математическим ожиданием (средним значением $\bar{y}$) величины Y:

$$M(\overline{Y}) = M(Y) = \bar{y} \qquad (6.24)$$

А дисперсия $D(\overline{Y}) = D(f(X))$ величины $\overline{Y}$ составляет лишь часть дисперсии $D(Y)$ величины Y:

$$D(Y) = D(\overline{Y}) + M(Y - \overline{Y})^2 \qquad (6.25)$$

При доказательстве равенств (6.24) и (6.25) ограничимся случаем, когда X и Y – дискретные случайные величины.

Итак, пусть X и Y – зависимые дискретные случайные величины, а таблица (6.26) – закон их совместного распределения:

Y \ X	x_1	x_2	.....	x_n	q_j
y_1	p_{11}	p_{21}	.....	p_{n1}	q_1
y_2	p_{12}	p_{22}	.....	p_{n2}	q_2
.....	.....	.....	.....	.....	.....
y_m	p_{1m}	p_{2m}	.....	p_{nm}	q_m
p_i	p_1	p_2	.....	p_n	1

(6.26)

Здесь $(x_1, x_2, \ldots x_n)$ и $(y_1, y_2, \ldots y_m)$ – возможные значения величин X и Y соответственно, а p_{ij} - вероятности того, что в результате испытания парой случайных величин (X, Y) будет принята пара значений $(x_i; y_j)$ $(i = 1, 2, \ldots n;\ j = 1, 2, \ldots m)$. Кстати, сумма всех вероятностей p_{ij}, как сумма вероятностей событий, составляющих полную группу событий, должна равняться единице:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} = 1 \qquad (6.27)$$

Действительно, события, состоящие в том, что $(X = x_i;\ Y = y_j)$, являются несовместными. Причем одно из них обязательно произойдет. То есть эти события действительно образуют полную группу событий.

В последней строке таблицы (6.26) просуммированы вероятности p_{ij} по строкам (внутри каждого столбца). А в последнем столбце этой таблицы просуммированы вероятности p_{ij} по столбцам (внутри каждой строки):

$$p_i = \sum_{j=1}^{m} p_{ij}\ \ (i = 1, 2, \ldots n); \quad q_j = \sum_{i=1}^{n} p_{ij}\ \ (j = 1, 2, \ldots m); \quad \sum_{i=1}^{n} p_i = \sum_{j=1}^{m} q_j = 1 \qquad (6.28)$$

Вероятности p_i - это, очевидно, вероятности значений x_i величины X, а вероятности q_i - это вероятности значений y_j величины Y. То есть на базе закона сов-

местного распределения случайных величин X и Y можно записать и законы распределения каждой из этих величин в отдельности:

X	x_1	x_2	…	x_n
p	p_1	p_2	…	p_n

Y	y_1	y_2	…	y_m
p	q_1	q_2	…	q_m

(6.29)

Среднее значение $\bar{y}_{x_i}$ величины Y для каждого возможного значения x_i величины Y следует находить по формуле:

$$\bar{y}_{x_i} = f(x_i) = \frac{\sum_{j=1}^{m} y_i p_{ij}}{p_i} \quad (i = 1, 2, \dots n) \tag{6.30}$$

Действительно, согласно (6.1)

$$\bar{y}_{x_i} = M(Y / X = x_i) \quad (i = 1, 2, \dots n) \tag{6.31}$$

То есть $\bar{y}_{x_i}$ - это условное математическое ожидание величины Y при $X = x_i$. А следовательно, оно должно быть найдено как сумма произведений значений y_j величины Y на соответствующее им вероятности этих значений при условии, что $X = x_i$. То есть

$$\bar{y}_{x_i} = \sum_{j=1}^{m} y_j \cdot p(Y = y_j / X = x_i) \tag{6.32}$$

А условные вероятности $p(Y = y_j / X = x_i)$ можно найти из формулы вероятности произведения двух зависимых событий (формула (4.5) главы 1):

$$p(X = x_i; Y = y_j) = p(X = x_i) \cdot p(Y = y_j / X = x_i);$$

$$p_{ij} = p_i \cdot p(Y = y_j / X = x_i) \Rightarrow p(Y = y_j / X = x_i) = \frac{p_{ij}}{p_i} \tag{6.33}$$

Из формул (6.32) и (6.33) и следует формула (6.30).

Подсчитав значения $\bar{y}_{x_i}$, можем составить и закон распределения случайной величины $\bar{Y} = f(X)$:

$\bar{Y} = f(X)$	$\bar{y}_{x_1}$	$\bar{y}_{x_2}$	…	$\bar{y}_{x_n}$
p	p_1	p_2	…	p_n

(6.34)

(вероятности значений $\bar{y}_{x_i}$ величины $\bar{Y}$ те же, что и вероятности значений x_i величины X).

Ну, а теперь можем перейти к доказательству равенств (6.24) и (6.25). Сначала докажем (6.24):

$$M(\bar{Y}) = M(f(X)) = \sum_{i=1}^{n} \bar{y}_{x_i} p_i = \sum_{i=1}^{n} \frac{\sum_{j=1}^{m} y_j p_{ij}}{p_i} \cdot p_i =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} y_j p_{ij} = \sum_{j=1}^{m}\sum_{i=1}^{n} y_j p_{ij} = \sum_{j=1}^{m} y_j \sum_{i=1}^{n} p_{ij} = \sum_{j=1}^{m} y_j q_j = M(Y) = \bar{y} \tag{6.35}$$

Равенство (6.24) доказано.

Для доказательства равенства (6.25) образуем случайную величину $Y - \bar{Y}$ и запишем закон её распределения:

$Y-\bar{Y}$	$y_j-\bar{y}_{x_i}$
p	p_{ij}

$(i=1,2,\dots n;\quad \mathrm{j}=1,2,\dots \mathrm{m})$ (6.36)

Математическое ожидание этой случайной величины равно нулю - это следует из (6.24). Покажем ещё, что

$$M\left[(Y-\bar{Y})\cdot(\bar{Y}-\bar{y})\right]=0 \tag{6.37}$$

Закон распределения случайной величины $(Y-\bar{Y})\cdot(\bar{Y}-\bar{y})$ имеет вид:

$(Y-\bar{Y})(\bar{Y}-\bar{y})$	$(y_j-\bar{y}_{x_i})(\bar{y}_{x_i}-\bar{y})$
p	p_{ij}

$(i=1,2,\dots n;\quad \mathrm{j}=1,2,\dots \mathrm{m})$ (6.38)

Отсюда следует:

$$M\left[(Y-\bar{Y})(\bar{Y}-\bar{y})\right]=\sum_{i=1}^{n}\sum_{j=1}^{m}(y_j-\bar{y}_{x_i})(\bar{y}_{x_i}-\bar{y})p_{ij}=\sum_{i=1}^{n}(\bar{y}_{x_i}-\bar{y})\sum_{j=1}^{m}(y_j-\bar{y}_{x_i})p_{ij}=$$

$$=\sum_{i=1}^{n}(\bar{y}_{x_i}-\bar{y})\left(\sum_{j=1}^{m}y_j p_{ij}-\bar{y}_{x_i}\sum_{j=1}^{m}p_{ij}\right)=\sum_{i=1}^{n}(\bar{y}_{x_i}-\bar{y})\,(\bar{y}_{x_i}p_i-\bar{y}_{x_i}p_i)=0. \tag{6.39}$$

А теперь, опираясь на доказанные равенства (6.24) и (6.37), можно доказать и равенство (6.25):

$$D(Y)=D(Y-\bar{y})=M(Y-\bar{y})^2-\left[M(Y-\bar{y})\right]^2=M(Y-\bar{y})^2=$$

$$=M\left[(Y-\bar{Y})+(\bar{Y}-\bar{y})\right]^2=M\left[(Y-\bar{Y})^2+2(Y-\bar{Y})(\bar{Y}-\bar{y})+(\bar{Y}-\bar{y})^2\right]=$$

$$=M(Y-\bar{Y})^2+2M((Y-\bar{Y})(\bar{Y}-\bar{y}))+M(\bar{Y}-\bar{y})^2=M(Y-\bar{Y})^2+D(\bar{Y})$$

Равенство (6.25) доказано. Это равенство дает разложение общей дисперсии зависимой от X случайной величины Y на сумму двух слагаемых: дисперсии функции $\bar{Y}=f(X)$ и среднего квадрата отклонения Y от этой функции. Иначе говоря, общий разброс значений y величины Y вокруг её среднего значения $\bar{y}$ складывается из разброса значений $\bar{y}_x=f(x)$ величины $\bar{Y}=f(X)$ вокруг того же $\bar{y}$, и разброса значений y вокруг $\bar{y}_x=f(x)$. То есть формула (6.25) раскладывает общий разброс всех возможных значений y величины Y вокруг её математического ожидания $M(Y)=\bar{y}$ на разброс вокруг $\bar{y}$ точек $\bar{y}_{x_i}=f(x_i)$ кривой регрессии, и на разброс значений y (облака точек, изображающих значения y) вокруг кривой регрессии $\bar{y}_x=f(x)$.

Введем теперь отношение

$$\eta(Y,X)=\frac{\sigma(\bar{Y})}{\sigma(Y)}=\frac{\sigma(f(X))}{\sigma(Y)}=\frac{\sqrt{D(f(X))}}{\sqrt{D(Y)}}=\sqrt{1-\frac{M(Y-f(X))^2}{D(Y)}} \tag{6.40}$$

которое будет называть *корреляционным отношением Y к X*. Очевидно, что всегда

$$0\le\eta(Y,X)\le 1 \tag{6.41}$$

Из определения $\eta(Y,X)$ следует, что $\eta(Y,X)=0$ при $D(\bar{Y})=D(f(X))=0$, то есть при условии, что $\bar{Y}=f(X)$=Const. Причем эта константа, естественно, равна $\bar{y}$. Но тогда уравнение регрессии Y на X имеет вид $\bar{y}_x=f(x)=\bar{y}$ и, следовательно,

случайная величина Y не зависит корреляционно (в среднем) от величины X. А если $\eta(Y,X)=1$, то в этом случае из (6.40) следует, что $M(Y-f(X))^2=0$, откуда вытекает, что $Y=f(X)$. То есть при $\eta(Y,X)=1$ случайные величины X и Y связаны жесткой функциональной зависимостью $Y=f(X)$, причем $f(X)\neq \text{Const}$.

Из сказанного следует, что чем ближе корреляционное отношение $\eta(Y,X)$ к единице, тем ближе корреляционная зависимость Y от X к функциональной зависимости. А это значит, тем эта корреляционная зависимость теснее. Наоборот, чем ближе $\eta(Y,X)$ к нулю, тем она слабее.

Таким образом, корреляционное отношение $\eta(Y,X)$ случайной величины Y к случайной величине X является мерой и наличия, и тесноты *любой* (а не только линейной) корреляционной зависимости величины Y от величины X.

Естественно, можно ввести в рассмотрение и корреляционное отношение $\eta(X,Y)$ величины X к величине Y.

$$\eta(X,Y)=\frac{\sigma(\overline{X})}{\sigma(X)}=\frac{\sigma(g(Y))}{\sigma(X)}=\frac{\sqrt{D(g(Y))}}{\sqrt{D(X)}}=\sqrt{1-\frac{M(X-g(Y))^2}{D(X)}} \qquad (6.42)$$

которое оценивает наличие и тесноту корреляционной зависимости величины X от Y, где $\bar{x}_y=g(y)$ - уравнение регрессии X на Y.

Отметим, что в отличие от коэффициента линейной корреляции, которой симметричен относительно X и Y ($\rho(X,Y)=\rho(Y,X)$), корреляционное отношение таким свойством, судя по (6.40) и (6.42), не обладает:

$$\eta(X,Y)\neq\eta(Y,X) \qquad (6.43)$$

Можно еще доказать, что всегда

$$|\rho(X,Y)|\le\eta(X,Y) \ \text{и} \ |\rho(X,Y)|\le\eta(Y,X) \qquad (6.44)$$

При этом в случае равенства

$$\eta(Y,X)=|\rho(Y,X)| \qquad (6.45)$$

имеет место *точная линейная корреляционная зависимость Y от X*. Это значит, что при условии (6.45) приближенное уравнение регрессии (6.22) Y на X становится *точным*.

Аналогично в случае

$$\eta(X,Y)=|\rho(X,Y)| \qquad (6.46)$$

становится точным соответствующее уравнение регрессии $\bar{x}_y-\bar{x}=k_*(y-\bar{y})$ X на Y.

Пример. Дискретные случайные величины X и Y заданы следующим законом их совместного распределения:

Y \	0	1	2	q_j
0	0,10	0,16	0,18	0,44
1	0,06	0,20	0,30	0,56

p_i	0,16	0,36	0,48	1

Требуется:

1) Найти коэффициент линейной корреляции $\rho(Y,X)$.

2) Найти корреляционное отношение $\eta(Y,X)$.

3) Построить линию регрессии $\bar{y}_x = f(x)$ величины *Y* на величину *X*.

Решение. Запишем сначала законы распределения величин *X* и *Y* по отдельности:

X	0	1	2
p	0,16	0,36	0,48

Y	0	1
p	0,44	0,56

Отсюда, в частности, следует (получите это самостоятельно):

$$M(X) = \bar{x} = 1{,}32; \quad D(X) = 0{,}5376; \quad \sigma(X) \approx 0{,}733;$$

$$M(Y) = \bar{y} = 0{,}56; \quad D(Y) = 0{,}2464; \quad \sigma(Y) \approx 0.496;$$

Теперь найдем $\rho(Y,X)$. Для этого, согласно (6.7), предварительно нужно найти корреляционный момент $\mu(X,Y)$. Его найдем по формуле (6.5), используя совместный закон распределения (таблицу) величины *X* и *Y*:

$$\mu(X,Y) = M(XY) - M(X)M(Y) = \sum_{i=1}^{3}\sum_{j=1}^{2} x_i y_j p_{ij} - \bar{x}\cdot\bar{y} =$$

$$= 1\cdot 1\cdot 0{,}20 + 2\cdot 1\cdot 0{,}30 - 1{,}32\cdot 0{,}56 = 0.0608$$

Тогда:

$$\rho(X,Y) = \frac{\mu(X,Y)}{\sigma(X)\sigma(Y)} \approx 0.167$$

Величина $\rho(X,Y) \neq 0$. Таким образом, величины *X* и *Y* линейно коррелированы, а значит и зависимы. Вместе с тем величина $\rho(X,Y)$ невелика (она гораздо ближе к нулю, чем к 1 или к -1). Поэтому корреляционная зависимость *Y* от *X* или слабая, или существенно нелинейная, или то и другое вместе.

Чтобы лучше выяснить этот вопрос, подсчитаем корреляционное отношение $\eta(Y,X)$ величины *Y* к величине *X*. Для этого сначала для каждого значения *x* величины *X* подсчитаем среднее значение $\bar{y}_x$ величины *Y*. Используя формулы (6.30), получим:

$$\bar{y}_{x_1=0} = \frac{0\cdot 0{,}10 + 1\cdot 0{,}06}{0{,}16} = 0{,}375; \quad \bar{y}_{x_2=1} = \frac{0\cdot 0{,}16 + 1\cdot 0{,}20}{0{,}36} \approx 0{,}556;$$

$$\bar{y}_{x_3=2} = \frac{0\cdot 0{,}18 + 1\cdot 0{,}30}{0{,}48} = 0{,}625$$

Полученные данные позволяют записать таблицу вида (6.34) - закон распределения функции $\bar{Y} = f(X)$ случайной величины *X*:

$\bar{Y} = f(X)$	0,375	0,556	0,625
p	0,16	0,36	0,48

Из этой таблицы находим:

$$M(\bar{Y}) = M(f(X)) = 0{,}56; \quad \sigma(\bar{Y}) = \sigma(f(X)) \approx 0{,}08665;$$

$$\eta(Y,X) = \frac{\sigma(\bar{Y})}{\sigma(Y)} = \frac{0{,}08665}{0{,}496} \approx 0{,}175.$$

Величина $\eta(Y,X)$ оказалась большей, чем $\rho(X,Y)$- так и должно, согласно (6.44), быть. Однако и она невелика, что свидетельствует о малой тесноте корреляционной зависимости *Y* и *X*. А так как различие между $\rho(X,Y)$ и $\eta(Y,X)$ незначительное, то корреляционная зависимость *Y* от *X* близка к линейной.

Этот вывод должна подтвердить линия регрессии $\bar{y}_x = f(x)$. Ее следует строить по трем точкам:

x_i	0	1	2
$\bar{y}_{x_i}$	0,375	0,556	0,625

Как легко убедиться, ломаная, соединяющая эти три точки, действительна близка к прямой линии.

Понятие о множественной корреляции.

Случайная величина *Y* может корреляционно (в среднем) зависеть не от одной, а от нескольких случайных величин $(X_1; X_2; ...X_n)$. Например, урожайность *Y* любой культуры очевидным образом корреляционно (в среднем) зависит от количества X_1 внесённых под неё удобрений, количества X_2 выпавших осадков, температуры X_3 воздуха и т.д. Такая корреляционная зависимость называется *множественной корреляцией*.

Пусть, например, случайная величина *Z* (дискретная или непрерывная) корреляционно зависит от некоторых двух случайных величин *X* и *Y*. Наличие такой корреляционной зависимости означает, что среднее значение $\bar{z}_{xy}$ величины *Z*, соответствующее значениям (*x; y*) величин *X* и *Y*, зависит от этих (*x; y*):

$$\bar{z}_{xy} = f(x; y) \neq Const \qquad (6.47)$$

Уравнение (6.47) называется *уравнением регрессии Z на X и Y*. В частности, если $f(x; y)$ – линейная функция своих аргументов, то есть если

$$\bar{z}_{xy} = ax + by + c$$

то корреляционная зависимость *Z* от *X* и *Y* называется *линейной*. В противном случае она называется *нелинейной*.

При исследовании множественной корреляционной зависимости, как и при исследовании парной (*Y* от *X)* корреляционной зависимости, ставятся те же две основные задачи:

1) Нахождение *уравнения регрессии*, выражающего зависимость среднего значения одной случайной величины от значений других случайных величин.

2) *Оценка тесноты* исследуемой корреляционной зависимости.

Исследование множественной корреляции – задача несравненно более сложная, чем исследование парной корреляции. При этом исследовании приходится выяснить наличие, характер и степень тесноты зависимости между каж-

дой парой рассматриваемых случайных величин, находить некие коэффициенты их групповой взаимозависимости, и т. д. В практическом плане множественную корреляцию (впрочем, как и парную) исследуют в математической статистике – науке, которой посвящена вторая часть этой книги. Это исследование обычно производят с помощью специальных программ корреляционно–регрессионного анализа на ЭВМ.

Упражнения.

1. Доказать следующие свойства коэффициента линейной корреляции:

а) $\rho(X,X)=1$; б) $\rho(X+C_1,Y+C_2)=\rho(X,Y)$;

в) $\rho(C_1X,C_2Y)=\begin{cases}\rho(X,Y), & \text{если } C_1 \text{ и } C_2 - \text{одного знака}\\ -\rho(X,Y), & \text{если } C_1 \text{ и } C_2 - \text{разных знаков}\end{cases}$

2. Выше было доказано, что если корреляционное отношение $\eta(Y,X)=0$, то случайная величина *Y* корреляционно от X не зависит. А если $\eta(Y,X)=1$, то *Y* является функцией от *X*. Доказать, что верны и обратные этим утверждения.

3. Используя данные примера, приведённого в конце параграфа, выполнить его задания, поменяв *X* и *Y* местами.

Ответ: 1) $\rho(Y,X)=\rho(X,Y)\approx 0.167$; 2) $\eta(X,Y)=\rho(X,Y)\approx 0.167$; 3) Линия регрессии $\bar{x}_y=g(y)$ - прямая, проходящая через две точки:

y_j	0	1
$\bar{x}_{y_j}$	1,1818	1,4286

§7. Закон больших чисел.

Закон больших чисел является центральным законом теории вероятностей в силу того, что он формулирует принципиальную связь между закономерностью и случайностью. А именно, он утверждает, что большое число случайностей ведет к закономерности. В наиболее общей форме он выражается *теоремой Чебышева:*

Пусть (X_1; X_2; … X_n; …) независимые случайные величины (их, предполагается, бесконечное число). И пусть их дисперсии равномерно ограничены (то есть дисперсии всех этих случайных величин не превосходят некоторой константы *C*):

$$D(X_k)\le C \qquad (k=1,2,3,\ldots) \tag{7.1}$$

Тогда как бы ни было мало положительное число ε, выполняется предельное вероятностное соотношение:

$$\lim_{n\to\infty} p\left(\left|\frac{X_1+X_2+\ldots+X_n}{n}-\frac{M(X_1)+M(X_2)+\ldots+M(X_n)}{n}\right|\prec\varepsilon\right)=1 \tag{7.2}$$

Или, что одно и то же, вероятность

$$p\left(\left|\frac{X_1+X_2+...+X_n}{n}-\frac{M(X_1)+M(X_2)+...+M(X_n)}{n}\right|\prec\varepsilon\right)\xrightarrow[n\to\infty]{}1 \qquad (7.3)$$

Таким образом, теорема Чебышева утверждает, что если рассматривать достаточное большое число n независимых случайных величин (X_1; X_2; … X_n), то почти достоверным (с вероятностью, близкой к единице) можно считать событие, состоящее в том, что отклонение среднего арифметического этих случайных величин от среднего арифметического их математических ожиданий будет по абсолютной величине сколь угодно малым.

Доказательство. Введем в рассмотрение новую случайную величину $\overline{X}$ - среднее арифметическое случайных величин (X_1; X_2; … X_n):

$$\overline{X}=\frac{X_1+X_2+...+X_n}{n} \qquad (7.4)$$

Пользуясь свойствами математического ожидания и дисперсии, для $\overline{X}$ получим:

$$M(\overline{X})=\frac{M(X_1)+M(X_2)+...+M(X_n)}{n};\quad D(\overline{X})=\frac{D(X_1)+D(X_2)+...+D(X_n)}{n^2} \qquad (7.5)$$

Учитывая условия (7.1), устанавливаем, что

$$D(\overline{X})\le\frac{C+C+...+C}{n^2}=\frac{nC}{n^2}=\frac{C}{n} \qquad (7.6)$$

Таким образом, при $n\to\infty$ дисперсия $D(\overline{X})\to 0$. То есть при $n\to\infty$ разброс значений случайной величины $\overline{X}$ вокруг её математического ожидания $M(\overline{X})$ неограниченно уменьшается. А это значит, что при $n\to\infty$ величина $\overline{X}\to M(\overline{X})$, то есть $\overline{X}-M(\overline{X})\to 0$. Или, если сказать точнее, к нулю стремиться вероятность того, что случайная величина $\overline{X}$ будет хоть как-то отклоняться от своего математического ожидания – константы $M(\overline{X})$. А именно, при любом сколь угодно малом положительном числе ε

$$p(\left|\overline{X}-M(\overline{X})\right|\prec\varepsilon)\xrightarrow[n\to\infty]{}1 \qquad (7.7)$$

Итак, согласно доказанной теореме Чебышева, среднее арифметическое $\overline{X}$ большого числа независимых случайных величин (X_1; X_2; … X_n), являясь случайной величиной, фактически утрачивает характер случайности, становясь, по сути, неизменной константой $M(\overline{X})$. Эта константа равна среднему арифметическому математических ожиданий величин (X_1; X_2; … X_n). В этом и состоит закон больших чисел.

Остановимся на одном важном частном случае теоремы Чебышева. А именно, рассмотрим случай, когда независимые случайные величины (X_1; X_2; … X_n) имеют одинаковые законы распределения, а, следовательно, и одинаковые числовые характеристики:

$$M(X_k)=a;\quad D(X_k)=D;\quad \sigma(X_k)=\sqrt{D}=\sigma\quad (k=1,2,3,...) \qquad (7.8)$$

Тогда для случайной величины $\overline{X}$, согласно (7.5), имеем:

$$M(\overline{X})=a;\ D(\overline{X})=\frac{D}{n};\ \sigma(\overline{X})=\frac{\sigma}{\sqrt{n}}\to 0\ \text{при}\ n\to\infty \qquad (7.9)$$

Предельное вероятностное соотношение (7.7) в этом случае примет вид:

$$p(\left|\overline{X}-a\right|\prec\varepsilon)\xrightarrow[n\to\infty]{}1 \qquad (7.10)$$

Вывод, следующий из (7.10), имеет большое значение для борьбы со случайными ошибками при производстве различного рода измерений.

Пусть, например, требуется измерить некоторую величину *a*. Произведем не одно, а несколько (*n*) независимых повторных измерений значения этой величины. Любым измерениям присуща случайная ошибка, связанная с несовершенством измерительного прибора, всевозможными случайными помехами измерению, и т.д. Поэтому результаты (X_1; X_2; … X_n) отдельных последовательных измерений искомого значения *a*, вообще говоря, давать не будут - они будут случайными величинами. Причем величинами, имеющими одинаковые распределения, ибо измерения производятся повторно, то есть при неизменных внешних условиях. Тогда для величины $\overline{X}$ - среднего арифметического из результатов всех *n* измерений - будет выполняться предельное вероятностное соотношение (7.10). А значит, это среднее арифметическое $\overline{X}$ при $n \to \infty$ утрачивает характер случайности, превращаясь в *a* – истинное значение измеряемой величины. Об этом, кстати, свидетельствуют и формулы (7.9), согласно которым:

$$M(\overline{X}) = a; \qquad \sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}} \to 0 \text{ при } n \to \infty \tag{7.11}$$

То есть проведя достаточно большое число повторных измерений искомой величины *a*, в каждом из которых возможна случайная ошибка измерения, и находя затем среднее арифметическое $\overline{X}$ результатов этих измерений, мы по формуле

$$a \approx \overline{X} \tag{7.12}$$

можем получить значение *a практически без случайной ошибки*.

Этот вывод - следствие закона больших чисел. В данном случае этот закон проявляется в том, что при суммировании результатов измерений в (7.4) случайные ошибки отдельных измерений, встречающиеся в принципе одинаково часто как со знаком плюс, так и со знаком минус, в целом будут взаимно уничтожаться. А оставшаяся ошибка еще разделится на *n*, то есть еще уменьшится в *n* раз. Так что при больших значениях *n* величина $\overline{X}$ будет практически точно равна измеряемой величине *a*. Этим выводом, естественно, широко пользуются на практике.

Примечание. В величине $\overline{X}$ взаимно уничтожаются лишь *случайные ошибки* измерений, то есть ошибки, связанные с действием случайных факторов (помех). Но систематические (постоянно действующие) ошибки, то есть ошибки, присущие каждому измерению, естественно, остаются и в $\overline{X}$. Например, сбитая (не отрегулированная) в приборе стрелка вызывает постоянную (систематическую) ошибку в каждом измерении, а значит вызывает её и в средней арифметической $\overline{X}$ из результатов этих измерений. Систематические ошибки надо исключать еще до производства измерений и не допускать в процессе измерений.

Потом, если α – цена деления измерительного прибора, то все повторные измерения производятся с точностью до α. Но тогда, естественно, и среднее

арифметическое $\overline{X}$ из результатов всех измерений можно указывать лишь тоже с точностью до α, то есть с точностью, определяемой точностью прибора.

Поэтому не стоит думать, что сделав достаточно большое число повторных измерений величины *a* и найдя затем среднее арифметическое $\overline{X}$ из результатов этих измерений, мы получим *точное* значение *a*. Мы его получим лишь в пределах точности измерительного прибора. Да и то, если исключим систематическую ошибку измерения.

Приведем еще один важный частный случай закона больших чисел. Пусть $X=k$ – число появлений некоторого события A в n повторных испытаниях (X – случайная величина). И пусть $p(A)=p$ и $p(\overline{A})=1-p=q$ – вероятности появления и непоявления события A в одном испытании. Рассмотрим случайную величину $\overline{X}=\frac{k}{n}$ - относительную частоту появления события A в n испытаниях. Введем также n случайных величин ($X_1, X_2,\ldots X_n$), которые представляют собой число появление события A в первом, втором, … n-ом испытаниях. Тогда $k = X_1+X_2+\ldots+X_n$, а

$$\overline{X}=\frac{k}{n}=\frac{X_1+X_2+\ldots+X_n}{n} \qquad (7.13)$$

Случайные величины ($X_1, X_2,\ldots X_n$) имеют одинаковые и простые законы распределения:

X_k	0	1
p	q	p

$(k=1,2,\ldots n)$ (7.14)

Из (7.14) находим:

$$M(X_k)=a=0\cdot q+1\cdot p=p \qquad (7.15)$$

Предельное вероятностное соотношение (7.10) для рассматриваемого случая примет вид:

$$p\left(\left|\frac{k}{n}-p\right|<\varepsilon\right)\xrightarrow[n\to\infty]{}1 \qquad (7.16)$$

То есть при достаточно большом числе n повторных испытаний можно почти наверняка (с вероятностью, близкой к единице) ожидать, что относительная частота $\frac{k}{n}$ появления события A практически совпадет с вероятностью $p(A)=p$ появления события A в одном испытании. На этом выводе основано нахождение вероятностей многих случайных событий, чьи вероятности каким-то другим путем (теоретически) найти не удается.

Например, пусть испытание – это подбрасывание деформированной (несимметричной) монеты, а событие A для этого испытания – это выпадение герба. Вероятность $p(A)=p$ события A по классической формуле $p(A)=\frac{m}{n}$ или по какой-то другой теоретической формуле найти затруднительно, ибо в такой формуле должны быть как-то отражены характеристики деформации монеты. Поэтому реальный путь, ведущий к цели, здесь один: повторно бросать монету (чем больше число бросаний n, тем лучше) и определять опытным путем отно-

сительную частоту $\frac{k}{n}$ появления герба. Если n велико, то в соответствии с законом больших чисел можно с большой вероятностью утверждать, что $p(A) = p \approx \frac{k}{n}$.

Закон больших чисел проявляет себя во многих природных и общественных явлениях.

Пример 1. Как известно, газ, помещенный в закрытый сосуд, оказывает давление на стенки сосуда. Согласно законам газового состояния, при неизменной температуре газа это давление постоянно. Давление газа имеет своей причиной хаотические удары отдельных его молекул о стенки сосуда. Скорости и направления движения у всех молекул разные, поэтому разными являются и силы ударов различных молекул о стенки сосуда. Однако давление газа на стенки сосуда определяется не силой ударов отдельных молекул, а их *средней* силой. Но она, как средняя из огромного числа независимо действующих сил, согласно закону больших чисел, будет сохранять практически неизменное значение. Поэтому практически неизменным оказывается и давление газа на стенки сосуда.

Пример 2. Страховая компания, занимающаяся, например, автострахованием, выплачивает по разным страховым случаям (автомобильным авариям и ДТП) разные страховые суммы. Однако величина среднего значения этой страховой суммы, как среднего значения из многих различных n независимых страховых сумм, согласно закону больших чисел, будет практически неизменной. Ее можно определить, исследовав реальную статистику страховых случаев. Чтобы страховой компании не оказаться в убытке, средний размер страхового взноса, взимаемого с клиентов этой компании, должен быть выше средней страховой суммы, которую выплачивает компания своим клиентам. Но этот взнос не должен быть и слишком высоким, чтобы компания была конкурентноспособна (могла конкурировать по привлекательности с другими страховыми компаниями).

Упражнения

1. К какому типу случайных величин применим закон больших чисел: дискретным? непрерывным? зависимым? независимым? любым?
2. В некотором эксперименте требуется определить (измерить) некоторую величину. Как исключить возможную случайную ошибку измерения?
3. Непрерывная случайная величина X задана плотностью вероятности

$$f(x) = \frac{1}{\pi(x^2+1)} \quad (-\infty \prec x \prec \infty)$$

(распределение Коши). Показать, что для средней арифметической $\overline{X}$ из независимых случайных величин ($X_1, X_2, \ldots X_n, \ldots$), имеющих распределение Коши, неприменим закон больших чисел. То есть средняя $\overline{X}$ при $n \to \infty$ не стремится по вероятности ни к какому конечному числу.

Часть II

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Математическая статистика – это прикладная, практическая наука, изучающая большие совокупности однотипных объектов (предметов, живых существ, чисел и т.д.) *выборочно*. В 19-ом веке она выделилась из теории вероятностей и с тех пор считается самостоятельной наукой.

Математическая статистика пользуется методами различных разделов математики. Но в первую очередь она пользуется методами теории вероятностей, которая служит для неё основной теоретической базой.

§1. Генеральная совокупность и выборка.

Пусть имеется большая совокупность однотипных объектов (зёрен в ворохе зерна, деревьев в лесу, жителей в стране, предметов массового производства, и т.д.), подлежащая изучению. При этом предметом изучения являются какие-то качественные или количественные параметры объектов, составляющих данную совокупность (скажем, пригодность объектов к использованию, их вес, сорт, размер, и т.д.), законы распределения этих параметров и многое другое (об этом конкретнее будет сказано позже).

Исходная совокупность объектов называется *генеральной совокупностью*, а число N объектов этой совокупности (обычно очень большое и точно не известное) называется *объёмом генеральной совокупности.*

Произвести сплошное обследование (обследование всех объектов) генеральной совокупности, в силу её огромного объёма, не представляется возможным. А если это обследование связано с порчей или даже уничтожением обследуемых объектов (скажем, нас интересует сила, при действии которой объект ломается), то оно и бессмысленно (исследовать все объекты генеральной совокупности – это значит все их переломать). Поэтому изучают только небольшую, случайно отобранную, часть этой совокупности (горсть зёрен из вороха, небольшую часть деревьев леса, случайно отобранных жителей страны, небольшую партию предметов массового производства, и т.д.).

Отобранная совокупность объектов называется *выборочной совокупностью* или, короче, *выборкой*. Количество n объектов, попавших в выборку, называется *объёмом выборки*. Как правило, объём n выборки много меньше объёма N генеральной совокупности ($n \ll N$). Объекты выборки подвергаются сплошному обследованию, а затем, по результатам этого обследования, делаются определенные выводы и обо всей генеральной совокупности.

Естественно, что обследование объектов выборки не даст полной и точной информации о всей генеральной совокупности (ведь обследуется лишь часть объектов этой совокупности). Поэтому любые выводы, касающиеся генеральной совокупности, к которым мы придем на основании исследования выборки, чреваты неточностями и даже ошибками. Но эти ошибки, естественно, будут тем менее вероятны и тем меньше по величине, чем больше будет n – объем выборки. Как именно от объема выборки зависит точность и надежность получаемых выводов о генеральной совокупности – эти вопросы тоже рассматриваются в математической статистике.

Кроме большого объема, для получения достаточно надежных и достоверных выводов о генеральной совокупности выборка должна еще адекватно представлять собой генеральную совокупность. Или, как ещё говорят, она должна быть *репрезентативной*. Это значит, что нельзя отбирать преимущественно лучшие или, наоборот, худшие объекты. Правильным (репрезентативным) будет такой отбор, при котором шансы быть отобранными у всех объектов генеральной совокупности будут одинаковыми. А это будет иметь место лишь в том случае, когда выборку объектов из генеральной совокупности осуществляют *случайно*.

Например, чтобы отбор гости зерна из вороха зерна был произведён репрезентативно, следует взять по щепотке зёрен из разных мест этого вороха (с разных краёв, с поверхности, с глубины, и т.д.). А если этот ворох лежит давно и уже слежался (однородность вороха нарушилась), то перед осуществлением выборки ворох этот хорошо бы и тщательно перемешать.

В тех случаях, когда объекты генеральной совокупности пронумерованы (например, это автомобили, выпускаемые автозаводом, или отдельные части этих автомобилей – моторы, кузова, и т.д.), для случайного отбора каких-то n объектов такой генеральной совокупности можно воспользоваться так называемой *таблицей случайных чисел*. То есть номера отбираемых объектов можно взять из этой таблицы, открыв страницу таблицы наугад. Эту таблицу получают с помощью ЭВМ, и она содержится во многих справочниках по математической статистике. Кстати, числа, содержащиеся в таблице случайных чисел – это просто наборы цифр дробной части случайной величины X, равномерно распределённой на отрезке [0;1].

После того, как выборка произведена, исследуют *каждый объект* этой выборки. То есть выясняют (измеряют, устанавливают) значения тех количественных или качественных признаков отобранных объектов, которые представляют исследовательский интерес в генеральной совокупности. Например, если исследуется выборочным путём ворох зерна, то качественным признаком каждого отобранного зерна может быть годность его к посеву или к использованию в мукомольной промышленности. А количественным признаком – вес зерна, количественно выраженная влажность, процентное содержание белка, клейковины и т.д. Другой пример: если генеральная совокупность представляет собой некоторые изделия массового производства, то качественным признаком каждого отобранного изделия может быть его стандартность, а количественным

– контролируемый размер изделия, или его вес, или время до выхода его из строя, и т.д.

Будем пока считать, что у объектов генеральной совокупности исследуется лишь один признак X, и этот признак – *количественный* (то есть его можно выразить некоторым числом). Это может быть вес, сорт, размер, и т.д. Кстати, при необходимости и качественный признак объектов (например, их годность к своему назначению) можно сделать количественным, если считать, что этот признак X=1, если объект годен, и считать X=0, если объект не годен.

Итак, пусть из изучаемой генеральной совокупности сделана случайная выборка объёмом n. И пусть оказалось, что у n_1 объектов, попавших в выборку, значение исследуемого признака X оказалось равным x_1, у n_2 объектов – значение x_2, …, у n_m объектов – значение x_m. Тогда таблица

x_i	x_1	x_2	…	x_m
n_i	n_1	n_2	…	n_m

$$(n_1 + n_2 + ... + n_m = n) \qquad (1.1)$$

содержащая указанные данные, называется *статистическим распределением выборки*. При этом числа (x_1; x_2;. . . x_m), представляющие собой все встретившиеся в выборке значения исследуемого признака X, называются *вариантами*, а количества (n_1; n_2;. . . n_m) объектов, имеющих соответствующие варианты, называются *частотами*.

Статистическое распределение выборки автоматически имеет вид (1.1), если исследуемый признак X является дискретной (прерывистой) величиной. Например, если исследуется экзаменационная оценка по какому-либо предмету большого количества студентов, то эта оценка X по своей природе является величиной дискретной (принимает лишь значения 2; 3; 4; 5). И если выборка составляет, например, 25 человек, то её статистическое распределение может быть, например, следующим:

x_i	2	3	4	5
n_i	2	8	10	5

(2+8+10+5=25) (1.2)

Статистическое распределение выборки (1.1) для наглядности изображают графически – в виде так называемого *полигона частот*, представляющего собой ломаную линию с узлами в точках $M_i(x_i; n_i)$ $(i = 1, 2, ...m)$ - см. рис.3.1.

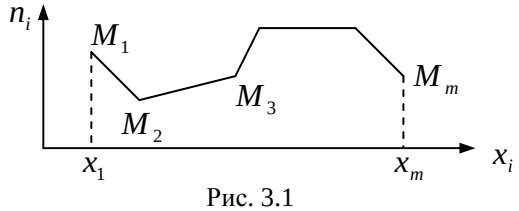


Рис. 3.1

Если же исследуемый признак X является непрерывной величиной, то статистическое распределение выборки обычно оформляют в виде таблицы (1.3.):

x_i	$x_1 - x_2$	$x_2 - x_3$	…	$x_m - x_{m+1}$
n_i	n_1	n_2	…	n_m

$(n_1 + n_2 + ... + n_m = n)$ (1.3)

Здесь $(x_1\text{-}x_2)$, $(x_2\text{-}x_3)$, . . . $(x_m\text{-}x_{m+1})$ – интервалы (обычно одинаковые по длине), на которые разбивают весь интервал $(x_1;\ x_{m+1})$ значений признака X в выборке, а $(n_1;\ n_2;\ .\ .\ .;\ n_{m)}$ – частоты для соответствующих интервалов. Например, если исследуется масса X (г) клубней картофеля, выращенного на некотором поле, то статистическое распределение выборки для 100 клубней, случайно отобранных из выращенного урожая, может быть таким:

x_i	0-40	40-80	80-120	120-160	160-200
n_i	12	20	28	25	15

(1.4)

Графически статистическое распределение выборки вида (1.3) изображается уже не полигоном, а так называемой *гистограммой частот* (рис.3.2.):

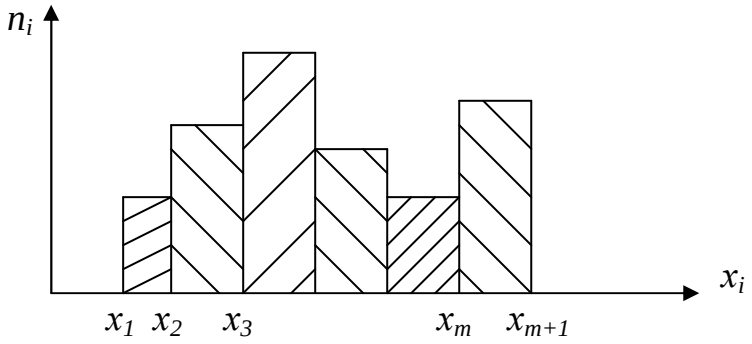


Рис.3.2

Отметим, что часто на оси ординат полигонов и гистограмм откладывают не частоты n_i, а *относительные частоты* $\frac{n_i}{n}$.

Перейдём теперь к основным числовым характеристикам статистического распределения выборки. Ими являются:

1. *Среднее значение признака X в выборке,* обозначаемое $\bar{x}_в$ и называемое *выборочной средней.*

2. Величина $D_в$, которая характеризует *среднее значение квадратов отклонений вариант* x_i *от выборочной средней* $\bar{x}_в$. Она называется *выборочной дисперсией.*

3. Величина $\sigma_в$, которая характеризует *среднее значение отклонения вариант* x_i *от выборочной средней* $\bar{x}_в$ *без учёта знака этого отклонения.* Она называется *выборочным средним квадратическим отклонением.*

4. Величина $V_в\%$, называемая *выборочным коэффициентом вариации*. Этот коэффициент характеризует *долю в процентах, которую составляет среднее отклонение от среднего по отношению к самому среднему.*

Все названные основные числовые характеристики выборки определяются по формулам:

$$\bar{x}_в = \frac{\sum_{i=1}^{m} x_i n_i}{n};\ D_в = \frac{\sum_{i=1}^{m} (x_i - x_в)^2 n_i}{n};\ \sigma_в = \sqrt{D_в};\ V_в\% = \frac{\sigma_в}{\bar{x}_в} \cdot 100\% \qquad (1.5)$$

Эти формулы можно использовать, если статистическое распределение выборки имеет вид (1.1), то есть является дискретным. А если оно имеет вид (1.3), то есть является непрерывным (интервальным), то его предварительно преобразуют в дискретное, в котором середины интервалов принимаются за его новые дискретные варианты.

Заметим, что введённые выше числовые характеристики выборки введены с той же целью и имеют в принципе тот же смысл, что и числовые характеристики случайных величин – математическое ожидание (среднее значение), дисперсия, среднее квадратическое отклонение, коэффициент вариации, о которых шла речь в курсе теории вероятностей. И названия этих характеристик во многом совпадают.

Кстати, формулу для подсчёта выборочной дисперсии $D_в$ можно упростить, если раскрыть в ней квадрат разности, сумму разбить на три суммы и привести затем подобные. В итоге получим следующую *упрощённую формулу для выборочной дисперсии* (выкладки проделайте самостоятельно):

$$D_в = \overline{x_в^2} - (\bar{x}_в)^2 \qquad (1.6)$$

То есть получаем: *выборочная дисперсия равна средней из квадратов вариант выборки минус выборочная средняя в квадрате*. Здесь

$$\bar{x}_в = \frac{\sum_{i=1}^{m} x_i n_i}{n};\qquad \overline{x_в^2} = \frac{\sum_{i=1}^{m} x_i^2 n_i}{n} \qquad \textbf{(1.7)}$$

Пример 1. Дано статистическое распределение выборки

x_i	1	2	3	4
n_i	20	15	10	5

(20+15+10+5=n= 50)

Найти $\bar{x}_в$, $D_в$, $\sigma_в$, $V_в\%$.

Решение. Используя приведённые выше формулы, получим:

$$\bar{x}_в = \frac{1 \cdot 20 + 2 \cdot 15 + 3 \cdot 10 + 4 \cdot 5}{50} = \frac{100}{50} = 2$$

$$\overline{x_в^2} = \frac{1^2 \cdot 20 + 2^2 \cdot 15 + 3^2 \cdot 10 + 4^2 \cdot 5}{50} = \frac{250}{50} = 5$$

$$D_в = \overline{x_в^2} - (\bar{x}_в)^2 = 5 - 2^2 = 1;\ \sigma_в = \sqrt{D_в} = 1;\quad V_в\% = \frac{\sigma_в}{\bar{x}_в} \cdot 100\% = \frac{1}{2} \cdot 100\% = 50\%$$

Числовые характеристики выборки ($\bar{x}_в$, $D_в$, $\sigma_в$, $V_в\%$), если они найдены, служат для оценки соответствующих числовых характеристик ($\bar{x}_Г$, $D_Г$, $\sigma_Г$, $V_Г\%$) *генеральной совокупности*.

Отметим, что числовые характеристики генеральной совокупности – *фиксированные*, хотя и неизвестные, числа. А числовые характеристики выборки очевидным образом зависят от того, какие объекты генеральной совокупности

попали в выборку. От выборки к выборке эти объекты меняются. А так как выборка объектов производится случайно, то и числовые характеристики выборки – *случайные величины*. А значит, возникают естественные вопросы о законах распределения этих случайных величин, их числовых характеристиках и т.д. Обо всём это пойдёт речь в следующем параграфе.

Упражнения

1. В чём достоинства и в чём недостатки исследования всей генеральной совокупности и исследования выборки из неё?
2. Пусть X – месячная зарплата на сдельной работе одного рабочего на некотором предприятии. Она исследовалась по бухгалтерским ведомостям выборочно. Какой смысл в этом случае будут иметь величины ($\bar{x}_в$, $D_в$, $\sigma_в$, $V_в$%)? И какой смысл будут иметь ($\bar{x}_Г$, $D_Г$, $\sigma_Г$, $V_Г$%)?
3. Статистическое распределение выборки имеет следующий вид:

x_i	1-3	3-5	5-7	7-9
n_i	20	15	10	5

Найти числовые характеристики выборки.

Ответ: $\bar{x}_в$=4; $D_в$=4; $\sigma_в$=2; $V_в$%=50%.

§2. Числовые характеристики выборочной средней и выборочной дисперсии. Оценки числовых характеристик генеральной совокупности

Как отмечено в конце предыдущего параграфа, числовые характеристики выборки $\bar{x}_в$; $D_в$; $\sigma_в$; $V_в$% являются случайными величинами. В связи с этим возникает естественный и важный для практики вопрос о математическом ожидании, дисперсии и прочих числовых характеристиках этих случайных величин.

Начнём с важнейшей из этих величин – с выборочной средней $\bar{x}_в$. Будем считать, что объём N генеральной совокупности настолько велик, что объём n выборки можно считать малой величиной по сравнению с N. Поэтому последовательный отбор из генеральной совокупности каждого отбираемого объекта практически не нарушает состава генеральной совокупности – она как бы всё время остаётся целой.

Обозначим через (X_1; X_2; …X_n) случайные величины, выражающие значения исследуемого признака X при отборе первого, второго, … n-ого объектов выборки. С учётом предположения, сделанного выше относительно объёма ге-

неральной совокупности, можем считать, что случайные величины (X_1; X_2; …X_n) одинаково распределены и независимы. Распределение каждой из них совпадает с с распределением величины X_1 - с распределением признака X у первого отобранного объекта. Если (x_1; x_2; ... x_p) – список всех возможных значений исследуемого признака X в генеральной совокупности, то случайная величина X_1 имеет возможность принять любое из этих значений. А их вероятности будут, очевидно, равны ($\frac{N_1}{N}$; $\frac{N_2}{N}$;…$\frac{N_p}{N}$), где (N_1; N_2;…N_p) – количества объектов генеральной совокупности, имеющих соответственно значения (x_1; x_2; ... x_p). Таким образом, закон распределения величины X_1, а вместе с нею и остальных случайных величин (X_2; X_3; …X_n), будет иметь вид:

X_k	x_1	x_2	…	x_p
p	$\frac{N_1}{N}$	$\frac{N_2}{N}$	…	$\frac{N_p}{N}$

(k=1, 2, …n) (2.1)

Выборочная средняя $\bar{x}_в$ - это, очевидно, средняя арифметическая из случайных величин (X_1; X_2; …X_n):

$$\bar{x}_в = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{X_1}{n} + \frac{X_2}{n} + \frac{X_n}{n} \qquad (2.2)$$

Если объём n выборки достаточно большой (хотя и много меньше объёма N генеральной совокупности), то согласно (2.2) выборочная средняя $\bar{x}_в$ является суммой большого числа независимых случайных величин. А потому, согласно теореме Ляпунова (часть I, глава 2, §4), можем считать, что случайная величина $\bar{x}_в$ распределена приблизительно по нормальному закону. Причём это будет *при любом законе распределения* величин (X_1; X_2; …X_n), то есть *при любом законе распределения признака X в генеральной совокупности*. А если есть основания считать, что признак X в генеральной совокупности распределён нормально (что обычно и имеет место), то распределение случайной величины $\bar{x}_в$, как суммы независимых нормально распределённых случайных величин, будет нормальным при любом, в том числе и малом, объёме n выборки.

Действительно, распределения случайных величин (X_1; X_2; …X_n) при очень большом объёме генеральной совокупности можно считать совпадающими с распределением величины X_1. Но если величина X_1 приняла некоторое значение x_i, то это значит, что признак X принял это значение. То есть распределения X и X_1, а значит распределения X и (X_1; X_2; …X_n) совпадают. И все эти величины являются нормальными, если распределение величины X нормальное. Но тогда и все слагаемые ($\frac{X_1}{n}$;$\frac{X_2}{n}$;…$\frac{X_n}{n}$) в (2.2) распределены нормально, а вместе с ними, в силу их независимости, и величина $\bar{x}_в$ распределена нормально.

В общем, так или иначе, мы практически всегда можем считать величину $\bar{x}_в$ распределённой нормально (мы не можем этого утверждать лишь в случае,

когда выборка имеет малый объём n и при этом исследуемый признак X заведомо не распределён нормально).

Найдём параметры $a = M(\bar{x}_в)$ и $\sigma = \sigma(\bar{x}_в)$ нормально распределённой случайной величины $\bar{x}_в$. Для этого сначала вычислим математическое ожидание и дисперсию случайных величин $X_к$ (k=1, 2,…n). В соответствии с таблицей (2.1) и формулами (1.4) и (1.23) (часть I, глава 2) имеем:

$$M(X_k) = \sum_{i=1}^{p} x_i p_i = \sum_{i=1}^{p} x_i \cdot \frac{N_i}{N} = \frac{\sum_{i=1}^{p} x_i N_i}{N} = \bar{x}_Г \tag{2.3}$$

$$D(X_k) = \sum_{i=1}^{p} (x_i - \bar{x}_Г)^2 p_i = \sum_{i=1}^{p} (x_i - \bar{x}_Г)^2 \frac{N_i}{N} = \frac{\sum_{i=1}^{p} (x_i - \bar{x}_Г)^2 N_i}{N} = D_Г$$

Здесь $\bar{x}_Г$ и $D_Г$ – числовые характеристики генеральной совокупности (генеральная средняя и генеральная дисперсия). А тогда на основании (2.2) и свойств математического ожидания и дисперсии (см. часть I, глава 2) получим:

$$M(\bar{x}_в) = \bar{x}_Г; \quad D(\bar{x}_в) = \frac{D_Г}{n}; \quad \sigma(\bar{x}_в) = \sqrt{\frac{D_Г}{n}} = \frac{\sigma_Г}{\sqrt{n}} \tag{2.4}$$

Итак, нормально распределённая случайная величина $\bar{x}_в$ распределена с параметрами $a = \bar{x}_Г$ и $\sigma = \frac{\sigma_Г}{\sqrt{n}}$.

Свои числовые характеристики имеет и выборочная дисперсия $D_в$. Она уже заведомо не распределена нормально, ибо по природе своей имеет лишь неотрицательные значения. Из ее числовых характеристик приведем лишь важнейшую – математическое ожидание (среднее значение):

$$M(D_в) = \frac{n-1}{n} D_Г. \tag{2.5}$$

Кстати, если объем n выборки достаточно велик, то $\frac{n-1}{n} \approx 1$, и тогда можно считать, что $M(D_в) = D_Г$.

Докажем формулу (2.5). Согласно определению (1.5) выборочной дисперсии, ее можно выразить через введенные выше случайные величины ($X_1, X_2, \ldots X_n$) формулой:

$$D_в = \frac{(X_1 - \bar{x}_в)^2 + (X_2 - \bar{x}_в)^2 + \ldots + (X_n - \bar{x}_в)^2}{n} \tag{2.6}$$

Тогда

$$M(D_в) = \frac{1}{n}\left[M(X_1 - \bar{x}_в)^2 + M(X_2 - \bar{x}_в)^2 + \ldots + M(X_n - \bar{x}_в)^2\right] \tag{2.7}$$

Так как случайные величины ($X_1, X_2, \ldots X_n$) имеют одинаковые распределения, то все n слагаемых в сумме (2.7) одинаковы, и поэтому

$$M(D_в) = \frac{1}{n} \cdot n \cdot M(X_1 - \bar{x}_в)^2 = M(X_1 - \bar{x}_в)^2 \tag{2.8}$$

То есть

$$M(D_в) = M\left[X_1^2 - 2X_1 \cdot \bar{x}_в + (\bar{x}_в)^2\right] = M\left(X_1^2\right) - 2M\left(X_1 \cdot \bar{x}_в\right) + M(\bar{x}_в)^2 \tag{2.9}$$

Найдем каждое из трех слагаемых, входящих в (2.9).

1) Найдем $M(X_1^2)$. Так как

$$D(X_1) = M\left(X_1^2\right) - [M(X_1)]^2,$$

то

$$M\left(X_1^2\right) = D(X_1) + [M(X_1)]^2 = D_Г + (\bar{x}_Г)^2 \qquad (2.10)$$

2) Найдем $M(X_1 \cdot \bar{x}_в)$:

$$M(X_1 \cdot \bar{x}_в) = M\left(X_1 \cdot \frac{X_1 + X_2 + ... + X_n}{n}\right) = \frac{1}{n} M\left(X_1^2 + X_1X_2 + X_1X_3 + ... + X_1X_n\right) =$$

$$= \frac{1}{n}\left[M\left(X_1^2\right) + M(X_1X_2) + M(X_1X_3) + ... + M(X_1X_n)\right]$$

Так как величины ($X_1, X_2, \ldots X_n$) независимы, то

$$M(X_1 \cdot \bar{x}_в) = \frac{1}{n}\left[M\left(X_1^2\right) + M(X_1) \cdot M(X_2) + M(X_1) \cdot M(X_3) + ... + M(X_1) \cdot M(X_n)\right] =$$

$$= \frac{1}{n}\left[D_Г + \bar{x}_Г{}^2 + \bar{x}_Г \cdot \bar{x}_Г + \bar{x}_Г \cdot \bar{x}_Г + ... + \bar{x}_Г \cdot \bar{x}_Г\right] = \frac{1}{n}\left[D_Г + n \cdot (\bar{x}_Г)^2\right] = \frac{D_Г}{n} + (\bar{x}_Г)^2 \qquad (2.11)$$

3) Найдём $M(\bar{x}_в)^2$. Так как

$$D(\bar{x}_в) = M(\bar{x}_в)^2 - [M(\bar{x}_в)]^2,$$

то

$$M(\bar{x}_в)^2 = D(\bar{x}_в) + [M(\bar{x}_в)]^2 = \frac{D_Г}{n} + (\bar{x}_Г)^2 \qquad (2.12)$$

Подставляя выражения (2.10), (2.11) и (2.12) в (2.9), мы и получим доказываемое равенство (2.5).

Анализируя формулы (2.4) для выборочной средней $\bar{x}_в$, видим, что математическое ожидание (среднее значение) выборочной средней равно средней генеральной $\bar{x}_Г$. При этом разброс этих значений $\bar{x}_в$ вокруг $\bar{x}_Г$ будет уменьшаться с увеличением объема *n* выборки, ибо, согласно (2.4),

$$D(\bar{x}_в) \to 0 \quad \text{при } n \to \infty \qquad (2.13)$$

Таким образом, если нам нужно по выборке оценить неизвестную генеральную среднюю $\bar{x}_Г$, то эта оценка будет такой:

$$\bar{x}_Г \approx \bar{x}_в \qquad (2.14)$$

Причем эта оценка будет тем точнее (надежнее), чем больше будет объем *n* выборки.

Анализируя теперь формулу (2.5), видим, что $M(D_в) \neq D_Г$. То есть возможные значения выборочной дисперсии $D_в$ (значения $D_в$ для разных выборок) группируются не вокруг генеральной дисперсии $D_Г$, а вокруг несколько меньшего числа $\frac{n-1}{n} D_Г$. То есть $D_в$ является *смещенной оценкой* $D_Г$. Для устранения этого смещения введем так называемую *исправленную выборочную дисперсию* $s_в^2$:

$$s_в^2 = \frac{n}{n-1} D_в = \frac{\sum_{i-1}^{m} (x_i - \bar{x}_в)^2 n_i}{n-1}. \qquad (2.15)$$

При этом $s_в = \sqrt{s_в^2}$ называется *исправленным выборочным средним квадратическим отклонением*. Математическое ожидание (среднее значение) $s_в^2$ уже равно $D_Г$. Действительно:

$$M\left(s_в^2\right) = M\left(\frac{n}{n-1}D_в\right) = \frac{n}{n-1}M\left(D_в\right) = \frac{n}{n-1}\cdot\frac{n-1}{n}D_Г = D_Г. \qquad (2.16)$$

Таким образом, исправленная выборочная дисперсия $s_в^2$ имеет среднее значение, равное генеральной дисперсии и, таким образом, является *несмещенной оценкой для генеральной дисперсии* $D_Г$*:*

$$D_Г \approx s_в^2 \qquad (2.17)$$

Замечание. Исправленная выборочная дисперсия $s_в^2$, согласно её выражению (2.15), является суммой квадратов отклонений вариант x_i выборки от их среднего значения $\bar{x}_в$, *рассчитанной на одну степень свободы этой суммы*.

Действительно, объём выборки равен *n*, значит, и всех вариант x_i в выборке тоже *n*. Будь в сумме

$$\sum_{i=1}^{m}(x_i - \bar{x}_в)^2 n_i \qquad (2.18)$$

все эти варианты независимыми, эта сумма квадратов отклонений вариант x_i от выборочной средней $\bar{x}_в$ имела бы *n* степеней свободы - по числу независимых вариант x_i, участвующих в формировании этой суммы. Однако эти варианты в сумме (2.18) не являются независимыми, ибо через них по первой из формул (1.5) вычисляется выборочная средняя $\bar{x}_в$, фигурирующая в этой сумме. Формула (1.5) представляет собой линейное соотношение

$$\sum_{i=1}^{m} x_i n_i = n\cdot\bar{x}_в\,, \qquad (2.19)$$

связывающее варианты x_i. Из него одну из вариант x_i (любую) можно выразить через остальные *n*-1 вариант. Так что в сумме (2.18) содержится только *n*-1 независимых слагаемых, в силу чего она имеет не *n*, а *n*-1 степеней свободы. Так что исправленная выборочная дисперсия $s_в^2$ (несмещённая оценка генеральной дисперсии $D_Г$), согласно (2.15), действительно представляет собой сумму (2.18), рассчитанную на её одну степень свободы. Такое истолкование исправленной выборочной дисперсии (несмещённой оценки дисперсии генеральной) нами ещё позднее не раз будет использоваться.

Исходя из оценок (2.14) и (2.17), можем получить еще две оценки для неизвестных числовых характеристик генеральной совокупности:

$$\sigma_Г = \sqrt{D_Г} \approx s_в; \quad V_Г\% = \frac{\sigma_Г}{\bar{x}_Г}\cdot 100\% \approx \frac{s_в}{\bar{x}_в}\cdot 100\% \qquad (2.20)$$

Оценки (2.14), (2.17) и (2.20) числовых характеристик $\bar{x}_Г, D_Г, \sigma_Г, V_Г\%$ генеральной совокупности называются *точечными оценками*, ибо эти оценки осуществляются одним числом (точкой). Все эти оценки несмещённые, и они тем точнее (надёжнее), чем больше объем *n* выборки.

Кроме точечных оценок числовых характеристик генеральной совокупности, вводятся также и их *интервальные оценки.*

Пусть, например, $y_в$ - некоторая из выборочных числовых характеристик ($\bar{x}_в$, или $s_в^2$, или $s_в$, и т.д.), а $y_Г$ - соответствующая ей генеральная характеристика. И пусть $M(y_в) = y_Г$, так что мы имеем точечную несмещённую оценку $y_Г \approx y_в$. Нас, естественно, интересует точность этой оценки, то есть разность $y_в - y_Г$. Но так как по выборочным данным вычисляется лишь $y_в$, а $y_Г$ неизвестна, то разность эту точно найти нельзя. Её можно лишь попытаться оценить. А именно, можно лишь поставить вопрос: с какой вероятностью γ можно утверждать, что $|y_в - y_Г| < \delta$, где δ - некоторое заданное положительное число? Или, что одно и то же, какова вероятность γ того, что $y_в - \delta \prec y_Г \prec y_в + \delta$?

Интервал $(y_в - \delta;\ y_в + \delta)$ называется *доверительным интервалом* для оценки $y_Г$; число δ называется *точностью интервальной оценки* $y_Г$; вероятность γ называется *надежностью интервальной оценки* $y_Г$. Все эти понятия связываются в следующем равенстве:

$$p(|y_в - y_Г| < \delta) = p(y_в - \delta < y_Г < y_в + \delta) = \gamma \qquad (2.21)$$

Геометрическая иллюстрация этого равенства изображена на рис. 3.3. Этот рисунок иллюстрирует смысл равенства (2.21): с вероятностью γ неизвестная $y_Г$ содержится в своем доверительном интервале $(y_в - \delta;\ y_в + \delta)$.

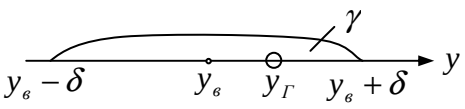


Рис. 3.3

Очевидно, чем шире доверительный интервал (то есть чем больше δ), тем больше надежность (вероятность) γ того, что $y_Г$ принадлежит этому интервалу. И наоборот, чем у́же доверительный интервал (меньше δ), тем меньше вероятность γ того, что $y_Г$ содержится в этом интервале. Заметим, что широкий доверительный интервал означает малую точность оценки величины $y_Г$, а узкий – наоборот, высокую. Таким образом, чем выше точность оценки $y_Г$, тем меньше её надежность, а чем ниже точность – тем больше надежность, что вполне закономерно.

Нас, естественно, будет интересовать конкретная математическая связь между шириной доверительного интервала $(y_в - \delta;\ y_в + \delta)$ и вероятностью γ того, что $y_Г$ содержится в этом интервале. Очевидно, что наиболее важно ответить на этот вопрос, когда $y_Г = \bar{x}_Г$. Этим мы и ограничимся.

Как отмечалось выше, точечной оценкой для $\bar{x}_Г$ является $\bar{x}_в$, которая распределена нормально с математическим ожиданием $a = \bar{x}_Г$ и средним квадратическим отклонением $\sigma = \dfrac{\sigma_Г}{\sqrt{n}}$ (формулы (2.4)). Заменяя в (2.21) $y_Г$ на $\bar{x}_Г$, $y_в$ на $\bar{x}_в$ и пользуясь формулой (4.11) (часть I, глава 2) для нормально распределённых случайных величин, получим:

$$p\left(\left|\bar{x}_в - \bar{x}_Г\right| \prec \delta\right) = p\left(\bar{x}_в - \delta \prec \bar{x}_Г \prec \bar{x}_в + \delta\right) = 2\Phi\left(\frac{\delta\sqrt{n}}{\sigma_Г}\right) = \gamma \qquad (2.22)$$

Таким образом, точность δ, определяющая ширину доверительного интервала $(\bar{x}_в - \delta;\ \bar{x}_в + \delta)$ для оценки генеральной средней $\bar{x}_Г$, и надёжность γ этой оценки связаны друг с другом равенством:

$$2\Phi\left(\frac{\delta\sqrt{n}}{\sigma_Г}\right) = \gamma, \text{ откуда } \delta = t_\gamma \frac{\sigma_Г}{\sqrt{n}}, \text{ где } 2\Phi(t_\gamma) = \gamma \qquad (2.23)$$

Неизвестное среднее квадратическое отклонение генеральной совокупности $\sigma_Г$ можно в (2.23) заменить, согласно (2.20), его точечной оценкой $s_в$. Однако эта замена будет достаточно точной, а значит, и оправданной лишь при достаточно большом объеме *n* выборки (скажем, при *n*>30).

Пример 1. Выборочным путём были получены следующие данные об урожайности ржи в некотором зерновом регионе (таблица 2.24). Найти вероятность того, что средняя урожайность ржи, полученная в выборке, отличается в ту или в другую сторону (то есть по абсолютной величине) от средней урожайности на всей площади региона, занятой под рожь, не более чем на 0,2 ц/га.

Урожайность (*ц/га*)	Площадь (*га*)
12-14	18
14-16	57
16-18	109
18-20	136
20-22	83
22-24	66
24-26	31
Итого	500

(2.24)

Решение. Площадь региона, занятая рожью, нам неизвестна. Но она нам и не нужна. Важно лишь, чтобы она была намного больше тех 500 га, которые попали в выборку, что мы и будем предполагать. В соответствии с условием задачи нам требуется найти вероятность (надёжность) γ того, что $\left|\bar{x}_в - \bar{x}_Г\right| \prec 0{,}2$, где $\bar{x}_в$ - средняя урожайность ржи в выборке, а $\bar{x}_Г$ - средняя урожайность ржи во всем регионе. Эту надежность γ найдем по формуле (2.23). В ней следует положить $\delta = 0{,}2$, $n = 500$ (n-велико!), а величину $\sigma_Г$ заменим на $s_в$:

$$\gamma = 2\Phi\left(\frac{0{,}2 \cdot \sqrt{500}}{s_в}\right) \qquad (2.25)$$

Исправленное выборочное среднее квадратическое отклонение $s_в$ найдем из статистического распределения выборки (2.24). Для этого сначала приведем его к дискретному виду:

x_i (урожайность в ц/га)	13	15	17	19	21	23	25
n_i (площадь в га)	18	57	109	136	83	66	31

(2.26)

Применяя затем формулы (1.5), получим:

$$\bar{x}_в = 19{,}124; \quad D_в = 8{,}985$$

А тогда, согласно (2.15), получаем:

$$s_в^2 = \frac{500}{499} \cdot 8{,}985 \approx 9{,}00; \quad s_в \approx 3.$$

Подставляя найденное значение $s_в$ в (2.25), получим искомую вероятность (надежность) γ:

$$\gamma = 2\Phi(1{,}49) = 2 \cdot 0{,}4319 \approx 0.86$$

Итак, с надежностью $\gamma = 0.86$ (с 86%-ой надежностью) можем утверждать, что средняя урожайность ржи $\bar{x}_Г$ во всем регионе отличается от средней урожайности $\bar{x}_в \approx 19{,}1$ ц/га на обследованных выборочно 500 га по абсолютной величине не более, чем на 0,2 ц/га. Или, что одно и тоже, с 86%-ой надежностью можем утверждать, что $\bar{x}_Г$ находится в следующем доверительном интервале:

$$18{,}9 \text{ ц/га} < \bar{x}_Г < 19{,}3 \text{ ц/га}.$$

Если объем выборки небольшой (n<30), то пользоваться формулой (2.23), просто заменив в ней $\sigma_Г$ на $s_в$, не рекомендуется, ибо $s_в$ может значительно отличаться от $\sigma_Г$, а значит, могут получаться слишком грубые результаты. Но если исследуемый признак X распределен нормально, то доказано, что при любых, в том числе малых, объемах n выборки случайная величина

$$T = \frac{\bar{x}_в - \bar{x}_Г}{\frac{s_в}{\sqrt{n}}} \tag{2.27}$$

имеет распределение Стьюдента с $k = n - 1$ степенями свободы (см [2], глава 16, §16). И поскольку плотность вероятности $f_k(t)$ такой случайной величины известна и является четной функцией своего аргумента t (см. часть I, глава 2, §4), то для любого t_γ вероятность γ осуществления неравенства $-t_\gamma < T < t_\gamma$ найдется по формуле (2.28), следующей из формулы (3.6) главы 2, часть I:

$$p(-t_\gamma \prec T \prec t_\gamma) = p\left(-t_\gamma \prec \frac{\bar{x}_в - \bar{x}_Г}{\frac{s_в}{\sqrt{n}}} \prec t_\gamma\right) = 2\int_0^{t_\gamma} f_k(t)dt = \gamma \tag{2.28}$$

Или, что одно и то же:

$$p(\bar{x}_в - \delta < \bar{x}_Г < \bar{x}_в + \delta) = \gamma, \text{ где } \delta = t_\gamma \frac{s_в}{\sqrt{n}}, \tag{2.29}$$

а величина t_γ связана с γ и $k = n - 1$ последним равенством (2.28). Составлена специальная таблица (см. таблицу 4 Приложения) – так называемая таблица критических точек распределения Стьюдента, позволяющая по заданным $k = n - 1$ и γ находить t_γ. А значит, в соответствии с (2.29), находить величину δ, определяющую доверительный интервал $(\bar{x}_в - \delta;\ \bar{x}_в + \delta)$ для оценки генеральной средней $\bar{x}_Г$ с надежностью (вероятностью) γ. И это – для любых значениях n, в том числе и для малых. При больших же n (n>30) указанный доверительный интервал, найденный посредством вычисления δ как по формуле (2.23) при замене в ней $\sigma_Г$ на $s_в$, так и по формуле (2.29), оказывается практически одинаковым.

Пример 2. Девять независимых повторных измерений некоторой величины *a* дали следующие результаты:

1,24; 1,26; 1,25; 1,23; 1,25; 1,24; 1,24; 1,25; 1,24

Оценить с помощью доверительного интервала истинное значение *a* измеряемой величины с надёжностью $\gamma = 0{,}95$ (95%-ой надёжностью).

Решение. Будем рассматривать результаты всех девяти повторных измерений величины *a* как выборочные значения случайной величины *X* – результата отдельного измерения этой величины. Тогда статистическое распределение выборки будет иметь вид:

x_i	1,23	1,24	1,25	1,26
n_i	1	4	3	1

(n=1+4+3+1=9) (2.30)

А генеральной совокупностью в данном случае будет, очевидно, бесконечное множество всех возможных значений одного измерения ($N=\infty$). Случайная величина *X*, как мы знаем (§4, глава 2) распределена нормально. Её параметры *a* и σ должны быть приняты за $\bar{x}_Г$ и $\sigma_Г$. При этом *a* - это искомое значение измеряемой величины.

Так как объём *n* выборки невелик (n=9), то для интервальной оценки $\bar{x}_Г = a$ следует использовать равенства (2.29).

Исходя из статистического распределения выборки (2.30), найдём $\bar{x}_в$ и $s_в$:

$$\bar{x}_в = \frac{1{,}23\cdot 1 + 1{,}24\cdot 4 + 1{,}25\cdot 3 + 1{,}26\cdot 1}{9} = 1{,}2444$$

$$\overline{x_в^2} = \frac{1{,}23^2\cdot 1 + 1{,}24^2\cdot 4 + 1{,}25^2\cdot 3 + 1{,}26^2\cdot 1}{9} = 1{,}5487$$

$$D_в = \overline{x_в^2} - (\bar{x}_в)^2 = 1{,}5487 - (1{,}2444)^2 = 0{,}0001687$$

$$s_в^2 = \frac{9}{9-1}\cdot 0{,}0001687 = 0{,}0001897;\quad s_в = \sqrt{0{,}0001897} = 0{,}0138$$

Далее по заданным γ=0,95 и k =n-1=9-1=8 с помощью таблицы 4 Приложения найдём значение t_γ, входящие в выражение (2.29) для δ: t_γ=2,31. Таким образом,

$$\delta = t_\gamma \frac{s_в}{\sqrt{n}} = 2{,}31\cdot \frac{0{,}0138}{\sqrt{9}} = 0{,}0106.$$

Следовательно, искомый доверительный интервал $(\bar{x}_в - \delta;\ \bar{x}_в + \delta)$, содержащий с надёжностью γ=0,95 (с 95%-ой надёжностью) истинное значение $a = \bar{x}_Г$ измеряемой величины, будет таким: (1,234; 1,255).

В заключении этого параграфа рассмотрим следующий важный для практики вопрос: каков минимальный объём *n* выборки, обеспечивающий оценку неизвестной генеральной средней $\bar{x}_Г$ с заданной точностью δ при заданной надёжности γ?

Очевидно, что искомое минимальное значение $n_{\min}$ объёма выборки при не слишком широком доверительном интервале $(\bar{x}_в - \delta;\ \bar{x}_в + \delta)$, то есть при не

слишком большом δ и при не слишком малой надёжности γ того, что $\bar{x}_Г$ будет содержаться в этом интервале, следует ожидать достаточно большим. И это $n_{\min}$ будет тем больше, чем меньше будет δ (точнее оценка) и чем больше будет γ (надежнее оценка) генеральной средней $\bar{x}_Г$. Поэтому для нахождения этого значения $n = n_{\min}$ первоначально следует использовать формулы (2.23), применяемые при n>30, из которых следует:

$$n = n_{\min} = \left(\frac{t_\gamma \cdot s_в}{\delta} \right)^2, \text{ где } \; 2\Phi(t_\gamma) = \gamma \qquad (2.31)$$

Но если найденное значение $n_{\min}$ окажется небольшим ($n_{\min}$ <30), то тогда для его уточнения следует использовать последнее равенство (2.29), приводящее, кстати, к тому же выражению для $n_{\min}$, что и (2.31). Только t_γ следует находить не из равенства $2\Phi(t_\gamma) = \gamma$, то есть не из таблицы интеграла вероятности $\Phi(x)$, а подбирать из таблицы 4 Приложения для критических точек распределения Стьюдента.

Пример 3. Выборочным путем исследуется зерно, свезённое с убранного поля на элеватор. Требуется определить минимальный объём выборки, проводимой с целью определения средней массы одного зерна, чтобы с вероятностью 0,99 ошибка в определении этой средней массы не превысила по абсолютной величине 0,002 г. По данным предыдущих выборок установлено, что $s_в \approx 0{,}014$г.

Решение. По таблице функции $\Phi(x)$ (по таблице 2 Приложения) из равенства $2\Phi(t_\gamma) = \gamma$=0.99 находим t_γ:

$$\Phi(t_\gamma) = 0{,}495 \rightarrow t_\gamma = 2{,}58$$

Теперь по формуле (2.31) находим искомый минимальный объём выборки:

$$n_{\min} = \left(\frac{2{,}58 \cdot 0{,}014}{0{,}002} \right)^2 \approx 326$$

Упражнения

1. С целью исследования размера X некоторых однотипных изделий, выпускаемых заводом, было случайным образом отобрано 50 изделий. Их распределение по размеру (статистическое распределение выборки) имеет вид:

x_i (см)	107,8--108,0	108,0--108,2	108,2--108,4	108,4--108,6	108,6--108,8	108,8--109,0
n_i	1	4	16	18	8	3

Найти доверительный интервал, оценивающий с надёжностью $\gamma = 0{,}95$ средний размер изделий, выпускаемых заводом.

Ответ: (108,39; 108,51).

2. При определении экспериментальным путём значения некоторой величины *a* проведено 5 повторных опытов, которые дали следующие результаты:

0,640; 0,652; 0,656; 0,664; 0,670.

а) Какова вероятность того, что истинное значение *a* измеряемой величины отличается от среднего результата проведённых измерений не более, чем на 0,01?

б) В каком доверительном интервале $(\bar{x}_в - \delta;\ \bar{x}_в + \delta)$ с надёжностью γ=0,99 находится искомое значение *a*?

Ответ: а) $\gamma \approx 0{,}87$; б) (0,633; 0,680)

3. Проверку качества большой партии изделий проводят выборочным путём. Каков должен быть минимальный объём выборки, чтобы с надёжностью γ=0,99 можно было утверждать, что отклонение среднего срока службы изделия в выборке отличается от среднего срока службы во всей исследуемой партии не более, чем на 3 часа (в ту или в другую сторону)? По результатам предварительной (пробной) выборки получено $s_в$ = 10 час.

Ответ: 74.

§3. Статистическая проверка гипотез

3.1. Общие положения.

Одна из основных задач математической статистики состоит в решении вопроса о том, подтверждает или не подтверждает выборка то или иное предположение (гипотезу) о генеральной совокупности. Например:

а) Подтверждает ли выборка предложение о том, что исследуемый признак *X* объектов генеральной совокупности распределен нормально?

б) Подтверждают ли выборки из двух разных генеральных совокупностей существенность (значимость) различия генеральных средних или дисперсий этих совокупностей?

И т.д. Правильный ответ на вопрос (а) важен в связи с тем, что именно при нормальности распределения исследуемого признака *X* верны многие положения и применимы многие методы математической статистики. В частности, применимы полученные в §2 формулы для случая, когда объем выборки невелик (n<30). А ответ на вопрос (б) позволяет понять, действительно ли следует считать различными обе исследуемые генеральные совокупности с нормально распределенным признаком *X* объектов этой совокупности, или все же это просто две части единой генеральной совокупности.

Задачи статистической проверки гипотез в принципе решаются следующим образом. Относительно генеральной совокупности высказывается та или иная гипотеза H_0 *(нулевая гипотеза),* которая подлежит проверке. Заодно высказывается и *конкурирующая с ней гипотеза* H_1, которая будет признана верной, если будет отвергнута нулевая гипотеза H_0. Далее, из генеральной совокупности извлекается некоторая случайная выборка. Затем применяется некий критерий, с помощью которого по выборочным данным решают, следует ли от-

клонить гипотезу H_0 в пользу конкурирующей гипотезы H_1 или все же следует гипотезу H_0 принять.

Сразу отметим, что критерия, позволяющего точно (на 100%) узнать, верна гипотеза H_0 или нет, не существует в силу ограниченности и случайности выборки (выборка не содержит всей информации о генеральной совокупности). Действительно, абсолютно точную информацию обо всем, что касается генеральной совокупности (в том числе и о справедливости или несправедливости гипотезы H_0) мы бы получили, если бы исследовали *всю* генеральную совокупность. Но мы ведь исследуем *лишь выборку из неё*, поэтому не застрахованы от ошибки в любых своих выводах (об этом, напомним, мы уже говорили). Отклоняя или принимая гипотезу H_0, можно допустить ошибку двух видов:

1) Нулевая гипотеза H_0 отвергается несмотря на то, что она верна. Это ошибка 1-го рода. Ее вероятность принято обозначать буквой α.

2) Нулевая гипотеза H_0 принимается несмотря на то, что она неверна. Эта ошибка 2-го рода. Ее вероятность принято обозначать буквой β.

А вообще все возможные ситуации с принятием и непринятием гипотезы H_0 изображены ниже в таблице (в скобках указаны вероятности соответствующих ситуаций)

Заключение о гипотезе H_0	На самом деле гипотеза H_0	
	Верна	Неверна
Отвергается	Ошибка 1-го рода (α)	Правильное решение (1-β)
Принимается	Правильное решение (1-α)	Ошибка 2-го рода (β)

Естественно потребовать от упомянутого выше критерия, чтобы возможные ошибки 1-го и 2-го родов обе имели как можно меньшие вероятности. Но добиться этого трудно, ибо очевидно, что при фиксированном объеме n выборки и выбранном критерии уменьшение α автоматически ведет к увеличению β, и наоборот.

Проиллюстрируем это утверждение на таком житейском примере. Пусть гипотеза H_0 состоит в том, что переход пешеходом перекрестка на красный свет окончится для него плохо. Тогда отвергнуть эту гипотезу – это значит переходить перекресток. Принять эту гипотезу – это значит стоять на месте (ждать зеленый сигнал).

1) Если гипотеза H_0 верна, то переходить перекресток – это отвергнуть правильную гипотезу, то есть совершить ошибку 1- го рода (смертельную).

2) Если гипотеза H_0 неверна, то оставаться на месте – это принять неправильную гипотезу, т.е. совершить ошибку 2-го рода (потерять время на ожидание зеленого сигнала светофора)

Очевидно, что тяжесть последствий каждой из этих двух ошибок совершенно разная. И еще очевидно, что уменьшая вероятность α ошибки 1-го рода, пешеход автоматически будет увеличивать вероятность β ошибки 2-го рода.

Действительно, уменьшая ошибку 1-го рода, пешеход не будет переходить на красный свет даже в тех случаях, когда машины от перекрестка далеко и переход ему практически ничем бы не грозил.

Можно дать понятиям ошибки 1-го и 2-го родов и юридическую трактовку. Пусть гипотеза H_0 состоит в том, что подсудимый невиновен (это исходная юридическая гипотеза, вытекающая из принципа презумпции невиновности). Тогда ошибка 1-го рода будет состоять в том, что подсудимому будет вынесен обвинительный приговор, несмотря на то, что он невиновен. А ошибка 2-го рода будет состоять в том, что подсудимому будет вынесен оправдательный приговор, несмотря на то, что он виновен. Эти ошибки тоже разные по тяжести. Стараясь уменьшить одну из них (например, ошибку первого рода), судья чаще должен выносить оправдательный приговоры, что автоматически ведет к увеличению ошибки 2-го рода.

При практической проверке гипотез, выдвигая нулевую гипотезу H_0 и альтернативную ей гипотезу H_1, сразу задают допустимую вероятность α ошибки 1-го рода – вероятность отвергнуть нулевую гипотезу H_0, если она верна. Конкретное числовое значение для этой ошибки выбирают, исходя из реальной тяжести последствий такой ошибки. Кстати, принятое значение α называют еще *уровнем значимости*. Таким образом, *уровень значимости α – это вероятность отвергнуть нулевую гипотезу H_0 при условии, что она верна*. При решении инженерных, сельскохозяйственных, экономических задач по статистической проверке гипотез обычно в качестве уровня значимости α принимают значения α = 0,1; или α = 0,05; или α = 0,01. Часто, особенно в литературе прикладного характера, уровень значимости выражается еще в процентах. При этом только что названные уровни значимости называют 10% - ым, 5% - ым, 1% - ым уровнями значимости. Чем меньше мы выберем значение α, тем меньше будет риск отвергнуть нулевую гипотезу H_0, если она правильная. Но зато тем больше будет риск принять ее, если она неправильная (то есть если правильной будет альтернативная гипотеза H_1). Брать α=0 смысла нет, так как при α=0 риск отклонить проверяемую гипотезу H_0, если она верна, должен быть сведен к нулю. А это значит, что при α=0 гипотезу H_0, как бы мало вероятной она ни была, заведомо нужно принимать. То есть при α=0 теряется сам смысл статистической проверки гипотез.

Приняв гипотезу H_0, тем самым соглашаются с тем, что те результаты исследования выборки, которые этой гипотезе противоречат, являются несущественными (незначимыми) и являются лишь следствием случайности самой выборки. А отвергнув гипотезу H_0, тем самым признают, что противоречащие гипотезе H_0 результаты исследования выборки трудно объяснить лишь случайностью самой выборки, они слишком существенны (значимы) для того, чтобы можно было согласиться со справедливостью гипотезы H_0. Наше решение относительно значимости или незначимости выборочных данных, противоречащих гипотезе H_0, зависит от принятого уровня значимости α. Отсюда и его название – *уровень значимости*.

Отметим еще одно важное и очевидное обстоятельство: принятие или непринятие гипотезы H_0 зависит не только от величины уровня значимости, *но и*

от того, какова гипотеза H_1 – гипотеза, альтернативная гипотезе H_0. Действительно, гипотеза H_1 противостоит гипотезе H_0 – они конкурируют друг с другом за принятие. И устоит ли гипотеза H_0 против гипотезы H_1 – это, естественно, зависит и от того, что эта гипотеза H_1 собой представляет.

В самом деле, если пешеход выбирает, переходить ли ему на красный свет или дождаться через пару минут зеленого – это одна ситуация. А если он выбирает между переходом на красный свет и необходимостью перехода в другом месте (метров за 300 от данного места) – это ситуация совсем другая. Во втором случае пешеходы будет рисковать, переходя улицу, чаще. И тем самым будут увеличивать вероятность собственной ошибки первого рода (ведущей под колеса автомобиля).

Заметим еще, что *отклонение нулевой гипотезы H_0 – акт гораздо более обоснованный, чем принятие этой гипотезы.* Действительно, отклоняя гипотезу H_0, мы знаем вероятность того, что ошибаемся - ведь это вероятность ошибки первого рода. А она равна принятому уровню значимости α, то есть она известна и мала. А принимая гипотезу H_0, мы не знаем вероятности возможной собственной ошибки (ошибки второго рода) - эта вероятность β не контролируется, и она может быть не мала, особенно если слишком мал задаваемый уровень значимости α. Отклонив гипотезу H_0, мы при малых значениях α практически можем быть уверены, что эта гипотеза действительно не верна. А приняв ее, мы отнюдь не можем утверждать, что подтвердили ее правильность. Мы можем лишь позволить себе осторожно сказать, что экспериментальные выборочные данные не противоречат этой гипотезе. Но ведь точно так же при малых α они могут не противоречить и многим другим, отличным от H_0, гипотезам. А при $\alpha = 0$ любые противоречия гипотезе H_0 будут признаны несущественными (незначимыми) и гипотеза H_0, даже самая невероятная, будет принята!

Для борьбы с принятием неверных гипотез H_0 нужно или увеличивать α, уменьшая тем самым ошибку β принятия неверной нулевой гипотезы, или, что гораздо продуктивнее, не увеличивая α, увеличивать объем n выборки, приближая тем самым выборку ко всей генеральной совокупности, а сомнительные результаты – к точным.

В качестве критерия, в соответствии с которым решают, принять или не принять выдвинутую нулевую гипотезу H_0, используют специально подобранную случайную величину K, распределение которой, при условии справедливости гипотезы H_0, известно (величина K может иметь нормальное распределение, или распределение Стьюдента, или распределение Фишера-Снедекора, и т.д.). С учетом выдвинутой гипотезы H_0, альтернативной гипотезы H_1 и принятого уровня значимости α множество всех возможных этой величины разбивают на два подмножества:

a) Множество тех значений k величины K, при которых нулевая гипотеза H_0 принимается (это множество называется *областью принятия гипотезы)*;

b) Множество тех значений k величины K, при которых нулевая гипотеза отвергается (а следовательно, принимается альтернативная гипотеза H_1). Это подмножество называется *критической областью.*

Точки $k_{кр}$ (их может быть одна или несколько), разделяющие область принятия гипотезы H_0 и критическую область, называются *критическими*.

Считая сделанную выборку экспериментом, на основании данных этой выборки вычисляют экспериментальное значение $k_{эксп}$ величины K. И если оно попадет в область принятия гипотезы H_0, гипотезу H_0 принимают. А если в критическую область – то отвергают.

Заметим, что $k_{эксп}$ в силу случайности выборки может попасть в критическую область и, следовательно, гипотеза H_0 будет отвергнута несмотря на то, что на самом деле она верна. Тогда будет допущена ошибка 1-го рода – отвергнута правильная гипотеза. Вероятность такой ошибки, то есть *вероятность попадания* $k_{эксп}$ *в критическую область, должна быть равна принятому уровню значимости* α. Из этого условия, а также из соображений, касающихся возможной симметрии критической области, и находятся критические точки $k_{кр}$. При этом используется специальная таблица критических точек того распределения, которое имеет выбранный критерий K.

Наоборот, $k_{эксп}$ может попасть в область принятия гипотезы H_0, когда она неверна. Тогда будет допущена ошибка 2-го рода – принята неправильная гипотеза. Вероятность β такой ошибки обычно неизвестна (она, в отличие от ошибки α, не задается и не контролируется). Не обращая внимание на эту ошибку второго рода, при попадании $k_{эксп}$ в область принятия гипотезы H_0 эту гипотезу принимают.

А теперь перейдем к некоторым конкретным примерам статистической проверки гипотез.

3.2. Проверка гипотезы о значении математического ожидания нормально распределенной случайной величины.

Пусть нулевая гипотеза H_0 состоит в том, что математическое ожидание a нормальной распределенной случайной величины X равно некоторому заданному значению a_0. А альтернативной гипотезой H_1 является гипотеза о том, что $a \neq a_0$. Будем при этом пока предполагать, что параметр σ (среднее квадратическое отклонение величины X) известен. Такая задача будет, например, актуальной для станка-автомата, изготавливающего в массовом количестве некоторые детали, размер a_0 которых задан. Если станок настроен правильно, то размер X изготавливаемых им деталей будет случайной величиной, распределенной нормально с известным математическим ожиданием $a = a_0$ и известным средним квадратическим отклонением σ, определяемым классом точности станка. Но если станок настроен неправильно, у него будет $a \neq a_0$. Отобрав из продукции станка некоторую совокупность деталей и обмерив их, можно попытаться узнать, верна ли гипотеза H_0 о том, что $a=a_0$, или она должна быть отклонена в пользу альтернативной гипотезы H_1, утверждающей, что $a \neq a_0$.

Итак, гипотеза H_0 сформулирована. Попробуем найти критерий ее правильности или неправильности.

Пусть выборочная средняя $\bar{x}_в$, являющаяся точечной оценкой генеральной средней $\bar{x}_Г = a$, из выборки найдена. Предположим, что гипотеза H_0 верна, то есть что $\bar{x}_Г = a = a_0$. В качестве случайной величины K, с помощью которой будем проверять гипотезу H_0, возьмем нормально распределенную случайную величину $K = \bar{x}_в - a_0$, математическое ожидание которой, если верна гипотеза H_0, равно нулю. Выберем некоторый уровень значимости α. Если гипотеза H_0 верна, то отклонение $K = \bar{x}_в - a_0$ от нуля должно быть невелико, то есть практически должно находится в некоторых границах $(-k_{кр}; k_{кр})$. А выход его за эти границы следует считать опровержением гипотезы H_0. То есть оставшиеся вне интервала $(-k_{кр}; k_{кр})$ участки числовой оси Ok мы будем считать критической областью – областью непринятия гипотезы H_0. При условии справедливости H_0 ее непринятие – это ошибка первого рода. Поэтому вероятность попадания значения k величины $K = \bar{x}_в - a_0$ в критическую область должна быть равна вероятности совершить ошибку первого рода, то есть должна равняться принятому уровню значимости α. А вероятность попадания значения k в область $(-k_{кр}; k_{кр})$ принятия гипотезы H_0 тогда должна быть равна $\gamma = 1-\alpha$ (см. рис. 3.4).

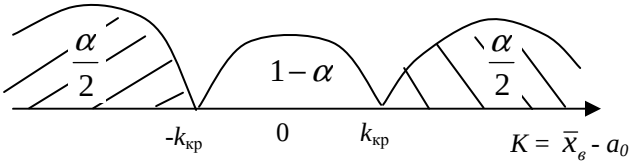


Рис. 3.4

Очевидно, что чем меньше будет выбран уровень значимости α, то есть чем меньшей будет установлена вероятность совершить ошибку первого рода – отвергнуть нулевую гипотезу H_0, если она верна, тем шире должен быть интервал $(-k_{кр}; k_{кр})$ принятия этой гипотезы. При $\alpha = 0$ должна быть вообще исключена возможность отвергнуть проверяемую гипотезу H_0, если она верна. Но тогда в любом случае ее нужно принять. А это будет, согласно рис. 3.4, если заштрихованной области непринятия гипотезы H_0 не будет вообще. То есть когда $(-k_{кр}; k_{кр}) = (-\infty; \infty)$.

Итак, подведем итог. При заданном $\alpha \neq 0$ гипотеза H_0 должна быть отвергнута, если экспериментальное значение $k=k_{эксп}$ нормально распределенной случайной величины $K = \bar{x}_в - a_0$ окажется вне интервала $(-k_{кр}; k_{кр})$. И она же может быть принята, если $k=k_{эксп}$ попадет в интервал $(-k_{кр}; k_{кр})$.

Согласно рис. 3.4 имеем:

$$p(-k_{кр} \prec \bar{x}_в - a_0 \prec k_{кр}) = 1-\alpha \qquad (3.2)$$

Сравнивая это равенство с равенством

$$p(-\delta \prec \bar{x}_в - a_0 \prec \delta) = 2\Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right) = \gamma, \qquad (3.2)$$

вытекающем из равенства (2.22), если в последнем заменить $\bar{x}_Г$ на a_0 и $\sigma_Г$ на σ, получим

$$2\Phi\left(\frac{k_{кр}\sqrt{n}}{\sigma}\right) = 1-\alpha, \qquad (3.3)$$

откуда

$$k_{кр} = t_{кр}\frac{\sigma}{\sqrt{n}}, \text{ где } 2\Phi(t_{кр}) = 1-\alpha \qquad (3.4)$$

После нахождения $k_{кр}$ находятся области принятия и непринятия гипотезы H_0:

1) Если $k_{эксп} = \bar{x}_в - a_0 \in (-k_{кр}; k_{кр})$ - гипотезу H_0 принимаем.

2) Если $k_{эксп} = \bar{x}_в - a_0 \notin (-k_{кр}; k_{кр})$ - гипотезу H_0 отвергаем.

Примечание 1. Если σ неизвестно, то при достаточно большой выборке (n>30) можно заменить в приведенных выше формулах σ на её точечную оценку $s_в$.

Примечание 2. Если σ неизвестно и выборка небольшая (n<30), то тогда в качестве критерия K проверки гипотезы H_0 используют случайную величину

$$T = \frac{x_в - a_0}{\frac{s_в}{\sqrt{n}}}, \qquad (3.5)$$

имеющую распределение Стьюдента с $k = n$-1 степенями свободы (сравните с (2.27)). А областью принятия гипотезы H_0 считается интервал $(-t_{кр}; t_{кр})$, в который с вероятностью $\gamma = 1$-α попадает экспериментальное значение $t_{эксп}$ случайной величины T:

$$p(-t_{кр} \prec T \prec t_{кр}) = 1-\alpha \qquad (3.6)$$

Величину $t_{кр}$, соответствующую данным значениям α и $k = n - 1$, находят по таблице критических точек распределения Стьюдента (см. Приложение, таблица 4). Кстати, по этой же таблице по $\gamma = 1-\alpha$ и $k = n - 1$ находят t_γ, входящее в равенство (2.28).

Примечание 3. В качестве гипотезы, альтернативной гипотезе H_0, могла быть принята не гипотеза H_1: $a \neq a_0$, а гипотеза H_1: $a > a_0$ или гипотеза H_1: $a < a_0$. Тогда критическая область непринятия гипотезы H_0 была бы не двусторонней, как на рис. 3.4, а односторонней с одной критической точкой $k_{кр}$, определяемой соответственно неравенствами $p(k_{кр} < k_{эксп} < \infty) = \alpha$ или $p(-\infty < k_{эксп} < k_{кр}) = \alpha$. Соответствующие изменения при проверке гипотезы H_0 (а именно, при нахождении табличного значения $k_{кр}$) продумайте самостоятельно (или посмотрите ниже примечание 2 к пункту 3.5).

Пример 1. Выборочным путем исследовались 30 молочных пакетов, в которых, согласно их маркировке, должно содержаться 0,5 л. = 500 мл. молока. Получено следующее статистическое распределение выборки:

x_i (мл.)	480 – 490	490 – 500	500 – 510	510 – 520
n_i	8	10	8	4

Проверить при уровне значимости α = 0,05 нулевую гипотезу H_0 о том, что автомат, заполняющий пакеты, настроен правильно (на норму 500 мл.), при альтернативной гипотезе H_1, что автомат настроен неправильно.

Решение. Пусть X (мл.) – реальное количество молока в пакете (X – случайная величина). Различие её значений для разных пакетов связано с действием большого числа случайных факторов, поэтому по теореме Ляпунова (глава

2, §4) можно считать, что она распределена нормально с некоторыми параметрами $a = M(X)$ и $\sigma = \sigma(X)$. Если гипотеза H_0 верна (автомат настроен правильно), то a = 500 мл. А если гипотеза H_0 неверна, то верна альтернативная гипотеза H_1: $a \neq 500$ мл.

Найдем числовые характеристики выборки – выборочную среднюю $\bar{x}_в$ и выборочную дисперсию $D_в$. Для этого сначала приведем статистическое распределение выборки от интервального вида к дискретному:

x_i (мл.)	485	495	505	515
n_i	8	10	8	4

Используя теперь формулы (1.5) – (1.7), получим:

$$\bar{x}_в \approx 497{,}7;\ D_в \approx 100{,}2$$

Далее, с помощью формулы (2.15) найдем исправленную выборочную дисперсию $s_в^2$ и исправленное выборочное среднее квадратическое отклонение $s_в$:

$$s_в^{\,2} = \frac{30}{29} \cdot 100{,}2 \approx 103{,}7; \quad s_в = \sqrt{103{,}7} \approx 10{,}2$$

А теперь рассмотрим нормально распределенную случайную величину $K = \bar{x}_в - a_0$, экспериментальное (выборочное) значение которой при $\bar{x}_в = 497{,}7$ и $a_0 = 500$ равно $k = k_{эксп} = -2{,}3$.

Если считать объем выборки $n = 30$ большим (хотя мы условились считать его большим лишь при $n > 30$), то область $(-k_{кр}; k_{кр})$ принятии гипотезы H_0 можно найти, найдя $k_{кр}$ по формуле (3.4), в которой нужно заменить σ на $s_в$. С учетом имеющихся данных находим:

$$2\Phi(t_{кр}) = 0{,}95 \Rightarrow t_{кр} = 1{,}96;\ k_{кр} = 1{,}96 \cdot \frac{10{,}2}{\sqrt{30}} \approx 3{,}7$$

Итак, интервал принятия гипотезы H_0 для значения случайной величины $K = \bar{x}_в - a_0$ таков: $-3{,}7 < K < 3.7$. Так как $k_{эксп} = -2{,}3$ принадлежит этому интервалу, то у нас нет оснований при заданном уровне значимости $\alpha = 0{,}05$ отвергать нулевую гипотезу H_0 о том, что автомат по разливу молока в пакеты настроен правильно.

Кстати, если величину α уменьшить, то интервал принятия гипотезы H_0 расширится, а значит, будет еще больше оснований эту гипотезу не отвергать.

А теперь, для сравнения, проведем проверку гипотезы H_0, считая объем $n = 30$ выборки небольшим, то есть используя формулы (3.5) и (3.6). Сначала найдем экспериментальное значение $t_{эксп.}$ величины T:

$$t_{эксп.} = \frac{-2{,}3}{\frac{10{,}2}{\sqrt{30}}} \approx -1{,}24$$

После этого по таблице 4 Приложения для $\alpha = 0{,}05$ и $k = n - 1 = 29$ найдем критическое значение $t_{кр}$ распределения Стьюдента: $t_{кр} = 2{,}045$. Таким образом, областью принятия гипотезы H_0 является интервал

$$(-t_{кр};\ t_{кр}) = (-2{,}045;\ 2{,}045).$$

И так как $t_{эксп.}$ = -1,24 содержится в этом интервале, то у нас нет оснований отвергать гипотезу H_0 о том, что автомат по разливу молока в пакеты настроен правильно.

3.3 Проверка гипотезы о равенстве дисперсий двух нормально распределенных случайных величин.

Гипотезы о дисперсиях имеют особенно большое значение в технике, так как дисперсия характеризует рассеяние значений случайной величины вокруг среднего значения этой величины. А значит, в технике она характеризует точность измерительных приборов, устойчивость работы машин, отлаженность технологических процессов, и т. д. Очевидно, предпочтительнее тот прибор, инструмент или процесс, который обеспечивает наименьшее рассеяние результатов измерений или действий. То есть обеспечивает наименьшую дисперсию.

Пусть X и Y – две нормально распределенные случайные величины, имеющие одно и то же наименование (размерность). Например, пусть X - результат измерения некоторой величины a_X одним прибором, а Y – результат измерения этой же (или какой-то другой) величины a_Y той же размерности, что и первая, другим прибором. Параметры $(a_X;\ \sigma_X)$ и $(a_Y;\ \sigma_Y)$ этих нормально распределенных случайных величин будем считать неизвестными. То есть нам не известны ни истинные значения a_X и a_Y измеряемых приборами величин, ни точности σ_X и σ_Y приборов. Требуется при заданном уровне значимости α проверить нулевую гипотезу H_0 о том, что дисперсии величин X и Y, а значит и точность обоих приборов, одинаковы: $\sigma_X^2 = \sigma_Y^2$. В качестве альтернативной гипотезы H_1 примем гипотезу, состоящую в том, что прибор, который на практике оказался менее точным (показал больший разброс результатов измерений) и на самом деле является менее точным.

Для статистической проверки нулевой гипотезы H_0 произведем n_X и n_Y повторных измерений величин X и Y, и по полученным выборочным данным найдем исправленные выборочные дисперсии s_X^2 и s_Y^2. Эти исправленные выборочные дисперсии являются, как мы знаем, точечными несмещенными оценками истинных дисперсий σ_X^2 и σ_Y^2 величин X и Y:

$$\sigma_X^2 \approx s_X^2;\ \ \sigma_Y^2 \approx s_Y^2 \qquad (3.7)$$

Таким образом, сравнение σ_X^2 и σ_Y^2 сводится к сравнению s_X^2 и s_Y^2. Если окажется, что $s_X^2 = s_Y^2$ (что маловероятно из-за случайности выборок), то гипотеза H_0, естественно, считается принятой. Но обычно $s_X^2 \neq s_Y^2$. И поэтому возникает вопрос: существенно ли (значимо) или несущественно (незначимо) отличается бóльшая из них от меньшей? То есть принимать ли гипотезу H_0 о равенстве дисперсий σ_X^2 и σ_Y^2, а значит, и о равной точности обоих приборов? Или все же склониться в сторону альтернативной гипотезы H_1 о том, что один из приборов точнее другого?

Для решения этого вопроса введем случайную величину F, представляющую собой отношение бóльшей исправленной дисперсии к меньшей:

$$F = \frac{s_б^{\,2}}{s_м^{\,2}} \tag{3.8}$$

(F – случайная величина, так как $s_б^2$ и $s_м^2$ – сами случайные величины). Доказано, что при условии справедливости гипотезы H_0 о равенстве σ_X^2 и σ_Y^2 введенная случайная величина F имеет распределение Фишера-Снедекора (см. §4, глава 2) со степенями свободы $k_1 = n_1$-1 и $k_2 = n_2$ -1, где n_1 – объем выборки с большей исправленной дисперсий, а n_2 – с меньшей. Напомним, что распределение Фишера-Снедекора зависит только от степеней свободы k_1 и k_2 и не зависит от других параметров.

Ясно, что при справедливости гипотезы H_0, утверждающей, что $\sigma_X^2 = \sigma_Y^2$, должны быть примерно равны $s_б^2$ и $s_м^2$, так что экспериментальное значение

$$f_{эксп.} = \frac{s_б^{\,2}}{s_м^{\,2}} \tag{3.9}$$

величины F не должно существенно превосходить единицу. То есть не должно превосходить некоторое значение $f_{кр}$, где $f_{кр} > 1$ (рис 3.5). Если оно его все же превзойдет, то это будет свидетельствовать о неправомерности принятия нулевой гипотезы H_0. То есть если окажется, что $f_{эксп.} > f_{кр.}$, то мы гипотезу H_0 отвергнем.

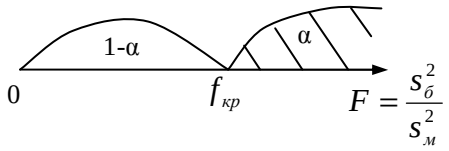


Рис. 3.5

Но если эта гипотеза все же верна, то отвергая её, мы совершим ошибку первого рода. Вероятность этой ошибки должна быть равна задаваемому уровню значимости α. А значит, для значения $f_{кр}$ получаем следующее равенство:

$$p(F \succ f_{кр}) = \alpha \tag{3.10}$$

Или что одно и то же:

$$p(0 \le F \prec f_{кр}) = 1 - \alpha \tag{3.11}$$

Величину $f_{кр} = f_{кр}(\alpha\ ;\ k_1\ ;\ k_2)$ для заданных ($\alpha\ ;\ k_1\ ;\ k_2$) можно найти но специальной таблице критических точек распределения Фишера-Снедекора (см. таблицу 5 Приложения). И если окажется, что экспериментальное значение (3.9) величины F окажется больше $f_{кр}$ (попадет в критическую область), то гипотезу H_0 о равенстве двух дисперсий (о равной точности двух приборов) отвергают, принимая тем самым альтернативную гипотезу H_1, о том, что один из приборов (с большей исправленной дисперсией) менее точен. А если окажется, что $f_{эксп} < f_{кр}$ (то есть если $f_{эксп}$ попадет в область принятия гипотезы H_0), то эту гипотезу H_0 о равной точности обоих приборов принимают. Принимают, считая тем самым несущественным (незначимым) то различие исправленных дисперсий $s_б^2$ и $s_м^2$, которое имеет место в проведенных двух сериях измерений этими приборами.

Пример 2. На двух токарных станках обрабатывается втулки. Отобраны две пробы: $n_1 = 10$ штук - из втулок, обработанных на первом станке, и $n_2 = 15$ штук - из втулок, обработанных на втором станке. По данным этих выборок найдены исправленные выборочные дисперсии: $s_1^2 = 9{,}6$ и $s_2^2 = 5{,}7$. Проверить при уровне значимости: а) $\alpha = 0.05$ и б) $\alpha = 0{,}01$ гипотезу H_0 об одинаковой

точности этих двух станков при альтернативной гипотезе H_1 о том, что второй станок точнее, чем первый.

Решение. Здесь

$$f_{эксп} = \frac{s_б^{\,2}}{s_м^{\,2}} = \frac{s_1^{\,2}}{s_2^{\,2}} = \frac{9{,}6}{5{,}7} = 1{,}68$$

По таблице критических точек распределения Фишера – Снедекора (таблица 5 Приложения) для $k_1 = 10 - 1 = 9$; $k_2 = 15 - 1 = 14$ и а) $\alpha = 0{,}05$; б) $\alpha = 0.01$ находим $f_{кр}$:

а) $f_{кр} = f_{кр}\,(0{,}05;\ 9;\ 14) = 2{,}65$;
б) $f_{кр} = f_{кр}\,(0{,}01;\ 9;\ 14) = 4{,}03$.

Так как и в варианте (а), и варианте (б) $f_{эксп} < f_{кр}$, то в обоих вариантах у нас нет оснований отвергать гипотезу H_0 о равенстве дисперсий размеров втулок, обрабатываемых на обоих станках. То есть гипотеза H_0 об одинаковой точности этих двух станков принимается. Тем самым имеющееся различие между s_1^2 и s_2^2 признается несущественным (незначимым) и относится на счет случайности проведенных выборок.

Примечание. Полученный вывод о незначимости различия выборочных дисперсий $s_1^2 = 9{,}6$ и $s_2^2 = 5{,}7$ получился, скорее всего, лишь из – за малости обеих выборок ($n_1 = 10$ и $n_2 = 15$). Если бы такие же значения s_1^2 и s_2^2 получились для больших выборок, то такое существенное различие между s_1^2 и s_2^2 было бы трудно объяснить случайностями самих выборок, и гипотеза H_0 о равной точности станков была бы отвергнута. Действительно, согласно таблице 5 Приложения при $k_1 = n_1 - 1 \to \infty$ и $k_2 = n_2 - 1 \to \infty$ значение

$$f_{кр} = f_{кр}\,(\alpha; k_1; k_2) \to 1$$

при всех уровнях значимости α. Поэтому при больших значениях k_1 и k_2 было бы $f_{эксп} > f_{кр}$, и гипотеза H_0 была бы отвергнута и в варианте (а), и в варианте (б).

3.4. Проверка гипотезы о равенстве математических ожиданий двух нормально распределенных случайных величин (большие независимые выборки)

Пусть X и Y – две независимые нормально распределенные случайные величины. Требуется с помощью выборочных данных, полученных при исследовании значений этих величин, проверить нулевую гипотезу H_0 о равенстве этих математических ожиданий при альтернативной гипотезе H_1 об их неравенстве. Такая задача, например, будет актуальной при выяснении вопроса о том, изготовляют ли два параллельно работающих станка одинаковые в среднем детали или все же разные. Или имеют два сорта сельскохозяйственной культуры одинаковую в среднем урожайность или все же разную.

Пусть n_X и n_Y – объемы выборок, а $(\bar{x}_в; s_X^2)$ и $(\bar{y}_в; s_Y^2)$ – выборочная средняя и исправленная выборочная дисперсия указанных выборок. Образуем случайную величину $Z = \bar{x}_в - \bar{y}_в$. При условии справедливости гипотезы H_0 случайная величина Z распределена нормально с

$$M(Z) = M(\bar{x}_в) - M(\bar{y}_в) = 0;\ \sigma(Z) = \sqrt{D(Z)} = \sqrt{D(\bar{x}_в) + D(\bar{y}_в)} = \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}, \qquad (3.12)$$

где σ_X^2 и σ_Y^2 – истинные (генеральные) дисперсии величин *X* и *Y*. При заданном уровне значимости α областью принятии гипотезы H_0 будет, очевидна, интервал

$$-z_{кр} \prec Z \prec z_{кр}, \qquad (3.13)$$

где $z_{кр}$ находится из условия:

$$p(-z_{кр} \prec Z \prec z_{кр}) = 1 - \alpha \qquad (3.14)$$

А это, на основании формулы (4.11) (часть I, глава 2), дает:

$$z_{кр} = t_{кр} \cdot \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}},\ \text{где}\ \Phi(t_{кр}) = \frac{1-\alpha}{2} \qquad (3.15)$$

Генеральные дисперсии σ_X^2 и σ_Y^2 обычно неизвестны. Но если объемы выборок n_X и n_Y достаточно велики, то есть они не менее 30 (что мы и предполагаем для больших выборок), то в (3.15) можно заменить σ_X^2 и σ_Y^2 их точечными оценками s_X^2 и s_Y^2.

Пример 3. По двум независимым выборкам объемами n_X=100 и n_Y=120 найдены выборочные средние $\bar{x}_в$= 32,4 и $\bar{y}_в$ = 30,1, а также исправленные выборочные дисперсии s_X^2=15,0 и s_Y^2=25,2. Исследуемые с помощью этих выборок случайные величины *X* и *Y* распределены нормально. При уровне значимости α = 0,05 проверить нулевую гипотезу H_0 о том, что *M(X) = M(Y)* при альтернативной гипотезе H_1 о том, что *M(X) ≠ M(Y)*.

Решение. Находим сначала опытное (экспериментальное) значение $z_{эксп}$ нормально распределенной величины *Z* = $\bar{x}_в$ - $\bar{y}_в$:

$$z_{эксп} = \bar{x}_в - \bar{y}_в = 32{,}4 - 30{,}1 = 2{,}3$$

Далее из равенства $\Phi(t_{кр}) = \frac{1-\alpha}{2} = 0{,}475$ с помощью таблицы интеграла вероятности *Ф(x)* находим $t_{кр:}$ $t_{кр}$=1,96. После этого по формуле (3.15) вычисляем $z_{кр}$: $z_{кр}$=1,176. Таким образом, интервалом принятия гипотезы H_0 для значений величины *Z* будет следующий интервал:

$$(-z_{кр};\ z_{кр}) = (-1{,}176;\ 1{,}176)$$

Так как экспериментальное значение $z_{эксп}$ = 2,3 величины *Z* не входит в этот интервал, то гипотезу H_0 о равенстве математических ожиданий величин *X* и *Y* отвергаем – она не подтверждается экспериментальными данными.

3.5 Проверка гипотезы о равенстве математических ожиданий двух нормально распределенных случайных величин (малые независимые выборки).

Рассмотрим ту же задачу, что и в предыдущем пункте 3.4, но только при условии, что объемы выборок n_X и n_Y невелики (меньше 30). В этом случае замена генеральных дисперсий σ_X^2 и σ_Y^2, входящих в (3.15), на исправленные вы-

борочные дисперсии s_X^2 и s_Y^2 может привести к большой ошибке в величине $z_{кр}$, а следовательно, к большой ошибке в установлении области $(-z_{кр}; z_{кр})$ принятия гипотезы H_0. Однако если есть уверенность в том, что неизвестные генеральные σ_X^2 и σ_Y^2 *одинаковы* (например, если сравниваются средние размеры двух партий деталей, изготовленных на одном и том же станке), то можно, используя распределение Стьюдента, и в этом случае построить критерий проверки гипотезы H_0 о равенстве математических ожиданий величин *X* и *Y*. Для этого вводят случайную величину

$$T = \frac{\bar{x}_в - \bar{y}_в}{\sqrt{s^2}} \cdot \sqrt{\frac{n_X \cdot n_Y}{n_X + n_Y}}, \qquad (3.16)$$

где

$$s^2 = \frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2} \qquad (3.17)$$

- среднее из исправленных выборочных дисперсий s_X^2 и s_Y^2, служащее точечной оценкой обеих одинаковых неизвестных генеральных дисперсий σ_X^2 и σ_Y^2. Как оказывается (см. [3], стр.180), при справедливости нулевой гипотезы H_0 случайная величина *T* имеет распределение Стьюдента с $k = n_X + n_Y - 2$ степенями свободы независимо от величин n_X и n_Y объемов выборок. Если гипотеза H_0 верна, то разница $\bar{x}_в - \bar{y}_в$ должна быть невелика. То есть экспериментальное значение $t_{эксп.}$ величины *T* должно быть невелико. А именно, должно заключаться в некоторых границах $(-t_{кр}; t_{кр})$. Выход же его за эти границы мы будем считать опровержением гипотезы H_0, и допускать это будем с вероятностью, равной задаваемому уровню значимости α.

Таким образом, областью принятия гипотезы H_0 будет являться некоторый интервал $(-t_{кр}; t_{кр})$, в который значения случайной величины *T* должны попадать с вероятностью 1- α:

$$p(-t_{кр} \prec T \prec t_{кр}) = 1 - \alpha \qquad (3.18)$$

Величину $t_{кр}$, определяемую равенством (3.18), для различных уровней значимости α и различных числах *k* степеней свободы величины *T* можно найти в таблице критических точек распределения Стьюдента (таблице 4 Приложения). Тем самым будет найден интервал $(-t_{кр}; t_{кр})$ принятия гипотезы H_0. И если экспериментальное значение $t_{эксп}$ величины *T* попадет в этот интервал – гипотезу H_0 принимают. Не попадает - не принимают.

Примечание 1. Если нет оснований считать равными генеральные дисперсии σ_X^2 и σ_Y^2 величин *X* и *Y*, то и в этом случае для проверки гипотезы H_0 о равенстве математических ожиданий величин *X* и *Y* допускается использование изложенного выше критерия Стьюдента. Только теперь у величины *T* число *k* степеней свободы следует считать равным не $n_X + n_Y - 2$, а равным (см. [1])

$$k = (n_X + n_Y - 2) \cdot \left(0{,}5 + \frac{s_X^2 \cdot s_Y^2}{s_X^4 + s_Y^4}\right) \qquad (3.19)$$

Если исправленные выборочные дисперсии s_X^2 и s_Y^2 различаются существенно, то второе слагаемое в последней скобке (3.19) невелико по сравнению с 0,5, так что выражение (3.19) по сравнению с выражением $k = n_X + n_Y - 2$ уменьшает число степеней свободы случайной величины T почти вдвое. А это ведет к существенному расширению интервала $(-t_{кр}; t_{кр})$ принятия гипотезы H_0 и, соответственно, к существенному сужению критической области непринятия этой гипотезы. И это вполне справедливо, так как степень разброса возможных значений разности $\bar{x}_в - \bar{y}_в$ будет, в основном, определяться разбросом значений той из величин X и Y, которая имеет большую дисперсию. То есть информация от выборки с меньшей дисперсией как бы пропадает, что и ведет к большей неопределенности в выводах о гипотезе H_0 .

Пример 4. По приведенным в таблице данным сравнить средние удои коров, получавших различные рационы. При проверке нулевой гипотезы H_0 о равенстве средних удоев принять уровень значимости α=0,05.

Рацион	Поголовье коров, получавших рацион (*голов*)	Среднесуточный удой в пересчете на базисную жирность (*кг/на голову*)	Среднеквадратическое отклонение суточной молочной продуктивности коров (*кг/на голову*)
№ 1	10 (n_X)	16,2 ($\bar{x}_в$)	3,8 (s_X)
№ 2	8 (n_Y)	17,7 ($\bar{y}_в$)	4,2 (s_Y)

Решение. Так как приведенные табличные данные получены на основании малых выборок объемами n_X=10 и n_Y=8, то для сравнения математических ожиданий среднесуточных удоев коров, получавших тот и другой кормовые рационы, мы должны использовать теорию, изложенную в этом пункте. Для этого в первую очередь выясним, позволяют ли найденные исправленные выборочные дисперсии s_X^2=$(3,8)^2$=14,44 и s_Y^2=$(4,2)^2$=17,64 считать равными генеральные дисперсии σ_X^2 и σ_Y^2. Для этого используем критерий Фишера-Снедекора (см. пункт 3.3). Имеем:

$$f_{эксп} = \frac{s_Y^2}{s_X^2} = \frac{17,64}{14,44} = 1,22$$

По таблице критических точек распределения Фишера-Снедекора для α=0,05; k_1=8-1=7 и k_2=10-1=9 находим $f_{кр}$:

$$f_{кр} = f_{кр}(0,05;7;9) = 3,29$$

И так как $f_{эксп} \prec f_{кр}$, то у нас нет оснований при данном уровне значимости α=0,05 отвергать гипотезу H_0 о равенстве генеральных дисперсий σ_X^2 и σ_Y^2.

Теперь, в соответствии с (3.17) и (3.16), подсчитаем экспериментальное значение величины T:

$$s^2 = \frac{(10-1)\cdot 14,44 + (8-1)\cdot 17,64}{10+8-2} = 15,84$$

$$t_{эксп} = \frac{16{,}2 - 17{,}7}{\sqrt{15{,}84}} \cdot \sqrt{\frac{10 \cdot 8}{10+8}} = -0{,}79$$

Далее, по формуле $k = n_X + n_Y - 2$ находим число k степеней свободы величины T: k=10+8-2=16. После этого для α=0,05 и k=16 по таблице критических точек распределения Стьюдента (таблица 4 Приложения) находим $t_{кр}$: $t_{кр}$=2,12. Таким образом, интервалом принятия гипотезы H_0 о равенстве средних удоев коров, получавших рационы № 1 и № 2, является интервал $(-t_{кр}; t_{кр})$ =(-2,12; 2,12). И так как $t_{эксп}$ = - 0,79 попадает в этот интервал, то у нас нет оснований отвергать гипотезу H_0. То есть мы вправе считать, что различие кормовых рационов не сказывается на среднесуточном удое коров.

Примечание 2. В рассмотренных выше пунктах 3.4 и 3.5 рассматривалась нулевая гипотеза H_0 о равенстве $M(X)=M(Y)$ при альтернативной гипотезе H_1 об их неравенстве: $M(X)\neq M(Y)$. Но альтернативная гипотеза H_1 может быть и другой, например, $M(Y)>M(X)$. На практике этот случай будет иметь место, когда вводится некоторое усовершенствование (положительный фактор), который позволяет рассчитывать на увеличение в среднем значений нормально распределенной случайной величины Y по сравнению со значениями нормально распределенной величины X. Например, в рацион коров введена новая кормовая добавка, позволяющая рассчитывать на увеличение среднего удоя коров; под культуру внесена дополнительная подкормка, позволяющая рассчитывать на увеличение средней урожайности культуры, и т.д. И хотелось бы выяснить, существенен (значим) или незначим этот введенный фактор. Тогда в случае больших объемов n_X и n_Y выборок (см. пункт 3.4) в качестве критерия справедливости гипотезы H_0 рассматривают нормально распределенную случайную величину

$$Z = \bar{y}_в - \bar{x}_в \tag{3.20}$$

При заданном уровне значимости α гипотеза H_0 о равенстве $M(X)$ и $M(Y)$ будет отвергнута, если экспериментальное значение $z_{эксп}$ величины Z будет положительным и бóльшим $z_{кр}$, где

$$p(Z \succ z_{кр}) = \alpha \tag{3.21}$$

Так как при справедливости гипотезы H_0 $M(Z)$=0, то

$$p(0 \prec Z \prec z_{кр}) + p(Z \succ z_{кр}) = \frac{1}{2}, \tag{3.22}$$

откуда следует:

$$\Phi(t_{кр}) + \alpha = \frac{1}{2}, \text{ то есть } \Phi(t_{кр}) = \frac{1-2\alpha}{2} \tag{3.23}$$

Из последнего равенства по таблице интеграла вероятностей $\Phi(x)$ находится $t_{кр}$, а затем, по первой из формул (3.15), находится критическое значение $z_{кр}$ величины $Z = \bar{y}_в - \bar{x}_в$. И если экспериментальное значение $z_{эксп}$ этой величины больше $z_{кр}$ - гипотезу H_0 о равенстве $M(X)$ и $M(Y)$ отвергают, утверждая тем самым существенность (значимость) нововведенного фактора. А если окажется, что $z_{эксп} < z_{кр}$, то этот фактор признается незначимым.

Аналогичное видоизменение претерпит проверка гипотезы H_0: $M(X)=M(Y)$ при альтернативной гипотезе H_1: $M(Y)>M(X)$ и в случае малых выборок (пункт 3.5). В качестве критерия проверки гипотезы H_0 вводят величину

$$T = \frac{\bar{y}_в - \bar{x}_в}{\sqrt{s^2}} \cdot \sqrt{\frac{n_X \cdot n_Y}{n_X + n_Y}}, \qquad (3.24)$$

имеющую, при справедливости гипотезы H_0 и при равенстве дисперсий σ_X^2 и σ_Y^2, распределение Стьюдента с тем же числом степеней свободы $k = n_X + n_Y - 2$, что и в пункте 3.5. Гипотеза H_0 будет отвергнута, если экспериментальное значение $t_{эксп}$ величины T окажется положительным и достаточно большим ($t_{эксп} > t_{кр}$), где $t_{кр}$ найдется по таблице критических точек распределения Стьюдента из равенства

$$p(-t_{кр} \prec T \prec t_{кр}) = 1 - 2\alpha, \qquad (3.25)$$

представляющего собой аналог равенства (3.18). Действительно, так как распределение Стьюдента симметрично относительно точки t=0, то

$$p(0 \prec T \prec t_{кр}) + p(T \succ t_{кр}) = \frac{1}{2} \qquad (3.26)$$

Или

$$\frac{1}{2} p(-t_{кр} \prec T \prec t_{кр}) + \alpha = \frac{1}{2}, \qquad (3.27)$$

откуда и следует равенство (3.25).

3.6. Проверка гипотезы о нормальности распределения случайной величины (критерий Пирсона).

При исследовании различных признаков объектов генеральной совокупности обычно заведомо предполагается, что исследуемый признак X является случайной величиной, распределенной нормально. Но это можно подтвердить (или отвергнуть) с помощью выборки значений величины X.

Итак, выдвигается нулевая гипотеза H_0: исследуемый признак X объектов генеральной совокупности распределен нормально. Альтернативная гипотеза H_1: это не так. Принимается некоторый уровень значимости α. Требуется указать критерий, по которому можно было бы решить, принимать или отвергать гипотезу H_0.

Такой критерий, в частности, разработал американский математик Э. Пирсон. Чтобы воспользоваться этим критерием, сначала, естественно, нужно из генеральной совокупности сделать некоторую выборку. Причем выборку достаточно большую, по крайней мере объемом не менее 50. Полученное статистическое распределение выборки вида (1.1) или (1.3) затем несколько изменяют, приводя его к виду с равноотстоящими вариантами x_i и достаточно большими частотами n_i ($n_i \geq 5$, кроме разве что крайних частот). Для этого полученные в выборке варианты, если они неравноотстоящие, несколько сдвигают, а близлежащие малочисленные варианты объединяют в одну с объединением

их частот. В дальнейшем будем считать, что такая предварительная работа проделана, так что статистическое распределение выборки

x_i	x_1	x_2	…	x_m
n_i	n_1	n_2	…	n_m

$(n_1+n_2+\ldots+n_m=n)$ (3.28)

уже удовлетворяет указанным выше требованиям. Будем также считать, что из (3.28) найдены $\bar{x}_в$ и $s_в$.

Пусть, в соответствии с гипотезой H_0, признак X (случайная величина X) распределена нормально с некоторыми неизвестными параметрами a и σ. В качестве значений этих параметров примем их известные точечные оценки $a \approx \bar{x}_в$ и $\sigma \approx s_в$. Далее, разобьем ось ox с нанесенными на нее m равноотстоящими с шагом h вариантами x_i (i=1, 2,…m) на m интервалов $(z_{i-1}; z_i)$, окружающими эти варианты – так, как показано на рис. 3.6.

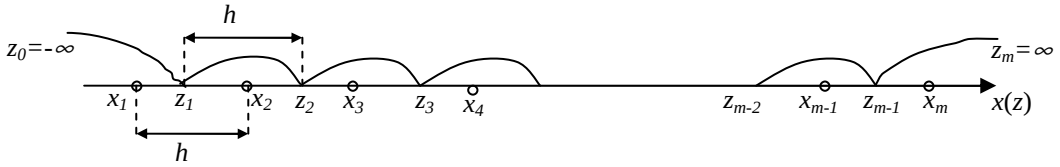


Рис. 3.6

Если мы найдем вероятности p_i попадания значения исследуемой величины X в интервалы $(z_{i-1}; z_i)$, окружающие варианты x_i, и умножим эти вероятности на объем n выборки, то получим частоты n_i^*, являющиеся математическими ожиданиями экспериментальных частот n_i. То есть получим теоретические частоты n_i^* вариант x_i:

x_i	x_1	x_2	…	x_m
n_i^*	n_1^*	n_2^*	⋯	n_m^*

$(n_i^* = np_i)$ (3.29)

Незначительное расхождение экспериментальных и теоретических частот n_i и n_i^* будет свидетельствовать в пользу принятия гипотезы H_0. А если же они будут сильно различаться, то это будет означать, что гипотеза H_0, скорее всего, неверна.

В качестве критерия близости (согласия) частот n_i и n_i^* Пирсоном было предложено использовать величину

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - n_i^*)^2}{n_i^*} \qquad (3.30)$$

Эта величина – случайная, так как ее значения от выборки к выборке меняются. Ясно, что все ее значения неотрицательны, и чем ближе она к нулю, тем ближе между собой n_i и n_i^*, а значит, тем в большей мере подтверждается гипотеза H_0.

Доказано, что при $n\to\infty$ закон распределения случайной величины χ^2, определенной равенством (3.30), стремится к закону распределения χ^2 (хи-квадрат, см. часть I, главу 2, §4) с $k=m-3$ степенями свободы. Причем даже независимо от того, по какому закону распределен признак X в генеральной совокупности. Поэтому случайная величина (3.30) и обозначена символом χ^2. А сам критерий Пирсона называется еще *«критерием «хи-квадрат»*.

Область принятия и критическую область проверяемой гипотезы H_0 о нормальности распределения случайной величины X определяют следующим образом. По заданному уровню значимости α и числу $k=m-3$ с помощью таблицы критических точек распределения «хи-квадрат» (таблица 3 Приложения) находят такое критическое значение $\chi^2_{кр}(\alpha;\, k)$ величины χ^2, чтобы (см. рис.3.7)

$$p(\chi^2 \succ \chi^2_{кр}(\alpha;k)) = \alpha \qquad (3.31)$$

И если $\chi^2_{эксп}$, вычисленное по формуле (3.30), окажется больше $\chi^2_{кр}(\alpha;\, k)$, то гипотезу H_0 о нормальности распределения случайной величины X отвергают. А если окажется, что $\chi^2_{эксп} < \chi^2_{кр}(\alpha;\, k)$, то гипотезу H_0 принимают.

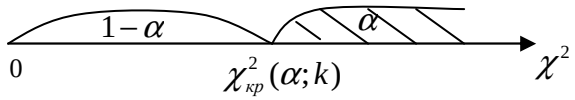


Рис.3.7

Нам осталось лишь указать, как найти вероятность p_i попадания значений нормально распределенной случайной величины X с параметрами $a\approx\bar{x}_в$ и $\sigma\approx s_в$ в интервалы $(z_{i-1};\, z_i)$, окружающие на рис. 3.6 варианты x_i (i=1,2,…m), ибо через эти вероятности p_i по формуле $n_i^* = np_i$ находятся теоретические частоты n_i^* (i=1,2,…m). Для нахождения указанных вероятностей следует применить формулу (4.10) (часть I, глава 2) для нормально распределенных случайных величин, согласно которой получаем:

$$p_i = p(z_{i-1} \prec X \prec z_i) \approx \Phi\left(\frac{z_i - \bar{x}_в}{s_в}\right) - \Phi\left(\frac{z_{i-1} - \bar{x}_в}{s_в}\right) \quad (i = 1,2,...m) \qquad (3.32)$$

Здесь, как это следует из рис. 3.6,

$$z_i = x_i + \frac{h}{2} \ \ (i = 1,2,...m-1); \quad z_0 = -\infty; \quad z_m = \infty \qquad (3.33)$$

Пример5. При исследовании некоторого признака X объектов генеральной совокупности выборочным путем обследовано 100 объектов. Данные выборки представлены в таблице:

x_i	n_i	x_i	n_i	x_i	n_i
1,00	1	1,19	2	1,37	6
1,03	3	1,20	4	1,38	2
1,05	6	1,23	4	1,39	1
1,06	4	1,25	8	1,40	2
1,08	2	1,26	4	1,44	3
1,10	4	1,29	4	1,45	3
1,12	3	1,30	6	1,46	2
1,15	6	1,32	4	1,49	4
1,16	5	1,33	5	1,50	2

При уровне значимости α=0,05 проверить гипотезу H_0 о нормальности распределения признака X в генеральной совокупности.

Решение. Варианты нашей выборки не равностоящие, причем малочисленные. Сведем их к небольшому числу равноотстоящих вариант с достаточно большими частотами. Для этого предварительно оформим данное статистическое распределение выборки в интервальном виде, чтобы затем принять середины получавшихся интервалов за новые равноотстоящие варианты.

Сначала определим минимальную и максимальную варианты исходного статистического распределения выборки:

$$x_{min}=1{,}00; \quad x_{max}=1{,}50$$

Таким образом, размах вариации

$$\Delta = x_{max} - x_{min} = 0{,}50$$

При этом n=100–количество вариант (объем выборки). А теперь разобьем промежуток вариации $[x_{min}; x_{max}]=[1{,}00;\ 1{,}50]$ точками x_i^* на несколько (m) промежутков одинаковой ширины h – так, как показано на рис. 3.8.

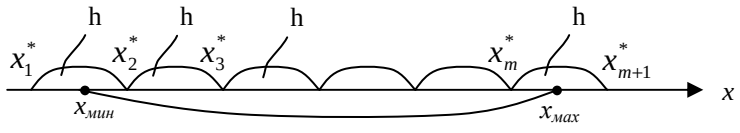


Рис. 3.8

Для выбора оптимального (не слишком большого и не слишком малого) шага разбиения h применим *формулу Стэрджеса*:

$$h \approx \frac{\Delta}{1+3{,}322 \cdot \lg n} \tag{3.34}$$

В нашем примере эта формула дает:

$$h \approx \frac{0{,}50}{1+3{,}322 \cdot \lg 100} = \frac{0{,}50}{1+3{,}322 \cdot 2} = 0{,}0654$$

Теперь, в соответствии с рис. 3.8, добавляя к размаху вариации Δ один шаг h (по полшага с обеих сторон), находим (округляя по недостатку до целого) оптимальное количество m интервалов, на которые мы разобьем интервал вариации:

$$m = \frac{\Delta + h}{h} = \frac{0{,}50 + 0{,}0654}{0{,}0654} = 8{,}64 \approx 8$$

А теперь, исходя из выбранного значения m=8 количества интервалов, получим и окончательное значение длины h каждого интервала:

$$m = \frac{\Delta + h}{h} \Rightarrow 8 = \frac{\Delta + h}{h} \Rightarrow 7h = \Delta \Rightarrow h = \frac{\Delta}{7} = \frac{0{,}50}{7} = 0{,}071428... \approx 0{,}0714$$

(окончательное значение h для борьбы с накоплением погрешностей взято с двумя дополнительными десятичными знаками по сравнению с исходными вариантами, то есть в нашей задаче – до десятитысячных).

После того, как количество интервалов m=8 и шаг разбиения h=0,0714 определены, запишем и сами интервалы $(x_i^*; x_{i+1}^*)$ (i=1,2,…8):

$(x_1^*\ x_2^*)=(x_{min}-h/2;\ x_{min}+\ h/2)=(1{,}00-0{,}0714/2;\ 1{,}00+0{,}0714/2)=(0{,}9643;\ 1{,}0357)$;
$(x_2^*; x_3^*)=(1{,}0357;\ 1{,}0357+h)=(1{,}0357;\ 1{,}1071)$;
$(x_3^*; x_4^*)=(1{,}1071;\ 1{,}1071+h)=(1{,}1071;\ 1{,}1785)$;
$(x_4^*; x_5^*)=(1{,}1785;\ 1{,}1785+h)=(1{,}1785;\ 1{,}2499)$;
$(x_5^*; x_6^*)=(1{,}2499;\ 1{,}2499+h)=(1{,}2499;\ 1{,}3213)$;
$(x_6^*; x_7^*)=(1{,}3213;\ 1{,}3213+h)=(1{,}3213;\ 1{,}3927)$;
$(x_7^*; x_8^*)=(1{,}3927;\ 1{,}3927+h)=(1{,}3927;\ 1{,}4641)$;
$(x_8^*; x_9^*)=(1{,}4641;\ 1{,}4641+h)=(1{,}4641;\ 1{,}5355)$.

В каждом из полученных восьми интервалов проведем подсчет количества вариант исходного распределения выборки, попавших в эти интервалы. Это будут новые частоты для новых вариант x_i – середин получившихся интервалов. Таким образом, приходим к следующей таблице:

№ интервала	Интервалы	Середины интервалов (новые варианты x_i)	Частоты n_i
1	0,9643 – 1,0357	1,0000	4
2	1,0357 – 1,1071	1,0714	16
3	1,1071 – 1,1785	1,1428	14
4	1,1785 – 1,2499	1,2142	10
5	1,2499 – 1,3213	1,2856	26
6	1,3213 – 1,3927	1,3570	14
7	1,3927 – 1,4641	1,4284	10
8	1,4641 – 1,5355	1,4998	6

Данные этой таблицы можно, для наглядности, представить и в виде гистограммы частот, и в виде полигона частот (рис.3.9 и ломаная на рис.3.10):

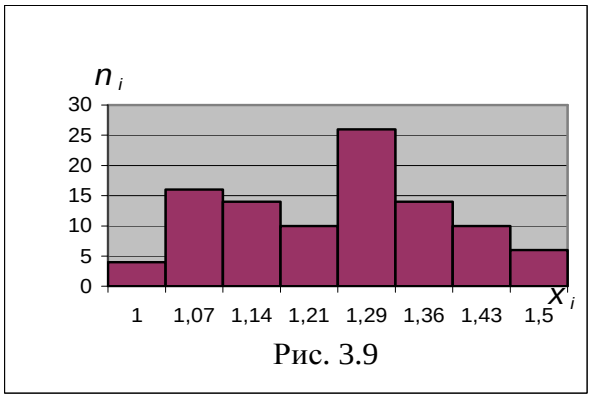


Рис. 3.9

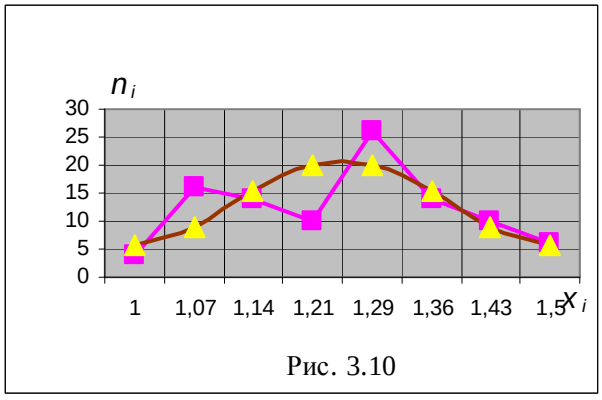


Рис. 3.10

И на гистограмме, и на полигоне частот обнаружился провал частот в их средней части, которого не должно быть, если исследуемый признак X распределен нормально. То есть этот провал является доводом против выдвинутой гипотезы H_0 о том, что признак X распределен нормально. Но этот провал мог образоваться и случайно – в силу случайности самой выборки. Так что пока неясно, принимать или отвергать гипотезу H_0. Обоснованный вывод мы сделаем, если

наряду с найденными экспериментальными частотами n_i вариант x_i найдем и их теоретические частоты n_i^*, которые следуют из гипотезы H_0, а затем сравним те и другие частоты по критерию Пирсона.

Для этого сначала по новым вариантам x_i и их частотам n_i найдем $\bar{x}_в$ и $s_в$ – выборочную среднюю и исправленное выборочное среднее квадратическое отклонение:

$$\bar{x}_в = \frac{\sum_{i=1}^{m} x_i n_i}{n} = \frac{\sum_{i=1}^{8} x_i n_i}{100} = 1{,}2499 \approx 1{,}25$$

$$D_в = \overline{x_в^2} - (\bar{x}_в)^2 = \frac{\sum_{i=1}^{8} x_i^2 n_i}{100} - (1{,}2499)^2 = 0{,}0182$$

$$s_в = \sqrt{\frac{n}{n-1} \cdot D_в} = \sqrt{\frac{100}{99} \cdot 0{,}0182} \approx 0{,}1356$$

Если исследуемая случайная величина X распределена нормально, то заменяя ее неизвестные параметры a и σ их точечными оценками $\bar{x}_в$ и $s_в$, получим:

$$a \approx \bar{x}_в = 1{,}25; \quad \sigma \approx s_в = 0{,}1356$$

Теперь, в соответствии с рис. 3.6, найдем по формулам (3.33) числа z_i (i=0, 1, 2,…8):

$z_o = -\infty$; $z_1 = x_1 + h/2 = 1{,}0000 + 0{,}0714/2 = 1{,}0357$;
$z_2 = x_2 + h/2 = 1{,}0714 + 0{,}0714/2 = 1{,}1071$; $z_3 = 1{,}1785$;
$z_4 = 1{,}2499$; $z_5 = 1{,}3213$; $z_6 = 1{,}3927$; $z_7 = 1{,}4641$; $z_8 = +\infty$

После этого по формуле (3.32) подсчитаем вероятности p_i попадания значений величины X в интервалы $(z_{i-1}; z_i)$ (i=1, 2,…8), а по ним по формулам $n_i^* = np_i$ (i=1, 2,…m) подсчитаем теоретические частоты n_i* наших новых вариант x_i (i=1, 2,…8). Подсчет этих теоретических частот оформим в виде таблицы, округляя, для точности, итоговые частоты n_i* до десятых, то есть до одного лишнего десятичного знака:

i	z_{i-1}	z_i	$\frac{z_{i-1}-\bar{x}_в}{s_в}$	$\frac{z_i-\bar{x}_в}{s_в}$	$\Phi\left(\frac{z_{i-1}-\bar{x}_в}{s_в}\right)$	$\Phi\left(\frac{z_i-\bar{x}_в}{s_в}\right)$	p_i	$n_i^*=np_i$ $=100\,p_i$
1	$-\infty$	1,0357	$-\infty$	-1,580	-0,5	-0,443	0,057	5,7
2	1,0357	1,1071	-1,580	-1,054	-0,443	-0,354	0,089	8,9
3	1,1071	1,1785	-1,054	-0,527	-0,354	-0,201	0,153	15,3
4	1,1785	1,2499	-0,527	0	-0,201	0	0,201	20,1
5	1,2499	1,3213	0	0,526	0	0,201	0,201	20,1
6	1,3213	1,3927	0,526	1,052	0,201	0,354	0,153	15,3
7	1,3927	1,4641	1,052	1,579	0,354	0,443	0,089	8,9
8	1,4641	$+\infty$	1,579	$+\infty$	0,443	0,5	0,057	5,7
$\sum$							$\sum p_i$=1	$\sum n_i^* = 100$

Полученные для одних и тех же вариант x_i теоретические частоты n_i* и экспериментальные частоты n_i можем теперь сравнить. Сначала сделаем это визуально:

n_i	4	16	14	10	26	14	10	6
n_i*	5,7	8,9	15,3	20,1	20,1	15,3	8,9	5,7

Как видим, есть и небольшие расхождения, и существенные (особенно в паре 10 и 20,1) Особенно наглядно видны эти расхождения, если наряду с полигоном реальных частот n_i (ломаной) построить и полигон теоретических частот n_i* - плавную кривую нормального распределения (см. рис. 3.10).

Сравним, однако, частоты n_i и n_i* на согласованность с помощью критерия Пирсона. Так как условие «все частоты n_i должны быть не менее 5, кроме разве что крайних» у нас выполняется, то объединять частоты не будем. В соответствии с формулой (3.30) подсчитаем экспериментальное значение величины χ^2:

$$\chi^2 = \chi^2_{эксп} = 13{,}35$$

А теперь по таблице критических точек распределения «хи-квадрат» (таблица 3 Приложения) для уровня значимости α=0,05 и числа степеней свободы $k=m$-3=8-3=5 найдем критическое значение величины χ^2:

$$\chi^2_{кр} = \chi^2_{кр}(0{,}05;\ 5) = 11{,}1$$

Сравнивая $\chi^2_{эксп}$ и $\chi^2_{кр}$ видим, что $\chi^2_{эксп} > \chi^2_{кр}$. Таким образом, гипотезу H_0 о нормальности исследуемого признака X отвергаем – она не подтверждается экспериментальными данными. То есть расхождения между реальными и теоретическими частотами являются слишком существенными, чтобы, при справедливости гипотезы H_0, их можно было отнести лишь на счет случайности самой выборки.

Впрочем, мы можем и ошибаться - гипотеза H_0 о нормальности распределения признака X объектов генеральной совокупности на самом деле может быть верна. В таком случае, отвергая ее, мы совершаем ошибку 1-го рода. И вероятность этой ошибки – это принятый нами уровень значимости α=0,05.

Упражнения

1. Станок-автомат должен изготовлять детали массой 20г. Известно, что средняя квадратичная ошибка в работе станка равна 0,1г. Из продукции станка наудачу отобрано 10 деталей. Средняя масса одной детали оказалась равной 20,08г. При уровне значимости а) α=0,05 и б) α=0,01 проверить гипотезу H_0 о том, что станок настроен правильно при альтернативной гипотезе H_1, что станок настроен неправильно (производит в целом увеличенные или уменьшенные детали).

Ответ: а) гипотеза H_0 отвергается; б) гипотеза H_0 принимается.

2. Решить предыдущую задачу по проверке гипотезы H_0 при альтернативной гипотезе H_1, состоящей в том, что станок производит в целом увеличенные по массе детали.

Ответ: и в варианте а), и в варианте б) гипотеза H_0 отвергается.

3. Исследовались ошибки X и Y двух однотипных приборов. Экспериментальные данные о допущенных ошибках измерений на этих приборах (они выявлены с помощью более точных приборов) таковы:

x_i	0,05	1,50	-1,35	-1,12	-0,52
n_i	1	1	1	1	1

y_i	1,82	0,10	-0,56	0,24	0,17	0,23	-0,31
n_i	1	1	1	1	1	1	1

При уровне значимости α=0,05 проверить гипотезу H_0 о равной точности обоих приборов при альтернативной гипотезе H_1 о более высокой точности второго прибора.

Ответ: гипотеза H_0 принимается.

4. В результате двух серий измерений с количеством измерений n_1=25 и n_2=50 получены следующие средние значения измеряемых величин a и b соответственно: $\overline{x}_в = 9{,}79$ и $\overline{y}_в = 9{,}60$. При уровне значимости α=0,01 проверить нулевую гипотезу H_0 о равенстве a и b при альтернативной гипотезе H_1 о неравенстве a и b, если известно, что оба измерения равноточны со средним квадратическим отклонением σ = 0,30.

Ответ: гипотеза H_0 отвергается.

5. Рассмотреть упражнение 3 с теми же данными, но при неизвестном σ. Взамен использовать исправленные выборочные дисперсии s_X^2=0,28 и s_Y^2=0,33 указанных двух серий измерений.

Ответ: гипотеза H_0 принимается.

6. Уровень исполнительского мастерства участников конкурса оценивался двумя судьями в баллах. Известны результаты оценки пяти первых участников конкурса:

Оценки первого судьи (x_i)	6	7	8	5	7
Оценки второго судьи (y_i)	7	6	8	7	8

При уровне значимости α=0,05 установить, значимо или незначимо различаются результаты оценок судей.

Ответ: оценки судей различаются незначимо.

Указание. Разность $Z=X-Y$ оценок судей в силу большого числа факторов, влияющих на эту разность, можно, очевидно, считать случайной величиной, распределенной нормально. Если оценки судей различаются незначимо, то математическое ожидание величины Z равно нулю. А если значимо, то $M(Z)\neq 0$. Поэтому задача состоит в том, чтобы при уровне значимости $\alpha=0{,}05$ проверить гипотезу о том, что $M(Z)= a = 0$ при альтернативной гипотезе H_1 о том, что $M(Z)= a \neq 0$ при объеме выборки $n=5$.

7. Решить предыдущую задачу 6 по проверке гипотезы H_0 о незначимости различия в оценках двух судей при альтернативной гипотезе H_1, состоящей в том, что второй судья ставит в целом более высокие оценки, чем первый.

Ответ: гипотеза H_0 принимается.

8. Для исследования массы X клубней кормовой свеклы из урожая случайным образом отобрано 100 клубней. Статистическое распределение выборки оказалось таковым:

x_i(кг)	0 - 1,0	1,0 – 2,0	2,0 – 3,0	3,0 – 4,0	4,0 – 5,0	5,0 – 6,0
n_i	20	20	28	22	12	8

Выдвинув гипотезу H_0 о нормальном распределении массы X клубней, найти соответствующие этой гипотезе теоретические частоты n_i^* (i=1,2,…6) для указанной выше выборки. С помощью критерия Пирсона при уровне значимости а) α=0,05 и б) α=0,01 подтвердить или отвергнуть гипотезу H_0.

Ответ:

n_i	20	20	28	22	12	8
n_i^*	8	18	29	26	14	5

Гипотеза H_0 отклоняется при обоих уровнях значимости.

§4. Дисперсионный анализ

В предыдущем параграфе мы рассмотрели схему проверки гипотезы о несущественности (незначимости) различия выборочных средних двух нормально распределенных случайных величин. Теперь рассмотрим обобщение этой задачи: требуется проверить гипотезу H_0 о несущественности (незначимости) различия выборочных средних m нормально распределенных случайных величин ($m>2$) при альтернативной гипотезе H_1 о том, что хотя бы некоторые из указанных выборочных средних различаются существенно (значимо).

Можно, конечно, опираясь на методику предыдущего параграфа, сравнить указанные выборочные средние попарно. Но для не слишком малых значений m это выливается в трудоемкую и громоздкую проблему. Гораздо эффективнее решается эта проблема с помощью так называемого *дисперсионного анализа*,

при котором сравниваются между собой не выборочные средние, а некие обобщенные дисперсии, о которых речь пойдет ниже.

На практике дисперсионный анализ применяют, когда хотят выяснить, оказывает или не оказывает влияние на нормально распределенную случайную величину X некоторый *качественный фактор* A, который имеет m различных качественных реализаций (уровней). Например, если X – урожайность некоторой сельскохозяйственной культуры, то качественным фактором A, чье влияние на X может исследоваться, может быть вид (марка) удобрения; или режим ухода за растениями (прополки, полива); или технология уборки, и т.д. А если X – прибыль предприятия, то качественным фактором A, влияющим (или не влияющим) на прибыль X, может быть технология производства; качество сырья; структура управления производством; система материального или морального стимулирования работников, и т.д.

Если исследуется влияние на величину X лишь одного качественного фактора A, то говорят об однофакторном дисперсионном анализе. А если сразу нескольких – то о многофакторном.

С помощью дисперсионного анализа исследуется значимость влияния на наблюдаемую величину X каждого из факторов, сравнивается их влияние между собой, устанавливается факт их взаимодействия, и т.д. Первоначально дисперсионный анализ был предложен Р. Фишером (1925 г.) для обработки результатов агрономических опытов. В дальнейшем этот метод стал использоваться всюду, где требуется математическая обработка результатов экспериментов.

Примечание. Если факторы, чье влияние на величину X исследуется, являются *количественными* (уровни $(a_1; a_2; ...a_m)$ каждого из таких факторов A – числа), то дисперсионный анализ можно применять и в этом случае. Но гораздо более емкие и глубокие выводы о характере влияния таких факторов на величину X мы сделаем, если вместо дисперсионного анализа применим корреляционно-регрессионный анализ, которому будет посвящен следующий параграф.

Главная идея дисперсионного анализа, как однофакторного, так и многофакторного, заключается в расчленении общего объема вариации (колеблемости) значений случайной величины X по источникам ее образования. В частности, при однофакторном дисперсионном анализе используется разложение

$$W_{общ} = W_A + W_{ост} \qquad (4.1)$$

где $W_{общ}$ – общий объем вариации изучаемой случайной величины X; W_A – объем вариации, обусловленный действием на величину X фактора A; $W_{ост}$ остаточный объем вариации, вызванный действием всех остальных неучтенных и считающихся случайными факторов (помех). Последующее сравнение рассчитанных на одну степень свободы объемов вариации W_A и $W_{ост}$, то есть сравнение соответствующих этим объемам исправленных дисперсий s_A^2 и $s_{ост}^2$ позволяет оценить существенность (значимость) влияния на изучаемую величину X данного фактора A.

При двухфакторном дисперсионном анализе выясняется не только значимость каждого из факторов A и B в отдельности, но и значимость их взаимодействия AB. Для это общая вариация представляется в виде

$$W_{общ}=W_{A+B}+W_{ост}=W_A+W_B+W_{AB}+W_{ост}\,, \qquad (4.2)$$

а затем анализируются соотношения между величинами W_A, W_B, W_{AB} и $W_{ост}$ (точнее, между исправленными дисперсиями s_A^2, s_B^2, s_{AB}^2 и $s_{ост}^2$).

Мы ограничимся рассмотрением лишь однофакторного дисперсионного анализа, ибо многофакторный анализ много сложнее и осуществляется обычно не вручную, а с помощью стандартных программ на ЭВМ.

Итак, пусть X – некоторая изучаемая случайная величина, и исследуется воздействие на нее некоторого фактора A, имеющего m различных фиксированных уровней (a_1; a_2;…; a_m) (качественных или количественных). В частности, если X – урожайность культуры, а A – удобрение, то качественными уровнями этого удобрения могут быть различные виды этого удобрения, а количественными уровнями – дозы вносимого удобрения. И т.д.

Будем считать, что проведено n измерений величины X при каждом из m уровней фактора A, так что таблица опытных данных (статистическое распределение выборки) состоит из $N=nm$ данных и имеет вид:

№ измерения	Уровни фактора А				
	a_1	a_2	a_3	…	a_m
1	x_{11}	x_{12}	x_{13}	…	x_{1m}
2	x_{21}	x_{22}	x_{23}	…	x_{2m}
…	…	…	…	…	…
n	x_{n1}	x_{n2}	x_{n3}	…	x_{nm}
Групповые средние	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	…	$\bar{x}_m$

В последней строке таблицы приведены средние значения величины X для соответствующих уровней фактора A (групповые средние). Они подсчитываются как средние по каждому столбцу:

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \ (j=1,2,...,m) \qquad (4.3)$$

Общая средняя $\bar{x}$ величины X может быть найдена по формуле

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij} \qquad (4.4)$$

Или проще, если найдены групповые средние (4.3):

$$\bar{x} = \frac{1}{m}\sum_{j=1}^{m} \bar{x}_j \qquad (4.5)$$

Рассмотрим теперь общий объем вариации (колеблемости) значений x_{ij} величины X вокруг ее общей средней $\bar{x}$:

$$W_{общ} = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij}-\bar{x})^2 \qquad (4.6)$$

Разложим, в соответствии с (4.1), $W_{общ}$ на W_A и $W_{ост}$:

$$W_{общ} = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(x_{ij} - \bar{x}\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\left[(x_{ij} - \bar{x}_j) + (\bar{x}_j - \bar{x})\right]^2 =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \bar{x}_j)^2 + 2\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \bar{x}_j)(\bar{x}_j - \bar{x}) + \sum_{i=1}^{n}\sum_{j=1}^{m}(\bar{x}_j - \bar{x})^2 = \quad (4.7)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \bar{x}_j)^2 + 2\sum_{j=1}^{m}(\bar{x}_j - \bar{x})\sum_{i=1}^{n}(x_{ij} - \bar{x}_j) + n\sum_{j=1}^{m}(\bar{x}_j - \bar{x})^2$$

Теперь учтем, что

$$\sum_{i=1}^{n}(x_{ij} - \bar{x}_j) = \sum_{i=1}^{n}x_{ij} - n\bar{x}_j = n\bar{x}_j - n\bar{x}_j = 0 \quad (4.8)$$

и введем обозначения:

$$W_A = n\sum_{j=1}^{m}(\bar{x}_j - \bar{x})^2; \;\; W_{ост} = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \bar{x}_j)^2 \quad (4.9)$$

Тогда выражение (4.7) для $W_{общ}$ примет вид:

$$W_{общ} = W_A + W_{ост} \quad (4.10)$$

Составляющая W_A общей вариации $W_{общ}$ действительна связана с действием фактора A, так как она определяется разбросом групповых средних $\bar{x}_j$, соответствующих разным уровням фактора A, вокруг общей средней $\bar{x}$. И если фактор A действительно влияет на величину X, то групповые средние $\bar{x}_j$ будут существенно разными, и величина W_A будет значительной. А если влияние фактора A несущественно, то групповые средние $\bar{x}_j$ будут мало отличатся между собой, и W_A будет малой. Впрочем, значительность или незначительность величины W_A можно установить лишь путем сравнения ее с $W_{ост}$, которая, очевидно, характеризует степень внутригруппового разброса значений x_{ij}. И сравниваются они следующим образом. На основании подсчитанных сумм $W_{общ}$, W_A и $W_{ост}$ находятся общая $s^2_{общ}$, факторная s^2_A и остаточная $s^2_{ост}$ исправленные дисперсии, которые вычисляются путем деления указанных сумм на соответствующее каждой из них число степеней свободы

$$k_{общ} = N - 1 = nm - 1; \;\; k_A = m - 1; \;\; k_{ост} = N - m = m(n-1) \quad (4.11)$$

Напомним (см. также §2), что число степеней свободы некоторой вычисляемой по опытным данным случайной величины – это количество этих данных, участвующих в ее формировании, за вычетом количества независимых друг от друга числовых характеристик, вычисляемых по тем же опытным данным и тоже участвующих в формировании указанной величины. Из сказанного и выражений для $W_{общ}$, W_A и $W_{ост}$ и следуют формулы (4.11).

Итак,

$$s^2_{общ} = \frac{W_{общ}}{N-1}; \;\; s^2_A = \frac{W_A}{m-1}; \;\; s^2_{ост} = \frac{W_{ост}}{N-m} \quad (4.12)$$

Теперь выдвинем нулевую гипотезу H_0 - фактор A несущественно (незначимо) влияет на величину X при альтернативной гипотезе H_1, что он на неё влияет существенно (значимо).

Если проверяемая гипотеза H_0 справедлива, то найденные групповые средние $\bar{x}_j$ должны несущественно (незначимо) отличаться друг от друга. То есть математические ожидания всех групповых средних должны быть одинаковы и равны общей средней $\bar{x}$. И будет это не так, если гипотеза H_0 несправедлива.

Для применимости дисперсионного анализа при проверке гипотезы H_0 должны выполняться *два обязательных условия*:

1. Должны незначимо различаться все внутригрупповые дисперсии. То есть все остальные случайные (неучтенные) факторы на любом из уровней фактора A должны действовать на величину X одинаково.

2. Величина X должна иметь нормальное распределение на любом из уровней фактора A.

При выполнении этих условий для проверки гипотезы H_0 можно использовать критерий Фишера-Снедекора

$$F = \frac{s_A^2}{s_{ост}^2} \tag{4.13}$$

Доказано (см. например [5], стр. 379-380), что при справедливости гипотезы H_0 (при незначимости фактора A) все три дисперсии (4.12) являются точечными несмещенными оценками генеральной дисперсии σ^2. То есть их математические ожидания равны σ^2:

$$M(s_{общ}^2) = M(s_A^2) = M(s_{ост}^2) = \sigma^2 \tag{4.14}$$

А значит, при справедливости гипотезы H_0, факторная дисперсия s_A^2 должна быть приблизительно такой же, как и остаточная дисперсия $s_{ост}^2$. И тогда будет $F \approx 1$. Но если гипотеза H_0 неверна (фактор A значим для величины X), то факторная дисперсия должна быть существенно больше остаточной, и тогда величина F должна быть существенно больше 1.

Доказано, что при указанных выше условиях и при справедливости гипотезы H_0 случайная величина F, определяемая формулой (4.13), распределена по закону Фишера-Снедекора со степенями свободы $k_A = m-1$ и $k_{ост} = N-m$. Поэтому для проверки гипотезы H_0 при заданном уровне значимости α достаточно с помощью таблицы критических точек распределения Фишера-Снедекора для данных значений (α; k_A; $k_{ост}$) найти критическое значение $f_{кр}$ случайной величины F и сравнить его с экспериментальным значением $f_{эксп}$, подсчитанным по формуле (4.13). И если $f_{эксп} > f_{кр}$ то гипотезу H_0 отвергают. То есть считают, что фактор A существенно (значимо) влияет на величину X, а значит, различие найденных групповых средних $\bar{x}_j$ не случайно. А если $f_{эксп} < f_{кр}$, то гипотезу H_0 принимают. То есть считают, что фактор A является незначимым – его влияние на величину X в принципе такое же, как и влияние прочих неучтенных случайных факторов (помех). И значит, различие групповых средних $\bar{x}_j$ – случайно.

Критерий Фишера-Снедекора позволяет установить наличие или отсутствие существенных (значимых) различий между групповыми средними *в целом*, однако он не показывает, между какими средними разница существенна, а между какими нет. Поэтому если проведенный дисперсионный анализ привел к отказу от нулевой гипотезы H_0, предполагающей равенство групповых средних,

и показал, таким образом, существенность влияния проверяемого фактора *A* на величину *X*, то этот общий вывод необходимо дополнить проверкой существенности различий между парами средних. Наиболее просто это сделать, сравнив модули всех разностей групповых средних $\bar{x}_j$ с так называемой *HCP – наименьшей существенной разностью* (см. [1], § 6.3):

$$HCP = t_{кр} \cdot \sqrt{\frac{2s^2_{ост}}{m}} \tag{4.15}$$

Здесь $t_{кр}$ – критическое значение распределения Стьюдента для заданного уровня значимости α и числа степеней свободы $k_{ост} = N - m$. Разности, по модулю бóльшие *HCP*, считаются значимыми, меньшие – незначимыми.

Примечание 1. Формулы (4.7) и (4.9) не очень удобны для ручного подсчета сумм $W_{общ}$, W_A и $W_{ост}$, что затрудняет подсчет общей, факторной и остаточной дисперсий (4.12). Но эти формулы можно и существенно упростить.

Во-первых, легко показать (проделайте это самостоятельно), что указанные формулы можно преобразовать к виду:

$$\begin{aligned} W_{общ} &= \sum_{j=1}^{m} R_j - \frac{1}{N}\left(\sum_{j=1}^{m} L_j\right)^2; \\ W_A &= \frac{1}{n}\sum_{j=1}^{m} L_j^2 - \frac{1}{N}\left(\sum_{j=1}^{m} L_j\right)^2; \quad W_{ост} = W_{общ} - W_A, \end{aligned} \tag{4.16}$$

где

$$R_j = \sum_{i=1}^{n} x_{ij}^2; \quad L_j = \sum_{i=1}^{n} x_{ij} \quad (j = 1,2,...,m) \tag{4.17}$$

Во-вторых, учтем, что замена x_{ij} на y_{ij}, определяемая равенствами

$$y_{ij} = x_{ij} - C, \tag{4.18}$$

где *C* – любая константа, не изменит, очевидно, ни одну из вариаций $W_{общ}$, W_A и $W_{ост}$. При этом для упрощения счета удобно в качестве константы *C* взять число, близкое к общей средней $\bar{x}$, что позволит кардинально уменьшить числа x_{ij}, если они велики. В итоге для $W_{общ,}$ W_A и $W_{ост}$ получим окончательно:

$$\begin{aligned} W_{общ} &= \sum_{j=1}^{m} P_j - \frac{1}{N}\left(\sum_{j=1}^{m} T_j\right)^2; \\ W_A &= \frac{1}{n}\sum_{j=1}^{m} T_j^2 - \frac{1}{N}\left(\sum_{j=1}^{m} T_j\right)^2; \quad W_{ост} = W_{общ} - W_A \end{aligned} \tag{4.19}$$

Здесь

$$N = nm; \quad P_j = \sum_{i=1}^{n} y_{ij}^2; \quad \mathrm{T}_j = \sum_{i=1}^{n} y_{ij} \quad (j = 1,2,...,m) \tag{4.20}$$

Все эти суммы легко вычисляются вручную.

Отметим еще, что при замене (4.18) x_{ij} на y_{ij} не изменятся, очевидно, и все разности между групповыми средними $\bar{x}_j$, которые заменятся на разности между групповыми средними $\bar{y}_j$. Поэтому выяснение вопроса о том, какие из раз-

ностей между $\bar{x}_j$ значимы, а какие нет, можно заменить выяснением этого вопроса о разностях между $\bar{y}_j$.

Примечание 2. На практике не всегда удается гарантировать одинаковое количество n экспериментальных данных x_{ij} на каждом уровне a_i фактора A.

Пусть это количество n_j зависит от номера j уровня фактора. Тогда общий объем N выборки будет, очевидно, таков:

$$N = \sum_{j=1}^{m} n_j \tag{4.21}$$

Формулы (4.19) для этого случая примут вид:

$$W_{общ} = \sum_{j=1}^{m} P_j - \frac{1}{N}\left(\sum_{j=1}^{m} T_j\right)^2;$$

$$W_A = \sum_{j=1}^{m} \frac{1}{n_j} T_j^2 - \frac{1}{N}\left(\sum_{j=1}^{m} T_j\right)^2; \quad W_{ост} = W_{общ} - W_A \tag{4.22}$$

Здесь

$$P_j = \sum_{i=1}^{n_j} y_{ij}^2; \quad \mathrm{T}_j = \sum_{i=1}^{n_j} y_{ij} \quad (j = 1,2,...,m) \tag{4.23}$$

Общая, факторная и остаточная дисперсии должны подсчитываться по тем же формулам (4.12):

$$s_{общ}^2 = \frac{W_{общ}}{N-1}; \quad s_A^2 = \frac{W_A}{m-1}; \quad s_{ост}^2 = \frac{W_{ост}}{N-m} \tag{4.24}$$

Формулы (4.13) и (4.15) и схемы их использования остаются без изменений.

Номер измерения	Уровни фактора A		
	a_1	a_2	a_3
1	151	152	142
2	152	154	144
3	156	156	150
4	157	158	152
Групповые средние	154	155	147

Пример 1. Исследовать влияние фактора A на случайную величину X при указанной слева таблице экспериментальных данных. Предполагается, что выборки произведены из нормальных генеральных совокупностей с одинаковыми дисперсиями. Методом дисперсионного анализа при уровне значимости α = 0,05 проверить гипотезу H_0 о несущественности (незначимости) влияния данного фактора A на изучаемую величину X.

Решение. Для упрощения расчетов вычтем некую среднюю варианту, скажем C = 152, из каждого наблюдённого значения:

$$y_{ij} = x_{ij} - 152$$

После этого составим необходимую для реализации формул (4.19) и (4.20) расчетную таблицу:

Номер измерения	Уровни фактора A						$\sum_{j=1}^{3}$
	a_1		a_2		a_3		
	y_{i1}	y_{i1}^2	y_{i2}	y_{i2}^2	y_{i3}	y_{i3}^2	
1	-1	1	0	0	-10	100	

2	0	0	2	4	-8	64	
3	4	16	4	16	-2	4	
4	5	25	6	36	0	0	
$P_j = \sum_{i=1}^{4} y_{ij}^2$		42		56		168	$\sum_{j=1}^{3} P_j = 266$
$T_j = \sum_{i=1}^{4} y_{ij}$	8		12		-20		$\sum_{j=1}^{3} T_j = 0$
T_j^2	64		144		400		$\sum_{j=1}^{3} T_j^2 = 608$

На основании результатов этой таблицы и формул (4.19) получаем:

$$W_{общ} = 266 - \frac{1}{12}\cdot 0^2 = 266; \quad W_A = \frac{1}{4}\cdot 608 - \frac{1}{12}\cdot 0^2 = 152; \quad W_{ост} = W_{общ} - W_A = 114$$

После этого реализуем формулы (4.12):

$$s_{общ}^2 = \frac{266}{11} \approx 24{,}2; \; s_A^2 = \frac{152}{2} = 76; \; s_{ост}^2 = \frac{114}{9} \approx 12{,}7$$

Теперь сравним по критерию Фишера-Снедекора факторную $s_A^2 = 76$ и остаточную $s_{ост}^2 = 12{,}7$ дисперсии. Для этого сначала по формуле (4.13) найдем экспериментальное значение $f_{эксп}$ случайной величины F:

$$f_{эксп} = \frac{76}{12{,}7} \approx 5.98$$

Учитывая теперь, что число степеней свободы числителя $k_A = m - 1 = 3 - 1 = 2,$ знаменателя $k_{ост} = N - m = 12 - 3 = 9,$ и имея в виду заданный уровень значимости α = 0,05, по таблице критических точек распределения Фишера-Снедекора находим:

$$f_{кр} = F(0{,}05;\ 2;\ 9) = 4{,}26$$

И так как оказалось, что $f_{эксп} > f_{кр}$, то гипотезу H_0 о несущественности (незначимости) фактора A для рассматриваемой случайной величины X *отвергаем*. То есть признаём, что фактор A существенно (значимо) влияет на величину X. Иначе говоря, различие групповых средних $(\bar{x}_1 = 154;\ \bar{x}_2 = 155;\ \bar{x}_3 = 147)$ в целом не случайно, а вызвано изменением уровней (a_1; a_2; a_3) влияющего на величину X фактора A.

А теперь уточним, какие из групповых средних различаются значимо, а какие нет. Сначала найдем модули всех возможных разностей групповых средних:

$$|\bar{x}_2 - \bar{x}_1| = 1; \quad |\bar{x}_3 - \bar{x}_1| = 7; \quad |\bar{x}_3 - \bar{x}_2| = 8$$

После этого по формуле (4.15) найдем наименьшую существенную разность (*НСР*) этих групповых средних. Так как согласно таблице критических точек распределения Стьюдента (см. таблицу 4 Приложения) $t_{кр} = t_{кр}$ (0,05; 9) = 2,26, то

$$НСР = t_{кр} \cdot \sqrt{\frac{2s_{ост}^2}{m}} = 2{,}26 \cdot \sqrt{\frac{2 \cdot 12{,}7}{3}} \approx 6{,}6$$

Как видим, лишь модули разностей $|\bar{x}_3 - \bar{x}_1| = 7$ и $|\bar{x}_3 - \bar{x}_2| = 8$ превышают наименьшую существенную разность *НСР* = 6,6, а модуль разности $|\bar{x}_2 - \bar{x}_1| = 1$ ее не превышает. То есть существенно (значимо) различаются лишь групповые средние $(\bar{x}_1; \bar{x}_3)$ и $(\bar{x}_2; \bar{x}_3)$.

Упражнения.

1. По выборочным данным, представленным в таблице

Номер измерения	Уровни фактора *A*						
	a_1	a_2	a_3	a_4	a_5	a_6	a_7
1	75	104	96	92	76	92	89
2	86	89	88	89	89	87	85
3		92	105		90	88	93
4		90	90		77	82	
5		81	91		75	90	
6						86	
Групповые средние $\bar{x}_j$	80,5	91,2	94	90,5	81,4	87,5	89

при уровнях значимости а) α = 0,05 и б) α = 0,01 проверить гипотезу H_0 о незначимости влияния фактора *A* на изучаемую величину *X*. Представить расчетную таблицу и результаты расчета для $y_{ij} = x_{ij} - 88$.

Ответ. В расчетной таблице должен быть осуществлен подсчет величин, фигурирующих в формулах (4.21) – (4.24):

Номер изме-рения	Уровни фактора *A*														$\sum_{j=1}^{7}$
	a_1		a_2		a_3		a_4		a_5		a_6		a_7		
	y_{i1}	y_{i1}^2	y_{i2}	y_{i2}^2	y_{i3}	y_{i3}^2	y_{i4}	y_{i4}^2	y_{i5}	y_{i5}^2	y_{i6}	y_{i6}^2	y_{i7}	y_{i7}^2	
1	-13	169	16	256	8	64	4	16	-12	144	4	16	1	1	
2	-2	4	1	1	0	0	1	1	1	1	-1	1	-3	9	
3			4	16	17	289			2	4	0	0	5	25	
4			2	4	2	4			-11	121	-6	36			
5			-7	49	3	9			-13	169	2	4			
6											-2	4			
P_j		173		326		366		17		439		61		35	1417
T_j	-15		16		30		5		-33		-3		3		3
$\frac{T_j^2}{n_j}$	112,5		51,2		180		12,5		217,8		1,5		3		578,5

$$N = 28;\ \ W_{общ} = 1417 - \frac{1}{28} \cdot 3^2 = 1416{,}7;$$

$$W_A = 578{,}5 - \frac{1}{28} \cdot 3^2 = 578{,}2;\ \ W_{ост} = 1416{,}7 - 578{,}2 = 838{,}5$$

$$s_A^2 = \frac{578{,}2}{7-1} = 96{,}4;\ \ s_{ост}^2 = \frac{838{,}5}{28-7} = 39{,}9;\ f_{эксп} = \frac{96{,}4}{39{,}9} = 2{,}42;$$

а) $f_{кр} = F(0{,}05; 6; 11) = 2{,}57;$ б) $f_{кр} = F(0{,}01; 6; 11) = 3{,}81$

При обоих уровнях значимости а) $\alpha = 0{,}05$ и б) $\alpha = 0{,}01$ нет оснований отвергать гипотезу H_0 о незначимости влияния фактора A на исследуемую величину X – гипотеза H_0 принимается.

2. Подтвердить итоговые результаты предыдущего упражнения, если положить $y_{ij} = x_{ij} - 90$.

§5. Корреляционно-регрессионный анализ

Корреляционный и регрессионный анализы являются смежными разделами математической статистики и предназначены для изучения по выборочным данным корреляционной зависимости (зависимости «в среднем») случайных величин друг от друга.

Напомним, что вопрос о корреляционной зависимости между двумя случайными величинами X и Y мы уже рассматривали в части I «Теория вероятностей» (§6, глава 2). Там же было дано понятие и о корреляционной зависимости между несколькими случайными величинами. Однако все это мы рассматривали там с общих теоретических позиций (на языке теории вероятностей). А теперь рассмотрим этот же вопрос с практической точки зрения (на языке математической статистики).

5.1. Парная корреляция.

Начнем с наиболее простого случая. Пусть с помощью выборки объемом N изучались объекты генеральной совокупности, каждый из которых характеризуется двумя (парой) *количественных* признаков X и Y. Подчеркиваем: *количественных*, ибо если эти признаки качественные, то исследование их взаимозависимости – это задача дисперсионного анализа. Например, если объектами генеральной совокупности являются изделия некоторого массового производства, то их количественными признаками X и Y могут быть, например, каких-то два их контролируемых размера; или размер и вес; или затраты на производство и выручка от продажи, и т. д. Выборочные данные оформляют в виде таблицы (5.1):

y_j \ x_i	x_1	x_2	…	x_n	$m_j = \sum_{i=1}^{n} n_{ij}$
y_1	$\boldsymbol{n_{11}}$	$\boldsymbol{n_{21}}$	…	$\boldsymbol{n_{n1}}$	$\boldsymbol{m_1}$
y_2	$\boldsymbol{n_{12}}$	$\boldsymbol{n_{22}}$	…	$\boldsymbol{n_{n2}}$	$\boldsymbol{m_2}$
…	…	…	…	…	…
y_m	$\boldsymbol{n_{1m}}$	$\boldsymbol{n_{2m}}$	…	$\boldsymbol{n_{nm}}$	$\boldsymbol{m_m}$
$n_i = \sum_{j=1}^{m} n_{ij}$	$\boldsymbol{n_1}$	$\boldsymbol{n_2}$	…	$\boldsymbol{n_n}$	$\sum_{i=1}^{n} n_i = \sum_{j=1}^{m} m_j = N$

(5.1)

Эта таблица называется *корреляционной*. Из нее видно, что признак X у объектов выборки принимал значения $x_i = (x_1 ; x_2 ; \dots x_n)$, а признак Y – значения $y_j = (y_1 ; y_2 ; \dots y_m)$, причем пара значений $(x_1; y_1)$ встретилась у n_{11} объектов, пара

$(x_2; y_1)$ – у n_{21} объектов, и т.д. Числа $n_i = (n_1 ; n_2 ; \dots \ n_n)$ определяют общее количество объектов выборки со значениями признака X, равными $(x_1 ; x_2 ; \dots \ x_n)$ соответственно, а числа $(m_1; m_2; \dots \ m_m)$ – общее количество объектов со значениями $(y_1 ; y_2 ; \dots ; y_m)$ признака Y соответственно. При этом ясно, что

$$n_i = \sum_{j=1}^{m} n_{ij}; \ \ m_j = \sum_{i=1}^{n} n_{ij}; \ \ \sum_{i=1}^{n} n_i = \sum_{j=1}^{m} m_j = \sum_{i=1}^{n}\sum_{i=1}^{m} n_{ij} = N \qquad (5.2)$$

Корреляционная таблица (5.1) фактически является статистическим распределением выборки при исследовании двумерного (двухпараметрического) признака $Z=(X, Y)$ объектов генеральной совокупности.

Данные корреляционной таблицы для наглядности удобно изобразить и в виде так называемого *корреляционного поля*. Корреляционное поле – это нанесенное на плоскость *хоу* множество всех N экспериментальных точек с координатами $(x_i; y_j)$ с учетом их кратности n_{ij}. Это значит, что при построении корреляционного поля нужно показать, что в точке $(x_i; y_j)$ плоскости *хоу* содержится не одна, а n_{ij} точек. Чтобы это было видно на корреляционном поле, нужно эти точки немного отделить друг от друга. Они тогда образуют видимую компактную кучку из n_{ij} точек, окружающих точку $(x_i; y_j)$ (рис.3.11).

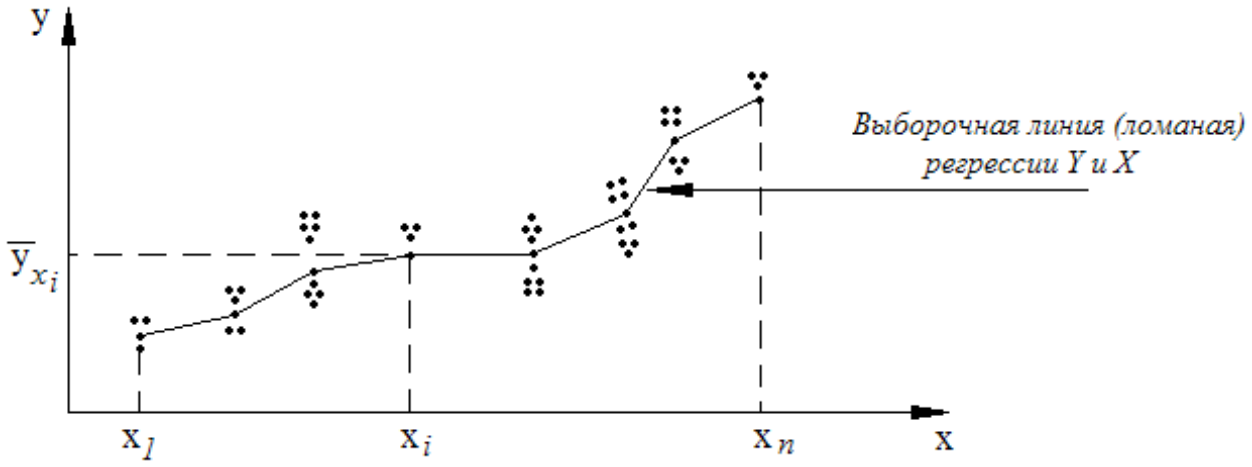


Рис. 3.11

Как мы знаем из части I (§6, глава 2), корреляционную зависимость (зависимость в среднем) величины Y от величины X характеризует так называемое *уравнение регрессии* $\bar{y}_x = f(x)$ величины Y на величину X. А график этого уравнения называется *линией регрессии* Y на X. Примерный (оценочный) вид этой линии мы получим, если по данным корреляционной таблицы (5.1) для каждого выборочного значения $X=x_i$ найдем среднее значение $\bar{y}_{x_i}$ величины Y

$$\bar{y}_{x_i} = \frac{\sum_{j=1}^{m} y_j n_{ij}}{n_i} \quad (i = 1, 2, ..., n) \tag{5.3}$$

и затем нанесем на корреляционное поле ломаную, соединяющую точки с координатами $(x_i; \bar{y}_{x_i})$ $(i = 1, 2, ... n)$ – рис. 3.11. Эта ломаная называется *выборочной линией регрессии Y на X*. Построив эту линию, визуально можем оценить и наличие, и тесноту корреляционной зависимости Y от X: чем меньше разброс точек корреляционного поля вокруг выборочной линии регрессии Y на X, тем эта зависимость теснее.

Задачи корреляционно-регрессионного анализа в математической статистике аналогичны тем, что были поставлены в теории вероятностей (§6, глава 2). Этих задач две.

Задача 1. Дать оценку истинного (генерального) уравнения регрессии $\bar{y}_x = f(x)$ величины Y на величину X. А следовательно, дать и оценку истинной (генеральной) линии регрессии Y на X. Мы говорим лишь об оценке, ибо найти точно по выборочным данным и генеральное уравнение регрессии, и генеральную линию регрессии, очевидно, невозможно.

Задача 2. Оценить степень тесноты корреляционной зависимости Y от X.

Начнем с рассмотрения первой из этих задач. Идея ее решения состоит в подборе возможно простого уравнения

$$\bar{y}_x^* = f^*(x), \tag{5.4}$$

график которого тем не менее будет достаточно близким к выборочной линии (ломаной) регрессии. То есть будет достаточно хорошим приближением этой ломаной, сглаживающим эту ломаную. Такое уравнение называется *выравнивающим* (или *сглаживающим*) выборочным уравнением регрессии. Именно оно и принимается за оценку истинного (генерального) уравнения регрессии

$\bar{y}_x = f(x)$, то есть за решение первой задачи корреляционно-регрессионного анализа.

Отметим, однако, что два указанных выше требования: а) простота сглаживающего уравнения регрессии (5.4), а значит, и простота сглаживающей линии регрессии, и б) близость сглаживающей линии регрессии к реальной ломаной регрессии, вообще говоря, противоречивы, ибо повысить указанную близость можно лишь за счет усложнения сглаживающего уравнения. Поэтому на практике стараются добиться некоей золотой середины: и чтобы сглаживающее уравнение регрессии (5.4) было не слишком сложным, и чтобы соответствующая ему сглаживающая линия регрессии тем не менее в целом была достаточно близкой к выборочной ломаной регрессии. Как из многих возможных вариантов выбрать лучший (найти эту золотую середину) – об этом будет сказано ниже.

В качестве наиболее простых форм сглаживающего уравнения (5.4) чаще всего принимаются следующие его формы:

1) Уравнение прямой

$$\bar{y}_x^* = kx + b \qquad (5.5)$$

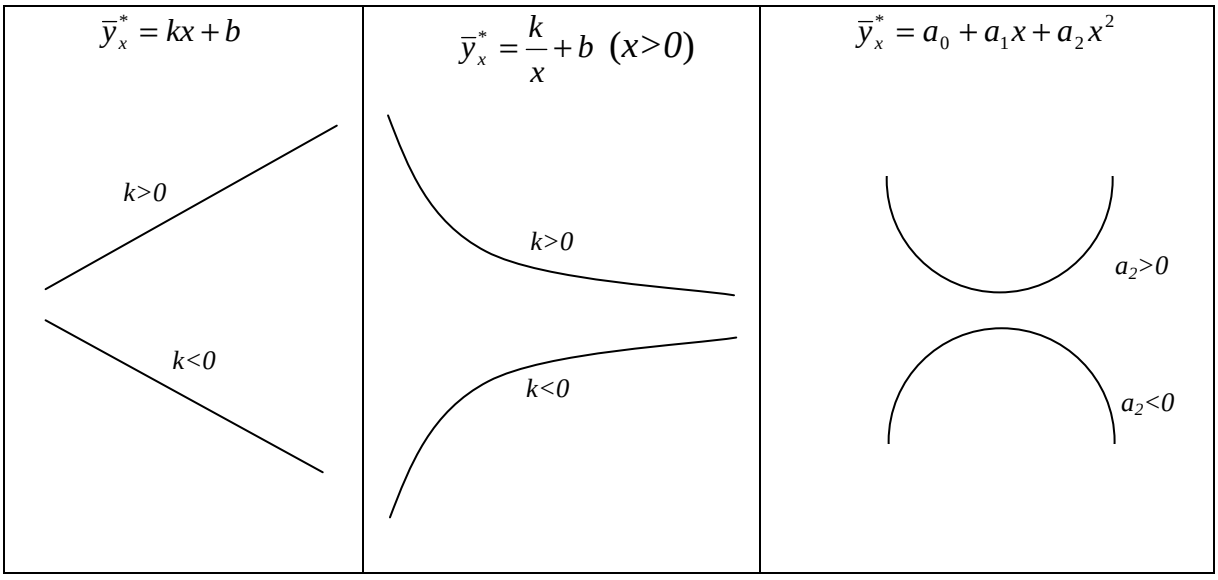
2) Уравнение гиперболы

$$\bar{y}_x^* = \frac{k}{x} + b \qquad (5.6)$$

3) Уравнение параболы

$$\bar{y}_x^* = a_0 + a_1 x + a_2 x^2 \qquad (5.7)$$

Напомним, что эти линии в принципе имеют следующий вид:



Линейная зависимость (5.5) $\bar{y}_x^*$ наиболее проста по форме, ее параметры k и b легко интерпретируются. В частности, коэффициент k указывает, на сколько в среднем увеличится (при $k>0$) или уменьшится (при $k<0$) величина Y, если значение x величины X увеличится на единицу. А параметр b указывает среднее значение величины Y при $x = 0$. Благодаря этим преимуществам, а также благодаря простоте вычисления параметров k и b уравнение (5.5) используется в ка-

честве приближенного (сглаживающего) уравнения регрессии даже в тех случаях, когда более логичным представляется использование уравнение кривой.

Гиперболы (5.6) – это либо монотонно возрастающая, либо монотонно убывающая кривая. Однако, в отличие от прямой, рост или убывание гиперболы имеет тенденцию к затуханию, практически сходя на нет при больших значениях x величины X. Параметр b при этом представляет собой предельное значение $\bar{y}_x^*$ при $x \to \infty$. Убывающей гиперболой, например, хорошо выражается зависимость себестоимости продукции растениеводства от урожайности, а возрастающей – зависимость продуктивности животных от расхода кормов.

Парабола (5.7) имеет вершину или впадину и применяется для приближенного описания зависимостей, в которых с изменением x убывание $\bar{y}_x^*$ может меняться на возрастание, и наоборот. Примером параболической зависимости является, например, зависимость средней урожайности от дозы удобрений, когда начальные дозы удобрений приводят к значительному увеличению урожая, последующие – к постепенно уменьшающимся прибавкам, а чрезмерные – к снижению урожая и даже к его гибели. Иногда график параболы используется только частично (только восходящая или только нисходящая ветвь). Параметры параболы, за исключением a_0, интерпретировать сложно. Ну, а параметр a_0 является, очевидно, оценкой среднего значения величины Y при $x = 0$.

Подходящую форму сглаживающего уравнения регрессии $\bar{y}_x^* = f^*(x)$ выбирают, исходя из общих теоретических соображений или, что чаще, по виду корреляционного поля (рис. 3.11). При этом, как уже говорилось выше, наиболее часто уравнение регрессии выбирается в одной из форм (5.5) – (5.7). А для нахождения параметров выбранного уравнения используется универсальный стандартный метод, называемый *методом наименьших квадратов*.

Суть этого метода в следующем. Из корреляционной таблицы (5.1) для каждого $X = x_i$ по формуле (5.3) находим выборочное среднее значение $\bar{y}_{x_i}$ величины Y. Далее, для выбранной формы уравнения регрессии записываем сглаживающие средние $\bar{y}_{x_i}^* = f^*(x_i)$. В итоге получаем следующую таблицу соответствий экспериментальных $\bar{y}_{x_i}$ и сглаживающих $\bar{y}_{x_i}^*$ средних значений величины X, принимающей значения x_i с частотами n_i $(i = 1, 2, \ldots n)$:

x_i	x_1	x_2	…	x_n
n_i	n_1	n_2	…	n_n
$\bar{y}_{x_i}$	$\bar{y}_{x_1}$	$\bar{y}_{x_2}$	…	$\bar{y}_{x_n}$
$\bar{y}_{x_i}^* = f^*(x_i)$	$f^*(x_1)$	$f^*(x_2)$	…	$f^*(x_n)$

(5.8)

Параметры выбранного сглаживающего уравнения регрессии $\bar{y}_x^* = f^*(x)$ считаются наилучшими, если они обеспечивают минимально возможные отклонения выборочных средних $\bar{y}_{x_i}$ от подсчитанных по уравнению регрессии сглаживающих средних $\bar{y}_{x_i}^* = f^*(x_i)$. В методе наименьших квадратов за меру отклонений $\bar{y}_{x_i}$ от $\bar{y}_{x_i}^* = f^*(x_i)$ принимается сумма квадратов их разностей. При этом должно

быть учтено, что в образовании каждого $\bar{y}_{x_i}$ участвуют n_i точек корреляционного поля. То есть в средней $\bar{y}_{x_i}$ как бы сливаются n_i значений величины *Y*. *С* учетом сказанного указанная сумма принимает вид:

$$Q = \sum_{i=1}^{n} (\bar{y}_{x_i} - \bar{y}^*_{x_i})^2 n_i = \sum_{i=1}^{n} [\bar{y}_{x_i} - f^*(x_i)]^2 n_i \to \min \qquad (5.9)$$

В эту сумму входят параметры выбранной функции $f^*(x)$. И эти параметры подбираются таким образом, чтобы сумма *Q* была минимально возможной. А это – стандартная задача математического анализа об исследовании функции нескольких переменных на экстремум (минимум или максимум), где *Q* – функция, а ее переменные – параметры функции $f^*(x)$. Решая эту задачу, находим наилучшие параметры функции $f^*(x)$, а вместе с ними получаем и искомое наилучшее (для выбранной формы) сглаживающее уравнение регрессии (5.4), являющееся оценкой истинного (генерального) уравнения регрессии $\bar{y}_x = f(x)$.

Заметим что и реальные выборочные средние $\bar{y}_{x_i}$, и сглаживающие средние $\bar{y}^*_{x_i}$ имеют одно и то же среднее значение – общую выборочную среднюю $\bar{y}$ величины *Y*. То есть

$$\frac{\sum_{i=1}^{n} \bar{y}_{x_i} \cdot n_i}{N} = \bar{y}; \quad \frac{\sum_{i=1}^{n} \bar{y}^*_{x_i} \cdot n_i}{N} = \bar{y} \qquad (5.10)$$

Первое из этих равенств следует из выражения (5.3):

$$\frac{\sum_{i=1}^{n} \bar{y}_{x_i} \cdot n_i}{N} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_j n_{ij}}{N} = \frac{\sum_{j=1}^{m} y_j \sum_{i=1}^{n} n_{ij}}{N} = \frac{\sum_{j=1}^{m} y_j m_j}{N} = \bar{y}$$

А второе равенство получим, рассудив следующим образом. Так как сглаживающая кривая $\bar{y}^*_x = f^*(x)$ получена методом наименьших квадратов в процессе минимизации суммы (5.9), то она наилучшим способом вписывается в выборочную линию регрессии, то есть имеет от нее в целом минимальное отклонение. Поэтому если эту сглаживающую линию поднять или опустить, то есть если заменить на линию $\bar{y}^*_x = f^*(x) + C$, где *C* – некоторая константа, то взамен функции (5.9) получим функцию

$$Q = Q(C) = \sum_{i=1}^{n} (\bar{y}_{x_i} - f^*(x_i) - C)^2 n_i,$$

значения которой при всех *C* больше величины (5.9). А свое наименьшее значение (экстремум) функция *Q(C)* должна иметь при *C = 0*. Но это значит, что

$$Q' = Q'(C) = -2\sum_{i=1}^{n} (\bar{y}_{x_i} - f^*(x_i) - C) n_i = 0 \quad \text{при} \quad C = 0$$

Отсюда следует:

$$\sum_{i=1}^{n} \bar{y}_{x_i} n_i - \sum_{i=1}^{n} f^*(x_i) n_i = 0; \; N\bar{y} - \sum_{i=1}^{n} f^*(x_i) n_i = 0; \; \frac{\sum_{i=1}^{n} f^*(x_i) n_i}{N} = \bar{y}.$$

То есть и второе равенство (5.10) доказано.

Равенства (5.10) можно использовать для контроля правильности подсчета и реальных выборочных средних $\bar{y}_{x_i}$, и сглаживающих выборочных средних $\bar{y}^*_{x_i}$

Рассмотрим, в частности, приложение метода наименьших квадратов к случаю, когда теоретические соображения или конфигурация корреляционного поля позволяют в качестве сглаживающего уравнения регрессии $\bar{y}^*_x = f^*(x)$ использовать уравнение прямой (5.5). Сформируем для этого случая сумму Q:

$$Q = Q(k;b) = \sum_{i=1}^{n} (\bar{y}_{x_i} - kx_i - b)^2 n_i \to \min$$

Вспоминая, что необходимым условием минимума (или максимума) функции нескольких переменных является равенство нулю всех ее частных производных первого порядка, получим следующую систему уравнений (так называемую *нормальную систему*) для нахождения параметров k и b функции (5.5):

$$\begin{cases} Q'_k = -2\sum_{i=1}^{n} (\bar{y}_{x_i} - kx_i - b) x_i n_i = 0 \\ Q'_b = -2\sum_{i=1}^{n} (\bar{y}_{x_i} - kx_i - b) n_i = 0 \end{cases} \tag{5.11}$$

Сократив на (-2) и разделив затем обе части каждого уравнения на N, получим:

$$\begin{cases} k \cdot \dfrac{\sum_{i=1}^{n} x_i^2 n_i}{N} + b \cdot \dfrac{\sum_{i=1}^{n} x_i n_i}{N} = \dfrac{\sum_{i=1}^{n} \bar{y}_{x_i} x_i n_i}{N} \\ k \cdot \dfrac{\sum_{i=1}^{n} x_i n_i}{N} + b \cdot \dfrac{\sum_{i=1}^{n} n_i}{N} = \dfrac{\sum_{i=1}^{n} \bar{y}_{x_i} n_i}{N} \end{cases} \tag{5.12}$$

Учтем, что

$$\begin{gathered} \frac{\sum_{i=1}^{n} x_i^2 n_i}{N} = \overline{x^2}; \quad \frac{\sum_{i=1}^{n} x_i n_i}{N} = \bar{x}; \quad \frac{\sum_{i=1}^{n} n_i}{N} = \frac{N}{N} = 1; \\ \frac{\sum_{i=1}^{n} \bar{y}_{x_i} \cdot x_i n_i}{N} = \left| см.(5.3) \right| = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j n_{ij}}{N} = \overline{xy}; \\ \frac{\sum_{i=1}^{n} \bar{y}_{x_i} \cdot n_i}{N} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_j \cdot n_{ij}}{N} = \frac{\sum_{j=1}^{m} y_j \sum_{i=1}^{n} n_{ij}}{N} = \frac{\sum_{j=1}^{m} y_j m_j}{N} = \bar{y} \end{gathered} \tag{5.13}$$

Тогда система (5.12) примет вид:

$$\begin{cases} k\overline{x^2} + b\bar{x} = \overline{xy} \\ k\bar{x} + b = \bar{y} \end{cases} \tag{5.14}$$

Решая ее, находим k и b:

$$k = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{\mu_{xy}}{\sigma_x^2}; \quad b = \bar{y} - k\bar{x} \tag{5.15}$$

Подставляя найденные значения k и b в уравнение (5.5), получим искомое сглаживающее линейное уравнение регрессии Y на X:

$$\bar{y}^*_x - \bar{y} = k(x - \bar{x}) \tag{5.16}$$

Отметим, что по своей форме оно точно такое же, как и уравнение (6.22) (часть I, глава 2, §6), полученное нами ранее в теории вероятностей. Совпадают эти уравнения не только по форме, но и по существу.

Действительно, в уравнении (6.22), согласно (6.23), (6.7) и (6.5) главы 2 фигурируют

$$\bar{y} = M(Y);\ \ \bar{x} = M(X);\ \ k = \frac{\sigma(Y)}{\sigma(X)}\rho(X,Y) = \\ = \frac{\sigma(Y)}{\sigma(X)} \cdot \frac{\mu(X,Y)}{\sigma(X)\sigma(Y)} = \frac{M(XY) - M(X)M(Y)}{\sigma^2(X)} \tag{5.17}$$

Сравнивая эти выражения $(\bar{x}; \bar{y}; k)$ с теми выражениями (5.13) и (5.15), что используются в только что полученном уравнении (5.16), видим, что разница между ними состоит лишь в том, что выражения (5.17) дают истинные (генеральные) значения параметров $(\bar{x}; \bar{y}; k)$, а выражения (5.13) и (5.15) дают выборочные значения этих же параметров.

Кстати, выборочный коэффициент линейной корреляции ρ_{xy}, являющийся выборочным значением истинного (генерального) коэффициента линейной корреляции $\rho(X, Y)$ случайных величин X и Y, находится по формуле:

$$\rho_{xy} = \frac{\mu_{xy}}{\sigma_x \sigma_y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y} \tag{5.18}$$

Или, что то же самое, по формуле, связывающей его с параметром k уравнения (5.16):

$$\rho_{xy} = \frac{\sigma_x}{\sigma_y} k; \quad k = \frac{\sigma_y}{\sigma_x} \rho_{xy} \tag{5.19}$$

При этом

$$\sigma_x = \sqrt{\overline{x^2} - (\bar{x})^2};\ \ \sigma_y = \sqrt{\overline{y^2} - (\bar{y})^2} \tag{5.20}$$

– выборочные значения среднеквадратических отклонений $\sigma(X)$ и $\sigma(Y)$ величин X и Y соответственно, а

$$\mu_{xy} = \overline{xy} - \bar{x} \cdot \bar{y} \tag{5.21}$$

– выборочное значение корреляционного момента (ковариации) $\mu(X,Y)$ величин X и Y.

Значение ρ_{xy} выборочного коэффициента линейной корреляции используется, в соответствии с §6 (часть I, глава 2), для оценки степени тесноты и линейности корреляционной связи между случайными величинами X и Y. Степень же тесноты *любой* (а не только линейной) корреляционной зависимости Y от X определяет, как мы знаем (§ 6, глава 2, часть I) *корреляционное отношение* $\eta(Y, X)$, чье выборочное значение η_{yx}, с учетом формулы (6.40) главы 2, находится по формуле:

$$\eta_{yx} = \frac{\sigma_{\bar{y}_x}}{\sigma_y} = \sqrt{\frac{\sum_{i=1}^{n} (\bar{y}_{x_i} - \bar{y})^2 n_i}{\sum_{j=1}^{m} (y_j - \bar{y})^2 m_j}} = \sqrt{1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} (y_j - \bar{y}_{x_i})^2 n_{ij}}{\sum_{j=1}^{m} (y_j - \bar{y})^2 m_j}} \tag{5.22}$$

Оно показывает долю, которую составляет $\sigma_{\bar{y}_x}$ по отношение к σ_y. То есть показывает, какую часть составляет средний разброс выборочных средних $\bar{y}_{x_i}$ вокруг общей средней $\bar{y}$ величины *Y* в выборке по отношению к среднему разбросу значений y_j величины *Y* в выборке вокруг той же общей средней $\bar{y}$.

Минимально возможное значение $(\eta_{yx})_{min} = 0$ указывает на то, что $\sigma_{\bar{y}_x} = 0$. То есть что разброс выборочных средних $\bar{y}_{x_i}$ вокруг общей средней $\bar{y}$ отсутствует. А это значит, что $\bar{y}_{x_i} = \bar{y}$ для всех x_i $(i = 1, 2, \ldots, n)$. Выборочная ломаная регрессии (рис. 3.11) становится в этом случае горизонтальной прямой. Выборочные средние $\bar{y}_x$ не меняются с изменением x, а значит, они от x не зависят. Тогда выборка свидетельствует о том, что, скорее всего, величина *Y* корреляционно (в среднем) не зависит от величины *X*. Мы говорим «скорее всего», потому что никаких окончательных выводов относительно генеральной совокупности исследование выборки дать не может - другая выборка может привести и к другим выводам.

Максимально же возможное значение $(\eta_{yx})_{max} = 1$ указывает на отсутствие разброса значений y_j величины *Y* относительно их средних значений $\bar{y}_{x_i}$ для каждого x_i $(i = 1, 2, \ldots, n)$. Это означает отсутствие разброса точек корреляционного поля вокруг выборочной линии регрессии (рис. 3.11). То есть в этом случае каждому x_i соответствует лишь одно значение $\bar{y}_{x_i} = y_i$. Иначе говоря, в этом случае выборочные данные свидетельствуют в пользу того, что величина *Y* жестко (функционально) зависит от величины *X*.

Если в формуле (5.22) заменить реальные условные средние $\bar{y}_{x_i}$ на сглаживающие условные средние $\bar{y}^*_{x_i} = f^*(x_i)$ и возвести полученное выражение в квадрат, то получим так называемый *выборочный коэффициент детерминации* d_{yx}:

$$d_{yx} = \frac{\sigma^2_{\bar{y}^*_x}}{\sigma^2_y} = \frac{\sum_{i=1}^{n} (\bar{y}^*_{x_i} - \bar{y})^2 n_i}{\sum_{j=1}^{m} (y_i - \bar{y})^2 m_j} \qquad (5.23)$$

Он показывает долю, которую составляет дисперсия $\sigma^2_{\bar{y}^*_x}$ сглаживающих средних $\bar{y}^*_{x_i}$ по отношению к общей дисперсии σ^2_y выборочных значений y_j исследуемого признака *Y*. То есть он показывает долю общего изменения (вариации) величины *Y*, объясняемую подобранным сглаживающим уравнением регрессии $\bar{y}^*_x = f^*(x)$. Его обычно выражают в процентах.

Выборочное корреляционное отношение η_{yx} не зависит, очевидно, от формы выбранного сглаживающего уравнения регрессии $\bar{y}^*_x = f^*(x)$, ибо его величина определяется исключительно выборочными данными. А вот выборочный коэффициент детерминации d_{yx} от этой формы зависит. Как можно доказать,

$$(d_{yx})_{\max} = \eta^2_{yx} \qquad (5.24)$$

И чем больше d_{yx} (чем ближе он к η^2_{yx}), тем лучше построенное сглаживающее уравнение регрессии объясняет вариацию (изменение) зависимой величины *Y*. А следовательно, тем удачнее это уравнение построено. При $d_{yx} = \eta^2_{yx}$ сглаживающая линия регрессии $\bar{y}^*_x = f^*(x)$ точно пройдет через все экспериментальные точки $(x_i; \bar{y}_{x_i})(i = 1, 2, ..., n)$ корреляционного поля, то есть через все узлы ломаной, изображенной на рис. 3.11. Это – идеальный вариант для сглаживающей линии. Правда, уравнение $\bar{y}^*_x = f^*(x)$ такой идеальной сглаживающей линии при большом числе узлов выборочной ломаной линии регрессии, как правило, слишком сложно. Поэтому на практике идут на существенное упрощение подбираемого сглаживающего уравнения регрессии, жертвуя при этом неизбежным снижением коэффициента детерминации. Если же усложнение сглаживающего уравнения регрессии не пугает, то среди различных подобранных сглаживающих уравнений лучшим считается то, которое обеспечивает наибольший коэффициент детерминации. На вычислительных машинах, кстати, и построение сглаживающих уравнений регрессии в различных формах, и выбор из них наилучшего (по коэффициенту детерминации) делается по специальной стандартной программе.

Кстати, если сглаживающее уравнение регрессии строить в линейной (наиболее простой) форме (5.5), то будем иметь:

$$d_{yx} = \rho^2_{yx} = \rho^2_{xy} \tag{5.25}$$

Действительно, для этого случая на основании (5.16) имеем:

$$d_{yx} = \frac{\sum_{i=1}^{n}[k(x_i - \bar{x}) + \bar{y} - \bar{y}]^2 n_i}{\sum_{j=1}^{m}(y_j - \bar{y})^2 m_j} = \frac{k^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 n_i}{\sum_{j=1}^{m}(y_j - \bar{y})^2 m_j} = \frac{k^2 N \sigma_x^2}{N \sigma_y^2} = \left(k \frac{\sigma_x}{\sigma_y}\right)^2 = \rho^2_{xy} \tag{5.26}$$

Если сглаживающее уравнение регрессии $\bar{y}^*_x = f^*(x)$ строится в нелинейной и достаточно сложной форме, то такое построение трудно произвести вручную, и его лучше поручить машине. Если же возможности воспользоваться машиной (персональным компьютером) нет, то для несложных нелинейных случаев, в частности для уравнений вида (5.6) и (5.7) все можно сделать и вручную. Делается это с помощью метода наименьших квадратов совершенно аналогично тому, как это было проделано выше при построении сглаживающего линейного уравнения (5.5), принявшего итоговую форму (5.16).

Пусть, например, сглаживающее уравнение регрессии строится в гиперболической форме (5.6). Заметим, что такое уравнение – это фактически уравнение (5.5), если в последнем заменить *x* на $\frac{1}{x}$. Поэтому для параметров *k* и *b* уравнения (5.6) мы можем воспользоваться формулами (5.15), заменив в них *x* на $\frac{1}{x}$:

$$k = \frac{\overline{\left(\frac{y}{x}\right)} - \overline{\left(\frac{1}{x}\right)} \cdot \bar{y}}{\overline{\left(\frac{1}{x^2}\right)} - \left[\overline{\left(\frac{1}{x}\right)}\right]^2}; \quad b = \bar{y} - k \cdot \overline{\left(\frac{1}{x}\right)} \tag{5.27}$$

Здесь

$$\overline{\left(\frac{1}{x}\right)} = \frac{\sum_{i=1}^{n} \frac{1}{x_i} n_i}{N}; \quad \overline{\left(\frac{1}{x^2}\right)} = \frac{\sum_{i=1}^{n} \frac{1}{x_i^2} n_i}{N};$$
$$\bar{y} = \frac{\sum_{j=1}^{m} y_i m_j}{N}; \quad \overline{\left(\frac{y}{x}\right)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{y_j}{x_i} n_{ij}}{N} \tag{5.28}$$

Наконец, рассмотрим еще построение сглаживающего уравнения регрессии в параболической форме (5.7). Формируя для этого случая сумму (5.9)

$$Q = Q(a_0; a_1; a_2) = \sum_{i=1}^{n} \left[\bar{y}_{x_i} - (a_0 + a_1 x_i + a_2 x_i^2)\right]^2 n_i \to \min \tag{5.29}$$

и отыскивая ее минимум, приходим к аналогичной (5.12) нормальной системе для нахождения параметров *(a_0; a_1; a_2)* сглаживающего уравнения регрессии (5.7). Приведём окончательный вид этой системы:

$$\begin{cases} a_0 N + a_1 \sum_{i=1}^{n} x_i n_i + a_2 \sum_{i=1}^{n} x_i^2 n_i = \sum_{i=1}^{n} \bar{y}_{x_i} n_i \\ a_0 \sum_{i=1}^{n} x_i n_i + a_1 \sum_{i=1}^{n} x_i^2 n_i + a_2 \sum_{i=1}^{n} x_i^3 n_i = \sum_{i=1}^{n} \bar{y}_{x_i} x_i n_i \\ a_0 \sum_{i=1}^{n} x_i^2 n_i + a_1 \sum_{i=1}^{n} x_i^3 n_i + a_2 \sum_{i=1}^{n} x_i^4 n_i = \sum_{i=1}^{n} \bar{y}_{x_i} x_i^2 n_i \end{cases} \tag{5.30}$$

Решая эту систему, находим *(a_0; a_1; a_2)*, а вместе с ними – и искомое параболическое уравнение регрессии (5.7).

Построив сглаживающее уравнение регрессии $\bar{y}_x^* = f^*(x)$ в нескольких различных формах (линейное, гиперболическое, параболическое и т.д.) и выбрав из них лучшее, мы тем не менее еще не можем быть уверены в пригодности такого уравнения для приближения им истинного (генерального) уравнения регрессии $\bar{y}_x = f(x)$. Дело в том, что построенная сглаживающая линия регрессии может на некоторых своих участках выходить за пределы корреляционного поля, особенно если полоса точек этого поля узкая (корреляционная зависимость *Y* от *X* близка к функциональной). Тогда на этих участках сглаживающая линия не будет соответствовать корреляционному полю (будем неадекватна ему), а значит, будет неадекватно ему и уравнение $\bar{y}_x^* = f^*(x)$ этой сглаживающей линии. Такое уравнение не может быть использовано для всех *x*, входящих в корреляционную таблицу, а значит, его применение чревато грубыми ошибками, если им мы будем приближать генеральное уравнение регрессии $\bar{y}_x = f(x)$. В этом случае уравнение $\bar{y}_x^* = f^*(x)$ считается *неадекватным выборочным данным* и применяться не должно.

Таким образом, после построения сглаживающего уравнения регрессии $\bar{y}_x^* = f^*(x)$ его еще нужно проверить на адекватность выборочным данным. Адекватность этого уравнения будет тем выше, чем лучше будет соответствующая ему сглаживающая линия регрессии вписываться в полосу точек корреляционного поля. То есть чем меньше будет разброс этих точек вокруг указанной линии.

Оценим величину этого разброса. Для этого подсчитаем сумму квадратов отклонений ординат y_i всех точек корреляционного поля от сглаживающей линии регрессии $\bar{y}_x^* = f^*(x)$:

$$Q_0 = \sum_{i=1}^{n}\sum_{j=1}^{m}(y_j - \bar{y}_{x_i}^*)^2 n_{ij} \qquad (5.31)$$

Проведем следующее преобразование этой суммы:

$$Q_0 = \sum_{i=1}^{n}\sum_{j=1}^{m}(y_j - \bar{y}_{x_i}^*)^2 n_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{m}\left[\left(y_j - \bar{y}_{x_i}\right) + \left(\bar{y}_{x_i} - \bar{y}_{x_i}^*\right)\right]^2 n_{ij} =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m}(y_j - \bar{y}_{x_i})^2 n_{ij} + 2\sum_{i=1}^{n}\sum_{j=1}^{m}(y_j - \bar{y}_{x_i})(\bar{y}_{x_i} - \bar{y}_{x_i}^*)n_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{m}(\bar{y}_{x_i} - \bar{y}_{x_i}^*)^2 n_{ij} =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m}(y_j - \bar{y}_{x_i})^2 n_{ij} + 2\sum_{i=1}^{n}(\bar{y}_{x_i} - \bar{y}_{x_i}^*)\sum_{j=1}^{m}(y_j - \bar{y}_{x_i})n_{ij} + \sum_{i=1}^{n}(\bar{y}_{x_i} - \bar{y}_{x_i}^*)^2\sum_{j=1}^{m}n_{ij}$$

Учитывая, согласно (5.2) и (5.3), что

$$\sum_{j=1}^{m}n_{ij} = n_i;\ \ \sum_{j=1}^{m}(y_j - \bar{y}_{x_i})n_{ij} = \sum_{j=1}^{m}y_j n_{ij} - \bar{y}_{x_i}\sum_{j=1}^{m}n_{ij} = \bar{y}_{x_i}n_i - \bar{y}_{x_i}n_i = 0,$$

получим окончательно:

$$Q_0 = Q_{повт} + Q_{адекв} \qquad (5.32)$$

Здесь

$$Q_{повт} = \sum_{i=1}^{n}\sum_{j=1}^{m}(y_j - \bar{y}_{x_i})^2 n_{ij};\quad \mathrm{Q_{адекв}} = \sum_{i=1}^{n}(\bar{y}_{x_i}^* - \bar{y}_{x_i})^2 n_i \qquad (5.33)$$

Сумма $Q_{повт}$ характеризует разброс выборочных значений y_j вокруг выборочных средних $\bar{y}_{x_i}$ при проведении повторных опытов для различных x_i, поэтому она так и обозначена: $Q_{повт}$. Она определяет степень влияния на величину Y различных неучтенных факторов (помех), не связанных с величиной X. Кстати, сумма $Q_{повт}$ не зависит, очевидно, от сглаживающего уравнения регрессии $\bar{y}_x^* = f^*(x)$, так что уменьшить или увеличить ее нельзя – она определяется исключительно выборочными данными. А вот вторая сумма $Q_{адекв}$ зависит от вида уравнения $\bar{y}_x^* = f^*(x)$. Она характеризует меру отклонений сглаживающих средних $\bar{y}_{x_i}^*$ от реальных (выборочных) средний $\bar{y}_{x_i}$. И чем эта сумма меньше, тем более адекватным будет, очевидно, сглаживающее уравнение регрессии $\bar{y}_x^* = f^*(x)$. Поэтому эта сумма так и обозначена: $Q_{адекв}$. Кстати, сумма $Q_{адекв}$ – это как раз та сумма Q (см. (5.9)), на минимизации которой основано построение сглаживающего уравнения регрессии.

Естественно, что если $Q_{адекв} = 0$, то сглаживающее уравнение регрессии полностью адекватно выборочным данным (корреляционной таблице (5.1)). А если $Q_{адекв} \neq 0$, что обычно и бывает на самом деле, то сравнивая $Q_{адекв}$ с $Q_{повт}$

выясняют, достаточно ли мала сумма $Q_{адекв}$, чтобы для данного уровня значимости α можно было бы принять нулевую гипотезу H_0 об адекватности сглаживающего уравнения регрессии $\bar{y}_x^* = f^*(x)$ при альтернативной гипотезе H_1 об его неадекватности. Это можно сделать по критерию Фишера-Снедекора, если заведомо известно (или подтверждено экспериментально), что зависимая случайная величина Y при любом значении x величины X распределена по нормальному закону и имеет не зависящую от x постоянную дисперсию.

Для этого делением сумм $Q_{повт}$ и $Q_{адекв.}$ на соответствующие им числа степеней свободы $k_{повт} = N - n$ и $k_{адекв.} = n - q$ (q – число коэффициентов уравнения регрессии) находят дисперсию повторности $s^2_{повт.}$ и дисперсию адекватности $s^2_{адекв.}$:

$$s^2_{повт} = \frac{Q_{повт}}{N-n};\ \ s^2_{адекв} = \frac{Q_{адекв}}{n-q} \tag{5.34}$$

После этого находят выборочное значение критерия F Фишера-Снедекора

$$f_{выб} = \frac{s^2_{адекв}}{s^2_{повт}} \tag{5.35}$$

и сравнивают его с критическим значением

$$f_{кр} = f_{кр}(\alpha; n-q; N-n), \tag{5.36}$$

взятом из таблицы 5 Приложения. И если $f_{выб} > f_{кр}$, то при данном уровне значимости α гипотезу H_0 об адекватности сглаживающего уравнения регрессии отвергают. То есть считают подобранное уравнение $\bar{y}_x^* = f^*(x)$ непригодным для приближения истинного (генерального) уравнения регрессии $\bar{y}_x = f(x)$. А если окажется, что $f_{выб} < f_{кр}$, то нет оснований отвергать гипотезу H_0. И только такое (адекватное) уравнение регрессии можно использовать в дальнейшем.

Кстати, если для имеющихся выборочных данных построено несколько различных сглаживающих выборочных уравнений регрессии, и все они адекватны выборочным данным, то лучшим среди них считается то, которое, не являясь заметно сложнее прочих, обеспечивает наибольший коэффициент детерминации d_{yx}.

Наряду с проверкой адекватности сглаживающего уравнения регрессии имеется возможность проверить и *значимость каждого его коэффициента в отдельности*. Это значит – имеется возможность установить, достаточно ли подсчитанное значение интересующего нас коэффициента для статистически обоснованного вывода а том, что он отличен от нуля. И если окажется, что коэффициент не значим, то его можно положить равным нулю. Это приведет к упрощению сглаживающего уравнения регрессии без существенного ущерба для его качества. Но на этом мы не останавливаемся. Отметим лишь, что такое исследование производится автоматически, если сглаживающее уравнение регрессии строится с помощью стандартной программы корреляционно-регрессионного анализа на ЭВМ. В сглаживающем уравнении регрессии, выдаваемом машиной, фигурируют лишь значимые коэффициенты, а заодно и указывается, адекватно ли все уравнение в целом.

Пример 1. На некотором предприятии исследовалась зависимость себестоимости Y единицы продукции (в условных единицах) от объема X произведенной за день продукции. Статистическое распределение выборки за 30 рабочих дней приведено в следующей таблице:

y_j \ x_i	5	10	15	20	25	$m_j = \sum_{i=1}^{5} n_{ij}$
10	–	–	–	1	4	5
11	–	3	6	4	1	14
12	1	3	2	–	1	7
13	3	–	1	–	–	4
$n_i = \sum_{j=1}^{4} n_{ij}$	4	6	9	5	6	$N = 30$

(5.37)

Требуется подобрать подходящую форму сглаживающего уравнения регрессии $\bar{y}_x^* = f^*(x)$, оценивающего корреляционную зависимость себестоимости Y единицы продукции от объема X продукции, произведенной за день, и построить это уравнение. Оценить степень тесноты указанной корреляционной зависимости, а также качество и адекватность построенного сглаживающего уравнения регрессии.

Решение. Сначала по данным корреляционной таблицы (5.37) построим корреляционное поле (рис. 3.13.).

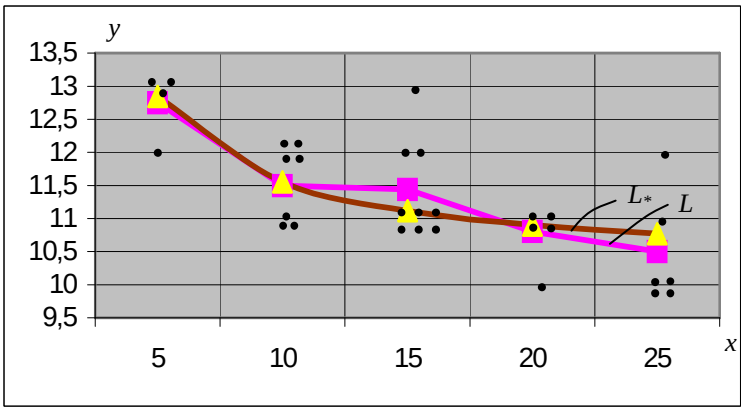


Рис. 3.13

Используя формулу (5.3), вычислим для каждого x_i выборочную среднюю $\bar{y}_{x_i}$:

$$\bar{y}_{x_1=5} = \frac{12\cdot 1 + 13\cdot 3}{4} = 12{,}75; \quad \bar{y}_{x_2=10} = \frac{11\cdot 3 + 12\cdot 3}{6} = 11{,}50;$$

$$\bar{y}_{x_3=15} = 11{,}44; \quad \bar{y}_{x_4=20} = 10{,}80; \quad \bar{y}_{x_5=25} = 10{,}50;$$

По точкам $(x_i; \bar{y}_{x_i})$ строим на корреляционном поле выборочную линию регрессии – ломанную L (ее узлы на рис. 3.13 обозначены квадратиками).

Теперь встает очередная задача – в какой форме $\bar{y}_x^* = f^*(x)$ искать сглаживающее равнение этой выборочной линии регрессии?

Обратим внимание на то, что с увеличением x выборочные средние $\bar{y}_x$ убывают, причем это убывание затухает. Так и должно быть (по смыслу рассматриваемых величин X и Y). Это дает основание строить сглаживающее выборочное уравнение регрессии в гиперболической форме (5.6). Коэффициенты k и b этого уравнения находятся по данным корреляционной таблицы (5.37) с помощью формул (5.27) и (5.28):

$$\overline{\left(\frac{1}{x}\right)} = \frac{\frac{1}{5}\cdot 4 + \frac{1}{10}\cdot 6 + \frac{1}{15}\cdot 9 + \frac{1}{20}\cdot 5 + \frac{1}{25}\cdot 6}{30} = 0{,}083$$

$$\overline{\left(\frac{1}{x^2}\right)} = \frac{\frac{1}{5^2}\cdot 4 + \frac{1}{10^2}\cdot 6 + \frac{1}{15^2}\cdot 9 + \frac{1}{20^2}\cdot 5 + \frac{1}{25^2}\cdot 6}{30} = 0{,}0094$$

$$\bar{y} = \frac{10\cdot 5 + 11\cdot 14 + 12\cdot 7 + 13\cdot 4}{30} = 11{,}33 \qquad (5.39)$$

$$\overline{\left(\frac{y}{x}\right)} = \frac{\frac{12}{5}\cdot 1 + \frac{13}{5}\cdot 3 + \frac{11}{10}\cdot 3 + \ldots + \frac{11}{25}\cdot 1 + \frac{12}{25}\cdot 1}{30} = 0{,}973$$

$$k = \frac{0{,}973 - 0{,}083\cdot 11{,}33}{0{,}0094 - (0{,}083)^2} \approx 13{,}0; \quad b = 11{,}33 - 13{,}0\cdot 0{,}083 \approx 10{,}25$$

Итак, сглаживающее выборочное уравнение регрессии в гиперболической форме (5.6) таково:

$$\bar{y}^*_x = f^*(x) = \frac{13{,}0}{x} + 10{,}25 \qquad (5.40)$$

Вычислим сглаживающие средние $\bar{y}^*_{x_i} = f^*(x_i)$ для всех x_i и сравним их с реальными выборочными средними $\bar{y}_{x_i}$:

x_i	5	10	15	20	25
n_i	4	6	9	5	6
$\bar{y}_{x_i}$	12,75	11,50	11,44	10,80	10,50
$\bar{y}^*_{x_i} = f^*(x_i)$	12,85	11,55	11,12	10,90	10,77

(5.41)

Впрочем, сначала убедимся, что и те, и другие средние подсчитаны правильно. Используя в качестве контроля формулы (5.10) убеждаемся, что обе суммы (5.10) дают один и тот же результат – общую среднюю $\bar{y}$ =11,33. То есть и реальные, и сглаживающие средние подсчитаны верно. И они весьма близки друг к другу. Это демонстрирует и рис. 3.13, где изображена гипербола (5.40) с указанием на ней точек $(x_i; \bar{y}^*_{x_i})$, помеченных треугольниками.

А теперь перейдем к получению ответов на остальные вопросы – о степени тесноты корреляционной зависимости Y от X и о качестве построенного уравнения (5.40).

Степень тесноты корреляционной зависимости Y от X оценивает выборочное корреляционное отношение η_{yx}. Подсчитывая его по формуле (5.22), получим: $\eta_{yx} \approx 0{,}75$. Величина η_{yx} оказалась весьма значительной (гораздо ближе к

1, чем нулю), что указывает на определенную и достаточно тесную корреляционную зависимость Y от X.

Подсчитаем еще, используя формулу (5.23), выборочный коэффициент детерминации d_{yx}: $d_{yx} \approx 0{,}52 \approx 52\%$. При этом, согласно (5.24),

$$(d_{yx})_{\max} = \eta^2_{yx} = 0{,}5625 = 56{,}25\ \% \qquad (5.42)$$

Значение $d_{yx} \approx 52\%$ указывает, что построенное сглаживающее выборочное уравнение регрессии (5.40) объясняет 52% общего объема вариации (изменения) величины Y в выборке и лишь немного не дотягивает до своего максимально возможного значения в 56,25 %. И это имеет место при весьма простом виде уравнения (5.40). Чтобы окончательно убедиться в высоком качестве этого уравнения, следует проверить его на адекватность выборочным данным. Для этого, используя формулы (5.33) – (5.35), подсчитаем выборочное значение критерия Фишера – Снедекора:

$$Q_{повт} = (12-12{,}75)^2 \cdot 1 + (13-12{,}75)^2 \cdot 3 + \ldots + (11-10{,}50)^2 \cdot 1 + (12-10{,}50)^2 \cdot 1 \approx 10{,}77$$

$$Q_{адекв} = (12{,}85-12{,}75)^2 \cdot 4 + (11{,}55-11{,}50)^2 \cdot 6 + \ldots + (10{,}77-10{,}50)^2 \cdot 6 \approx 1{,}464$$

$$s^2_{повт} = \frac{10{,}77}{30-5} \approx 0{,}431; \qquad s^2_{адекв} = \frac{1{,}464}{5-2} \approx 0{,}488; \qquad f_{выб} = \frac{0{,}488}{0{,}431} = 1{,}13 \qquad (5.43)$$

Далее задаем уровень значимости α (например α=0,05) и по таблице критических точек распределения Фишера – Снедекора находим:

$$f_{кр} = f_{кр}(0.05;\ 3;\ 25) = 2{,}99 \qquad (5.44)$$

И так как оказалось, что $f_{выб} < f_{кр}$, то у нас нет оснований отвергать гипотезу H_0 об адекватности уравнения (5.40) выборочным данным. В пользу этого свидетельствует и рис. 3.13: гипербола L_* нигде не выходит за пределы корреляционного поля.

Все задания примера 1 выполнены.

Кстати, если бы мы искали сглаживающее выборочное уравнение регрессии в линейной форме (5.5), то, используя (5.13) и (5.15), получили бы:

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i n_i}{N} = \frac{5 \cdot 4 + 10 \cdot 6 + 15 \cdot 9 + 20 \cdot 5 + 25 \cdot 6}{30} = 15{,}5$$

$$\overline{x^2} = \frac{\sum_{i=1}^{5} x_i^2 n_i}{N} = \frac{5^2 \cdot 4 + 10^2 \cdot 6 + 15^2 \cdot 9 + 20^2 \cdot 5 + 25^2 \cdot 6}{30} = 282{,}5$$

$$\bar{y} = \frac{\sum_{j=1}^{4} y_j m_j}{N} = \frac{10 \cdot 5 + 11 \cdot 14 + 12 \cdot 7 + 13 \cdot 4}{30} \approx 11{,}33$$

$$\overline{y^2} = \frac{\sum_{j=1}^{4} y_j^2 m_j}{N} = \frac{10^2 \cdot 5 + 11^2 \cdot 14 + 12^2 \cdot 7 + 13^2 \cdot 4}{30} \approx 129{,}3$$

$$\sigma_x^2 = \overline{x^2} - (\bar{x})^2 = 42{,}25; \quad \sigma_x = 6{,}50; \quad \sigma_y^2 = \overline{y^2} - (\bar{y})^2 = 0{,}823; \quad \sigma_y = 0{,}907$$

$$\overline{xy} = \frac{\sum_{i=1}^{5}\sum_{j=1}^{4} x_i y_j n_{ij}}{N} = 171{,}5;\quad k = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\sigma_x^2} = -0{,}0986;\quad b = \bar{y} - k\bar{x} = 12{,}86$$

И тогда вместо гиперболического (5.40) мы бы получили линейное уравнение

$$\bar{y}_x^* = f^*(x) = -0{,}0986x + 12{,}86 \qquad (5.46)$$

Если подсчитать по этому уравнению сглаживающие средние $\bar{y}_{x_i}^*$ и сравнить их с реальными выборочными средними $\bar{y}_{x_i}$, то получим следующую таблицу:

x_i	5	10	15	20	25
n_i	4	6	9	5	6
$\bar{y}_{x_i}$	12,75	11,50	11,44	10,80	10,50
$\bar{y}_{x_i}^* = f^*(x_i)$	12,37	11,87	11,38	10,89	10,40

(5.47)

Если для полученных сглаживающих средних $\bar{y}_{x_i}^*$ провести, на основе формулы (5.10), контроль, то он (проверьте это) сходится - опять получаем общую среднюю $\bar{y} = 11{,}33$.

Как видим, в таблице (5.47), как и в таблице (5.41), расхождение средних $\bar{y}_{x_i}$ и $\bar{y}_{x_i}^*$ невелико. То есть линейное уравнение (5.46), как и гиперболическое уравнение (5.40), тоже достаточно качественное.

Выясним все же, какое из них лучше. Для этого подсчитаем выборочный коэффициент детерминации d_{yx} и для линейного уравнения (5.46) . Используя формулу (5.23) получим: $d_{yx} \approx 0{,}50$ =50%. Впрочем, в линейном случае его можно было бы найти и по формуле (5.26), если предварительно найти выборочный коэффициент линейной корреляции ρ_{xy}:

$$\rho_{xy} = \frac{\sigma_x}{\sigma_y} k = \frac{6{,}50}{0{,}907}(-0{,}0986) \approx 0{,}71;\quad d_{yx} = (\rho_{xy})^2 \approx 0.50 = 50\%$$

Итак, линейное сглаживающее уравнение регрессии (5.46) объясняет примерно 50% всей вариации зависимой величины *Y*. Гиперболическое же уравнение (5.40) объясняло чуть больше – 52% этой вариации. То есть по этому показателю гиперболическое уравнение несколько лучше линейного. И оно, по существу, так же просто, как и линейное.

Выше мы показали, что гиперболическое уравнение адекватно выборочным данным. Покажем, что и линейное уравнение им адекватно. Используя опять формулы (5.33) – (5.35), получаем:

$$Q_{повт} = 10{,}77 \text{ (см. (5.43)}$$

$$Q_{адекв} = (12{,}37 - 12{,}75)^2 \cdot 4 + (11{,}87 - 11{,}50)^2 \cdot 6 + \ldots + (10{,}40 - 10{,}50)^2 \cdot 6 \approx 1{,}53 \qquad (5.49)$$

$$s_{повт}^2 = \frac{10{,}77}{30-5} \approx 0{,}431;\qquad s_{адекв}^2 = \frac{1{,}53}{5-2} \approx 0{,}511;\qquad f_{выб} = \frac{0{,}511}{0{,}431} \approx 1{,}18$$

При том же уровне значимости α=0,05, который был принят при проверке гипотезы об адекватности гиперболического уравнения регрессии, в соответствии с (5.36) получаем для линейного уравнения (5.46) то же самое критиче-

ское значение критерия Фишера – Снедекора, что было указано в (5.44): $f_{kp} = 2.99$. И так как опять $f_{выб} < f_{кр}$, то у нас нет оснований отвергать гипотезу H_0 и об адекватности линейного сглаживания уравнения регрессии (5.46) выборочным данным.

В общем, оба сглаживающие выборочные уравнения регрессии – гиперболическое (5.40) и линейное (5.46) – адекватны выборочным данным и оба достаточно хороши. Из них несколько лучшим является гиперболическое уравнение (5.40).

5.2. Множественная корреляция.

Перейдем теперь к случаю, когда выборочным путем изучается корреляционная зависимость (зависимость в среднем) одной величины от нескольких других (множественная корреляция). Ограничимся рассмотрением случая, когда изучается корреляционная зависимость некоторой величины Z от двух величин X и Y.

Будем считать, что при изучении этой зависимости проведено n различных опытов, в которых измерялись значения всех трех величин и в которых пары $(x_i; y_i)$ $(i=1,2\ldots n)$ значений величин X и Y варьировались следующим образом:

i	1	2	3	…	n
x_i	x_1	x_2	x_3	…	x_n
y_i	y_1	y_2	y_3	…	y_n

(5.50)

Далее, будем считать, что каждый из этих опытов (при фиксированных $(x_i; y_i)$) повторен некоторое (не обязательно одинаковое) число раз n_i. Повторные опыты при одних и тех же значениях $(x_i; y_i)$ дают, вообще говоря, различные значения z_{ij} $(j=1,2,\ldots n_i)$ величины Z. Пусть

$$\bar{z}_i = \bar{z}_{x_i y_i} = \frac{\sum_{j=1}^{n_i} z_{ij}}{n_i} \quad (i=1,2\ldots n) \tag{5.51}$$

- среднее их значение для каждого i-го опыта . В итоге совокупность всех опытных данных (корреляционная таблица) примет вид:

i (номер опыта)	1	2	3	….	n
n_i (повторность опыта)	n_1	n_2	n_3	….	n_n
x_i	x_1	x_2	x_3	….	x_n
y_i	y_1	y_2	y_3	….	y_n
$\bar{z}_i = \bar{z}_{x_i y_i}$	$\bar{z}_1$	$\bar{z}_2$	$\bar{z}_3$	….	$\bar{z}_n$

(5.52)

Заметим, что общее количество N всех проведенных опытов найдется, очевидно, по формуле:

$$N = \sum_{i=1}^{n} n_i \tag{5.53}$$

Основные задачи корреляционно – регрессионного анализа при множественной корреляции те же, что и при парной.

В частности, *первой основной задачей* по исследованию корреляционной зависимости Z от X и Y является построение выборочного уравнения регрессии

$$\bar{z}^{*}_{xy} = f^{*}(x; y), \tag{5.54}$$

наилучшим образом выравнивающего (сглаживающего) выборочные данные, а следовательно, являющегося наилучшим приближением *истинного (генерального)* уравнения регрессии

$$\bar{z}_{xy} = f(x; y) \tag{5.55}$$

Заметим, что геометрически этому генеральному уравнению регрессии соответствует уже не линия регрессии на плоскости *xoy*, а *поверхность регрессии* в пространстве ($x; y; z$)

Второй основной задачей является оценка тесноты корреляционной зависимости Z от X и Y.

Ограничимся рассмотрением случая, когда наилучшее сглаживающееся уравнение регрессии (5.54) строится в линейной форме

$$\bar{z}^{*}_{xy} = ax + by + c \tag{5.56}$$

То есть когда уравнение поверхности в пространстве (5.55) приближается (приближенно заменяется) уравнением плоскости в пространстве (5.56). Параметры ($a;b;c$) этого уравнения находятся, как обычно, методом наименьших квадратов:

$$Q = Q(a;b;c) = \sum_{i=1}^{n} [\bar{z}_i - (ax_i + by_i + c)]^2 \cdot n_i \quad \rightarrow \quad \min \tag{5.57}$$

Реализация необходимых условий минимума функции Q

$$\begin{cases} Q'_a = 0; \\ Q'_b = 0 \\ Q'_c = 0 \end{cases} \tag{5.58}$$

приводит к следующей системе трех линейных уравнений с тремя неизвестными ($a;b;c$) (к так называемой *нормальной системе*):

$$\begin{cases} a\sum_{i=1}^{n} x_i^2 n_i + b\sum_{i=1}^{n} x_i y_i n_i + c\sum_{i=1}^{n} x_i n_i = \sum_{i=1}^{n} \bar{z}_i x_i n_i \\ a\sum_{i=1}^{n} x_i y_i n_i + b\sum_{i=1}^{n} y_i^2 n_i + c\sum_{i=1}^{n} y_i n_i = \sum_{i=1}^{n} \bar{z}_i y_i n_i \\ a\sum_{i=1}^{n} x_i n_i + b\sum_{i=1}^{n} y_i n_i + cN = \sum_{i=1}^{n} \bar{z}_i n_i \end{cases} \tag{5.59}$$

Решая эту нормальную систему, находим (*a;b;c*), а вместе с ними находим и наилучшее линейное уравнение (5.56), сглаживающее выборочные данные корреляционной таблицы (5.52). В своем окончательном виде оно таково:

$$\bar{z}^{*}_{xy} - \bar{z} = a(x - \bar{x}) + b(y - \bar{y}) \tag{5.60}$$

Здесь

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i n_i}{N} \quad ; \quad \bar{y} = \frac{\sum_{i=1}^{n} y_i n_i}{N} \quad ; \quad \bar{z} = \frac{\sum_{i=1}^{n} \bar{z}_i n_i}{N} \tag{5.61}$$

- выборочные средние величин *X*, *Y* и *Z* соответственно, а коэффициенты *a* и *b* находится по формулам:

$$a = \frac{\rho_{xz} - \rho_{xy}\rho_{yz}}{1 - \rho^2_{xy}} \cdot \frac{\sigma_z}{\sigma_x} \quad ; \quad b = \frac{\rho_{yz} - \rho_{xz}\rho_{xy}}{1 - \rho^2_{xy}} \cdot \frac{\sigma_z}{\sigma_y} \tag{5.62}$$

При этом

$$\rho_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y} \; ; \; \rho_{xz} = \frac{\overline{xz} - \bar{x} \cdot \bar{z}}{\sigma_x \sigma_z} \; ; \; \rho_{yz} = \frac{\overline{yz} - \bar{y} \cdot \bar{z}}{\sigma_y \sigma_z} \tag{5.63}$$

- выборочные значения парных коэффициентов линейной корреляции между соответствующими парами величин *X*, *Y*, *Z*, а

$$\sigma_x = \sqrt{\overline{x^2} - (\bar{x})^2} \; ; \; \sigma_y = \sqrt{\overline{y^2} - (\bar{y})^2} \; ; \; \sigma_z = \sqrt{\overline{z^2} - (\bar{z})^2} \tag{5.64}$$

- выборочные значения среднеквадратических отклонений величин *X*, *Y* и *Z* соответственно. При подсчете выражений (5.63) и (5.64) используются аналогичные (5.13) формулы:

$$\begin{gathered} \overline{x^2} = \frac{\sum_{i=1}^{n} x_i^2 n_i}{N} \; ; \; \overline{y^2} = \frac{\sum_{i=1}^{n} y_i^2 n_i}{N} \; ; \; \overline{z^2} = \frac{\sum_{i=1}^{n} (\bar{z}_i)^2 n_i}{N} \; ; \\ \overline{xy} = \frac{\sum_{i=1}^{n} x_i y_i n_i}{N} \; ; \; \overline{xz} = \frac{\sum_{i=1}^{n} x_i \bar{z}_i n_i}{N} \; ; \; \overline{yz} = \frac{\sum_{i=1}^{n} y_i \bar{z}_i n_i}{N} \end{gathered} \tag{5.65}$$

Вторая основная задача – установление тесноты линейной корреляционной зависимости *Z* от *X* и *Y* – решается с помощью так называемого *совокупного коэффициента линейной корреляции* *R*(*Z*, *XY*) (или просто *R*), выборочное значение которого находится по формуле:

$$R_{z,xy} = R = \sqrt{\frac{\rho^2_{xz} + \rho^2_{yz} - 2\rho_{xz} \cdot \rho_{yz} \cdot \rho_{xy}}{1 - \rho^2_{xy}}} \tag{5.66}$$

Доказано, что:

1. Совокупный коэффициент линейной корреляции имеет возможные значения в промежутке [0;1].

2. Если R=0, то Z не может быть связана с X и Y линейной корреляционной зависимостью. Однако при этом возможна нелинейная корреляционная и даже функциональная зависимость Z от X и Y.

3. Если R=1, то Z связана с X и Y линейной функциональной зависимостью вида

$$Z=aX+bY+c \qquad (5.67)$$

Для выборочных данных (x_i; y_i; z_{ij}) последний случай означает, что все повторные (для разных j) значения z_{ij} совпадают и равны одному и тому же значению $\bar{z}_i$, а все пространственные точки (x_i; y_i; $\bar{z}_i$) располагаются на одной и той же плоскости – а именно, на плоскости (5.56) (или, что одно и то же, на плоскости (5.60)).

4. Если R отличен от своих крайних значений (0 и 1), то при приближении R к единице теснота линейной корреляционной зависимости Z от X и Y увеличивается. Это значит, что экспериментальные пространственные точки (x_i; y_i; z_{ij}) все теснее примыкают к плоскости (5.60).

Частные коэффициенты корреляции.

Если переменные X и Y коррелированны не только с величиной Z ($\rho_{xz} \neq 0$; $\rho_{yz} \neq 0$), но и между собой ($\rho_{xy} \neq 0$), то на величине выборочных парных коэффициентов линейной корреляции ρ_{xz} и ρ_{yz} сказывается не только коррелированность X с Z и Y с Z соответственно, но и коррелированность (взаимозависимость) величин X и Y между собой. Чтобы очистить корреляцию X с Z от влияния Y и корреляцию Y с Z от влияния X, рассматривают так называемую *частную корреляцию* между Z и одной из переменных X и Y при исключении влияния (*элиминировании*) другой переменной. То есть при условии, что эта другая переменная не меняется. Указанная частная корреляция оценивается с помощью так называемых *выборочных частных коэффициентов линейной корреляции* $\rho_{xz\,(y)}$ (здесь y =const) и $\rho_{yz(x)}$ (здесь x=const):

$$\rho_{xz(y)}=\frac{\rho_{xz}-\rho_{xy}\rho_{yz}}{\sqrt{(1-\rho_{xy}^2)\cdot(1-\rho_{yz}^2)}}\ ;\ \rho_{yz(x)}=\frac{\rho_{yz}-\rho_{xy}\rho_{xz}}{\sqrt{(1-\rho_{xy}^2)\cdot(1-\rho_{xz}^2)}} \qquad (5.68)$$

Частные коэффициенты линейной корреляции (5.68), как и парные коэффициенты ρ_{xz} и ρ_{yz}, определяемые формулами (5.63), по своим свойствам аналогичны обычному выборочному коэффициенту линейной корреляции ρ_{xy}, оценивающему наличие и тесноту линейной корреляционной связи между случайными величинами X и Y. Их возможные значения заключены в пределах [-1;1]. Когда, например, $\rho_{xz\,(y)}$=0, то выборка указывает на отсутствие линейной корреляционной зависимости Z от X (при постоянном Y), хотя нелинейная корреляционная и даже функциональная зависимость Z от X остаются возможными. Если же $\rho_{xz\,(y)}$= 1 или $\rho_{xz\,(y)}$= -1, то выборка указывает на строгую линейную функциональную зависимость Z от X (соответственно возрастающую или убывающую) при постоянном Y. И т.д.

Кстати, и парные ρ_{xz}, ρ_{yz}, и частные $\rho_{xz\,(y)}$, $\rho_{yz(x)}$ коэффициенты линейной корреляции по своей абсолютной величине не больше, чем множественный коэффициент линейной корреляции $R=R_{z,xy}$.

Множественный коэффициент, равно как и парные, и частные коэффициенты линейной корреляции используются лишь для оценки наличия и степени тесноты *линейной корреляционной зависимости Z от X и Y*. Степень же тесноты *любой* (а не только линейной) корреляционной зависимости Z от X и Y оценивается с помощью *совокупного выборочного корреляционного отношения* $\eta_{z,xy}$:

$$\eta_{z,xy} = \frac{\sigma_{\bar{z}}}{\sigma_z} = \sqrt{\frac{\sum_{i=1}^{n}(\bar{z}_i - \bar{z})^2 n_i}{\sum_{i=1}^{n}\sum_{j=1}^{n_i}(z_{ij} - \bar{z})^2}} \qquad (5.69)$$

Оно показывает, какую долю составляет средний разброс выборочных средних $\bar{z}_i$ вокруг общей средней $\bar{z}$ по отношению к среднему разбросу всех выборочных значений z_{il} величины Z вокруг той же общей средней $\bar{z}$. Формула (5.69) по своему смыслу аналогична формуле (5.22), записанной для случая, когда исследуется парная корреляционная зависимость одной величины от другой (Y от X). Аналогичны и свойства определяемых этими формулами корреляционных отношений. А именно:

а) возможные значения $\eta_{z,xy}$ заключены в пределах [0;1].

б) если $\eta_{z,xy}$=0, то это значит что все $\bar{z}_i = \bar{z}$ (i=1,2,…n). То есть все выборочные средние $\bar{z}_i$ величины Z одинаковы для всех выборочных пар (x_i; y_i) значений величин ($X;Y$). Но это говорит о том, что в выборке величина Z корреляционно (в среднем) не зависит от величин X и Y. Что, в свою очередь, свидетельствует в пользу аналогичной корреляционной независимости Z от X и Y и в генеральной совокупности.

в) если $\eta_{z,xy}$=1, то это значит, что $\sigma_{\bar{z}} = \sigma_z$. То есть весь разброс значений z_{ij} величины Z в выборке сосредоточен в разбросе средних выборочных $\bar{z}_i$. Разброс значений z_{ij} вокруг их средних $\bar{z}_i$, таким образом, отсутствует, что указывает на некую жесткую функциональную зависимость $Z=f(X;Y)$ величины Z от X и Y.

Если в формуле (5.69) заменить реальные выборочные средние $\bar{z}_i$ на сглаживающие средние $\bar{z}_i^* = f^*(x_i; y_i)$, вычисленные по подобранному сглаживающему уравнению регрессии (5.54), и возвести полученное выражение в квадрат, то получим так называемый *совокупный выборочный коэффициент детерминации*:

$$d_{z,xy} = \frac{\sigma^2_{\bar{z}^*}}{\sigma^2_z} = \frac{\sum_{i=1}^{n}(\bar{z}_i^* - \bar{z})^2 n_i}{\sum_{i=1}^{n}\sum_{j=1}^{n_i}(z_{ij} - \bar{z})^2} \qquad (5.70)$$

Совокупный выборочный коэффициент детерминации (5.70) показывает, какую долю вариации (изменения) исследуемой величины Z объясняет в выборке подобранное сглаживающее уравнение регрессии (5.54). Чем этот коэффициент больше, тем лучшим считается подобранное выборочное уравнение регрессии

(5.54), сглаживающее выборочные данные (5.52). Формула (5.70) и её смысл аналогичен формуле (5.23), полученной при исследовании парной корреляционной зависимости одной величины от другой (*Y* от *X*).

Кстати, имеют место и аналоги равенств (5.24) и (5.25). В частности:

$$(d_{z,xy})_{\max} = \eta^2_{z,xy} \qquad 5.71)$$

Если сглаживающее уравнение регрессии (5.54) строится в линейной форме (5.56), то совокупный выборочный коэффициент детерминации можно найти не только по общей формуле (5.70), но и по частной формуле:

$$d_{z,xy} = R^2_{z,xy} = R^2, \qquad (5.72)$$

являющейся аналогом формулы (5.25).

Адекватность выборочного сглаживающего уравнения регрессии.

Построив с помощью метода наименьших квадратов некое сглаживающее выборочное уравнение регрессии (5.54) – например, в линейной форме (5.56) или в какой-то нелинейной форме, мы затем должны проверить его на адекватность выборочным данным (5.52). То есть сделать то, что делалось выше для случая парной корреляции (формулы (5.31 - 5.36)). Для этого:

а) Находим суммы $Q_{повт}$ и $Q_{адекв}$:

$$Q_{повт} = \sum_{i=1}^{n}\sum_{j=1}^{ni}(z_{ij} - \bar{z}_i)^2; \quad Q_{адекв} = \sum_{i=1}^{n}(\bar{z}_i^* - \bar{z}_i)^2 n_i \qquad (5.73)$$

б) Находим дисперсии повторности и адекватности

$$s^2_{повт} = \frac{Q_{повт}}{k_{повт}} = \frac{Q_{повт}}{N-n}; \quad s^2_{адекв} = \frac{Q_{адекв}}{k_{адекв}} = \frac{Q_{адекв}}{n-q}, \qquad (5.74)$$

где *q*- число оцениваемых по выборочным данным коэффициентов подобранного сглаживающего уравнения регрессии (5.54) (в случае (5.56) $q = 3$);

в) Находим

$$f_{выб} = \frac{s^2_{адекв}}{s^2_{повт}} \text{ и } f_{кр} = f_{кр}(\alpha; n-q; N-n), \qquad (5.75)$$

где $f_{кр}$ - табличное критическое значение критерия Фишера – Снедекора. И если окажется, что $f_{выб} \prec f_{кр}$, то подобранное уравнение регрессии (5.54) признаем адекватным выборочным данным (при заданном уравнение значимости α). А если окажется, что $f_{выб} \succ f_{кр}$, то это уравнение считаем неадекватным, а следовательно, непригодным для оценки истинного (генерального) уравнения регрессии.

Наряду с проверкой адекватности выборочного сглаживающего уравнения регрессии *в целом* имеется возможность проверить и значимость *каждого его коэффициента в отдельности*. Это значит имеется возможность установить, достаточно ли велико по модулю подсчитанное значение интересующего нас коэффициента для статистически обоснованного вывода о том, что он отличен от нуля. То есть значим ли он (нужно его учитывать) или он незначим и учитывать его не нужно. И если окажется, что коэффициент незначим, то его можно

положить равным нулю. Это приведёт к упрощению подобранного уравнения регрессии без существенного ущерба для его качества.

Но на этом мы не останавливаемся. Отметим лишь, что такое исследование проводится автоматически, если выборочное сглаживающее уравнение регрессии строится с помощью стандартной программы корреляционно-регрессионного анализа на ЭВМ. В уравнении регрессии, выдаваемом машиной, фигурируют лишь значимые коэффициенты, а заодно и указывается, адекватно ли всё уравнение в целом.

Упражнения.

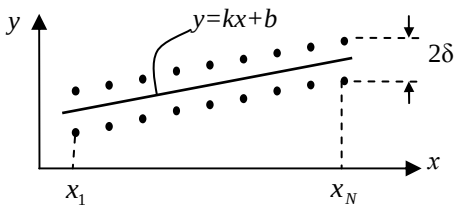


Рис. 3.14

1. Пусть при исследовании корреляционной зависимости X от Y получено $2N$ точек корреляционного поля, расположенных симметрично относительно некоторой прямой $y = kx + b$ и удаленных от неё в направлении оси oy на $\pm\delta$ (см. рис. 3.14). Требуется:

1) Найти выборочное сглаживающее уравнение регрессии $\bar{y}_x^* = f^*(x)$;

2) Найти выборочный коэффициент линейной корреляции ρ_{yx};

3) Найти выборочное корреляционное отношение η_{yx};

4) Найти выборочный коэффициент детерминации d_{yx};

5) Исследовать вопрос об адекватности выборочным данным найденного уравнения регрессии.

Ответ: 1) $\bar{y}_x^* = kx + b$ - точное сглаживающее выборочное уравнение регрессии;

2) $\rho_{yx} = \dfrac{1}{\sqrt{1+\left(\dfrac{\delta}{k\sigma_x}\right)^2}}$; 3) $\eta_{yx} = \rho_{yx}$; 4) $d_{yx} = \rho_{yx}^2 = \dfrac{1}{1+\left(\dfrac{\delta}{k\sigma_x}\right)^2}$;

5) Выборочное уравнение регрессии $\bar{y}_x = kx + b$ адекватно выборочным данным при любом уровне значимости α.

2. При исследовании корреляционной зависимости урожайности некоторой культуры Y (*ц/га*) от глубины орошения X (*см*) получена следующая корреляционная таблица:

y_j \ x_i	0	10	20	30	40	50	$m_j = \sum_{i=1}^{6} n_{ij}$
10	4	-	-	-	-	2	6
12	1	2	1	2	2	2	10

14	-	3	4	2	3	2	14
16	-	2	4	3	1	-	10
$n_i = \sum_{j=1}^{4} n_{ij}$	5	7	9	7	6	6	N=40

Построить корреляционное поле, соответствующее данной корреляционной таблице. С его помощью из уравнений (5.5) - (5.7) выбрать наиболее подходящую форму для построения выборочного сглаживающего уравнения регрессии $\bar{y}_x^* = f^*(x)$ и построить это уравнение. Найти выборочное корреляционное отношение η_{yx} и выборочный коэффициент детерминации d_{yx}. При уровне значимости $\alpha = 0{,}05$ проверить гипотезу об адекватности полученного выборочного уравнения регрессии.

Ответ. По конфигурации корреляционного поля, а также по смыслу рассматриваемых величин X и Y наиболее подходящей формой для выборочного уравнения регрессии является параболическая форма (5.7). Коэффициенты (a_0; a_1; a_2) уравнения (5.7) находятся из решения нормальной системы (5.30). Для формирования матрицы этой системы рекомендуется по данным заданной корреляционной таблицы составить следующую вспомогательную таблицу:

x_i	n_i	$\bar{y}_{x_i}$	$x_i n_i$	$x_i^2 n_i$	$x_i^3 n_i$	$x_i^4 n_i$	$\bar{y}_{x_i} n_i$	$\bar{y}_{x_i} x_i n_i$	$\bar{y}_{x_i} x_i^2 n_i$
0	5	10,4	0	0	0	0	52	0	0
10	7	14	70	700	7000	70000	98	980	9800
20	9	44/3	180	3600	72000	1440000	132	2640	52800
30	7	100/7	210	6300	189000	5670000	100	3000	90000
40	6	41/3	240	9600	384000	15360000	82	3280	131200
50	6	12	300	15000	750000	37500000	72	3600	180000
$\sum$	40	-	1000	35200	1402000	60040000	536	13500	463800

В итоге (убедитесь в этом) получаем следующее выборочное уравнение регрессии:

$$\bar{y}_x^* = f^*(x) = 10{,}96 + 0{,}2913 \cdot x - 0{,}0055 \cdot x^2$$

При этом

$$\eta_{yx} = 0{,}702; \quad d_{yx} = 0{,}446 = 44{,}6\%$$

Найденное выборочное уравнение регрессии адекватно выборочным данным, причем не только для уровня значимости $\alpha = 0{,}05$, но и вообще для любого уровня значимости. Действительно, используя формулы (5.33) - (5.36) и данные предыдущих таблиц получаем:

$$Q_{ад} = \sum_{i=1}^{6} \left(\bar{y}_{x_i}^* - \bar{y}_{x_i}\right)^2 n_i = 6{,}739; \quad Q_{повт} = \sum_{i=1}^{6}\sum_{j=1}^{4} (y_j - \bar{y}_{x_i})^2 n_{ij} = 81{,}96$$

$$s^2_{ад} = \frac{Q_{ад}}{n-q} = \frac{6{,}739}{6-3} = 2{,}246;\ \ s^2_{повт} = \frac{Q_{повт}}{N-n} = \frac{81{,}96}{40-6} = 2{,}411$$

$$f_{выб} = \frac{s^2_{ад}}{s^2_{повт}} = \frac{2{,}246}{2{,}411} = 0{,}93 < 1.$$

Но табличное значение критерия Фишера-Снедекора $f_{кр} = f_{кр}(\alpha; k_1; k_2) \succ 1$ для любых значений $(\alpha; k_1; k_2)$. Поэтому $f_{выб} \prec f_{кр}$ для любых значений α. А значит, выборочное уравнение регрессии адекватно выборочным данным при любом α.

3. Выборочным путем исследовалась корреляционная зависимость некоторой случайной величины *Z* от двух других случайных величин *X* и *Y*. Данные выборки (корреляционная таблица) представлены в первых четырех столбцах таблицы, приводимой ниже:

i	x_i	y_i	z_i	$x_i y_i$	$x_i z_i$	$y_i z_i$	x_i^2	y_i^2	z_i^2
1	25,5	17,3	85,5	441,15	2180,25	1479,15	650,25	229,29	7310
2	34,1	19,8	81,7	665,18	2785,97	1617,66	1162,81	392,04	6675
3	37,3	30,1	71,7	1122,73	2674,41	2158,17	1391,29	906,01	5141
4	44,7	31,9	62,7	1425,93	2802,69	2000,13	1998,09	1017,61	3931
5	44,6	38,3	66,4	1708,18	2961,44	2543,12	1989,16	1466,89	4409
6	41,0	26,5	70,6	1086,50	2894,60	1870,90	1681,00	702,25	4984
7	49,5	36,2	65,0	1791,90	3217,50	2353,00	2450,25	1310,44	4225
8	45,1	21,0	72,8	947,10	3283,28	1528,80	2034,01	441,00	5300
9	56,5	29,5	67,6	1666,75	3819,40	1994,20	3192,25	870,25	4570
10	35,4	29,4	90,0	881,46	3186,00	2241,00	1253,16	620,01	8100
11	54,9	25,0	60,2	1372,50	3304,98	1505,00	3014,01	625,00	3624
12	32,8	28,0	74,8	918,40	2453,44	2094,40	1075,84	784,00	5595
13	55,5	33,9	63,4	1881,45	3518,70	2149,26	3080,25	1149,21	4020
14	41,5	16,1	74,2	668,15	3079,30	1194,62	1722,25	259,21	5506
15	41,5	21,0	71,6	871,50	2971,40	1503,60	1722,25	441,00	5127
16	52,7	28,7	60,0	1512,49	3162,00	1722,00	2777,49	823,69	3600
17	37,9	20,3	81,1	769,37	3073,69	1646,33	1436,41	412,09	6577
18	43,9	19,9	71,5	873,61	3138,85	1422,85	1927,21	396,01	5112
19	35,0	22,6	77,2	791,00	2702,00	1744,72	1225,00	510,76	5960
20	27,8	20,1	91,2	558,78	2535,36	1833,12	772,84	404,01	8317
$\sum$	837,2	511,1	1459,2	21964,13	59745,26	36602,03	36555,62	13830,77	108083

Требуется:

1) Построить выборочное сглаживающее уравнение регрессии $\bar{z}^*_{xy} = f^*(x; y)$ в линейной форме (5.56).

2) Найти совокупный выборочный коэффициент линейной корреляции *R* и частные выборочные коэффициенты линейной корреляции $\rho_{xz(y)}$ и $\rho_{yz(x)}$.

3) Найти выборочный коэффициент детерминации $d_{z,xy}$.

4) Оценить адекватность выборочным данным построенного уравнения регрессии.

Ответ:

1) Искомое выборочное уравнение регрессии имеет вид

$$\bar{z}^{*}_{xy} = 113{,}25 - 0{,}7603x - 0{,}3313y$$

Его можно получить, или составив и решив для коэффициентов (*a;b;c*) уравнения (5.56) нормальную систему (5.59) (используя для этого результаты приведенной выше таблицы), или воспользовавшись формулами (5.60) - (5.65).

2) R=0,877; $\rho_{xz(y)}$= 0,792; $\rho_{yz(x)}$=0,374; 3) $d_{z(x,y)} = R^2$ = 0.769;

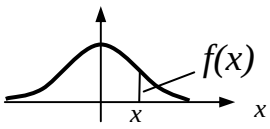
4) Адекватность полученного выборочного уравнения регрессии проверить невозможно, так как корреляционная таблица не содержит результатов повторных опытов для одних и тех же (x_i ; $y_{i)}$, а следовательно, невозможно найти дисперсию повторности $s^2_{повт}$..

ПРИЛОЖЕНИЕ

(таблицы)

Таблица 1.

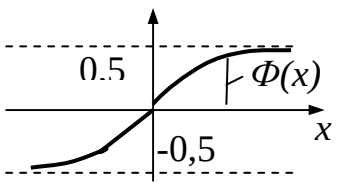
Значения функции $f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$



x	0	1	2	3	4	5	6	7	8	9
0,0	0,3989	0,3989	0,3989	0,3988	0,3986	0,3984	0,3982	0,3980	0,3977	0,3973
0,1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0,2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0.3	3814	3802	3790	3778	3765	3752	3739	3726	3712	3697
0.4	3683	3668	3652	3637	3621	3605	3589	3572	3555	3538
0,5	3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0,6	3332	3312	3292	3271	3251	3230	3209	3187	3166	3144
0.7	3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0,8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0.9	2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1,0	2420	2396	2371	2347	2323	2299	2275	2251	2227	2203
1,1	2179	2155	2131	2107	2083	2059	2036	2012	1989	1965
1,2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1.4	1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1.5	1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1.6	1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1,7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1.8	0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
1.9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2.0	0540	0529	0519	0508	0498	0488	0478	0468	0459	0449
2,1	0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2.2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2.3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2.4	0224	0219	0213	0208	0203	0198	0194	0189	0184	0180
2,5	0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2,6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2,8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2.9	0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3,0	0044	0043	0042	0040	0039	0038	0037	0036	0035	0034
3.1	0033	0032	0031	0030	0029	0028	0027	0026	0025	0026
3,2	0024	0023	0022	0022	0021	0020	0020	0019	0018	0018
3,3	0017	0017	0016	0016	0015	0015	0014	0014	0013	0013
3,4	0012	0012	0012	0011	0011	0010	0010	0010	0009	0009
3,5	0009	0008	0008	0008	0008	0007	0007	0007	0007	0006
3,6	0006	0006	0006	0005	0005	0005	0005	0005	0005	0004
3,7	0004	0004	0004	0004	0004	0004	0003	0003	0003	0003
3,8	0003	0003	0003	0003	0003	0002	0002-	0002	0002	0002
3,9	0002	0002	0002	0002	0002	0002	0002	0002	0001	0001

Таблица 2.

Значения функции $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{t^2}{2}}dt$



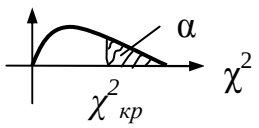
x	Φ(x)	x	Φ(x)	x	Φ(x)	x	Φ(x)	x	Φ(x)	x	Φ(x)
0,00	0,0000	0,39	0,1517	0,78	0,2823	1,17	0,3790	1,56	0,4406	1,95	0,4744
0,01	0,0040	0,40	0,1554	0,79	0,2852	1,18	0,3810	1,57	0,4418	1,96	0,4750

0,02	0,0080	0,41	0,1591	0,80	0,2881	1.19	0,3830	1,58	0,4429	1,97	0,4756
0,03	0,0120	0,42	0,1628	0,81	0.2910	1,20	0,3849	1,59	0.4441	1,98	0,4761
0,04	0,0160	0,43	0,1664	0,82	0,2939	1,21	0,3869	1,60	0,4452	1,99	0,4767
0,05	0,0199	0,44	0,1700	0,83	0,2967	1,22	0,3883	1,61	0,4463	2,00	0,4772
0,06	0,0239	0,45	0,1736	0,84	0,2995	1,23	0,3907	1.62	0,4474	2,02	0,4783
0,07	0,0279	0,46	0,1772	0,85	0,3023	1,24	0,3925	1,63	0,4484	2,04	0,4793
0,08	0,0319	0,47	0,1808	0,86	0,3051	1,25	0,3944	1,64	0.4495	2,06	0,4803
0,09	0,0359	0,48	0,1844	0,87	0,3078	1,26	0,3962	1,65	0,4505	2,08	0,4812
0,10	0,0398	0,49	0,1879	0,88	0,3106	1,27	0,3980	1,66	0,4515	2,10	0,4821
0,11	0,0438	0,50	0,1915	0,89	0,3133	1,28	0,3997	1,67	0,4525	2,12	0,4830
0,12	0,0478	0,51	0,1950	0,90	0,3159	1,29	0,4015	1,68	0,4535	2,14	0.4838
0,13	0,0517	0,52	0,1985	0,91	0,3186	1,30	0,4032	1,69	0,4545	2,16	0,4846
0,14	0,0557	0,53	0,2019	0,92	0,3212	1,31	0,4049	1,70	0,4554	2.18	0,4854
0,15	0,0596	0,54	0,2054	0,93	0,3238	1,32	0,4066	1,71	0,4564	2.20	0,4861
0,16	0,0636	0,55	0,2088	0,94	0,3264	1,33	0,4082	1,72	0,4573	2,22	0,4868
0,17	0,0675	0,56	0,2123	0,95	0,3289	1,34	0,4099	1,73	0,4582	2,24	0,4875
0,18	0,0714	0,57	0,2157	0,96	0,3315	1,35	0,4115	1,74	0,4591	2,26	0,4881
0,19	0,0753	0,58	0,2190	0,97	0,3340	1,36	0,4131	1,75	0,4599	2,28	0,4887
0,20	0,0793	0,59	0,2224	0,98	0,3365	1,37	0.4147	1,76	0,4608	2,30	0,4893
0,21	0,0832	0,60	0,2257	0.99	0,3389	1,38	0,4162	1,77	0,4616	2,32	0,4898
0,22	0,0871	0,61	0,2291	1,00	0,3413	1,39	0,4177	1,78	0,4625	2,34	0,4904
0,23	0,0910	0,62	0,2324	1,01	0,3438	1,40	0,4192	1,79	0,4633	2,36	0,4909
0,24	0,0948	0,63	0,2357	1,02	0,3461	1.41	0,4207	1,80	0,4641	2,38	0,4913
0,25	0,0987	0,64	0,2389	1,03	0,3485	1.42	0,4222	1,81	0,4649	2,40	0,4918
0,26	0,1026	0,65	0,2422	1,04	0.3508	1,43	0,4236	1,82	0,4656	2,42	0,4922
0,27	0,1064	0,66	0,2454	1,05	0,3531	1,44	0,4251	1,83	0,4664	2.44	0,4927
0,28	0,1103	0,67	0,2486	1,06	0,3554	1,45	0,4265	1,84	0,4671	2,46	0,4931
0,29	0,1141	0,68	0,2517	1,07	0,3577	1,46	0,4279	1,85	0.4678	2,48	0,4934
0,30	0,1179	0,69	0,2549	1,08	0,3599	1,47	0,4292	1,86	0,4686	2,50	0,4938
0,31	0,1217	0,70	0,2580	1,09	0,3621	1,48	0,4306	1,87	0,4693	2,60	0,4953
0,32	0,1255	0,71	0,2611	1,10	0,3643	1,49	0,4319	1,88	0,4699	2,70	0,4965
0,33	0,1293	0,72	0,2642	1.11	0,3665	1,50	0,4332	1,89	0,4706	2,80	0,4974
0,34	0,1331	0,73	0,2673	1,12	0,3686	1,51	0,4345	1,90	0,4713	2,90	0,4981
0.35	0,1368	0,74	0,2703	1,13	0,3708	1,52	0,4357	1,91	0,4719	3,00	0,4986
0.36	0,1406	0,75	0,2734	1.14	0,3729	1,53	0,4370	1,92	0,4726	3,50	0,4997
0,37	0,1443	0,76	0.2764	1.15	0,3749	1,54	0.4382	1,93	0,4732	4,00	0,4999
0,38	0,1480	0,77	0,2794	1,16	0,3770	1,55	0,4394	1,94	0,4738	5,00	0,5

Таблица 3.

Критические точки $\chi^2_{кр} = \chi^2_{кр}(\alpha; k)$ распределения «Хи–квадрат» с k степенями свободы, удовлетворяющие условию

$$p(\chi^2 > \chi^2_{кр}) = \alpha$$



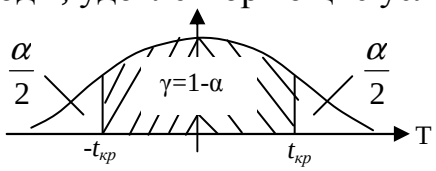
k	α			k	α		
	0,10	0,05	0,01		0,10	0,05	0,01
1	2,71	3,84	6,63	16	23,54	26,30	32,00
2	4,61	5,99	9,21	17	24,77	27,59	33,41
3	6,25	7,81	11,34	18	25,99	28,87	34,81
4	7,78	9,95	13,28	19	27,20	30,14	36,19
5	9,24	11,07	15,09	20	28,41	31,41	37,57
6	10,64	12,59	16,81	21	29,62	32,67	38,93
7	12,02	14,07	18,48	22	30,81	33,92	40,29
8	13,36	15,51	20,09	23	32,01	35,17	41,64

9	14,68	16,92	21,67	24	33,20	36,42	42,98
10	15,99	18,31	23,21	25	34,38	37,65	44,31
11	17,28	19,68	24,72	26	35,56	38,89	45,64
12	18,55	21,03	26,22	27	36,74	40,11	46,96
13	19,81	22,36	27,69	28	37,92	41,34	48,28
14	21,06	23,68	29,14	29	39,09	42,56	49,59
15	22,31	25,00	30,58	30	40,26	43,77	50,89

Таблица 4.

Критические точки $t_{кр} = t_{кр}(\alpha; k)$ распределения Стьюдента (T - распределения) с k степенями свободы, удовлетворяющие условию

$$p(-t_{кр} < T < t_{кр}) = \gamma = 1 - \alpha$$

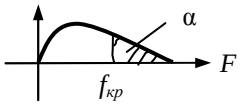


k	γ (α)			k	γ (α)		
	0,90 (0,10)	0,95 (0,05)	0,99 (0,01)		0,90 (0,10)	0,95 (0,05)	0,99 (0,01)
1	6,31	12,71	63,66	9	1,83	2,26	3,25
2	2,92	4,30	9,92	10	1,81	2,23	3,17
3	2,35	3,18	5,84	15	1,75	2,13	2,95
4	2,13	2,78	4,60	20	1,72	2,09	2,85
5	2,02	2,57	4,03	30	1,70	2,04	2,75
6	1,94	2,45	3,71	60	1,67	2,00	2,66
7	1,89	2,36	3,50	120	1,66	1,98	2,62
8	1,86	2,31	3,36	∞	1,64	1,96	2,58

Таблица 5.

Критические точки $f_{кр} = f_{кр}(\alpha; k_1; k_2)$ распределения Фишера–Снедекора (F – - распределения) с $(k_1; k_2)$ степенями свободы, где k_1 – число степеней большей дисперсии, а k_2 – число степеней меньшей дисперсии, удовлетворяющие условию

$$p(F > f_{кр}) = \alpha$$



k_2 \ k_1	α = 0,05																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	240	242	244	246	248	249	250	251	252	253	254
2	18,5	19,0	19,2	19,2	19,3	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	7,71	6,94	6.59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3.63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01

	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
17	4,45	3,59	3,20	2,%	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	4,24	3,39	2,99	2,76	2,60	4,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	4,18	3,33	2,93	2,70	2*155	2,43	2,35	2,28	2,22	2,18	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91	1,83	1,75	0,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	3,84	3,00	3,60	2,37	2,21	2,10	2,01	1,94	1,83	1,83	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

α = 0,01

k_2 \ k_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,49	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65

	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,80	3,66	3,51	3.43	3,35	3,27	3,18	3,09	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	2,96	2,81	2,66	2,58	2,50	2,42	2,33	2,23	2,13
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,90	2,75	2,60	2,52	2t,44	2,35	2,26	2,17	2,06
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,29	2,20	2,П	2,02	1,92	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1.95	1,86	1,76	1,66	1,53	1,38
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2.18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00

Литература

1. Аничин В. Л. Математическая статистика. Харьков, 1992.

2. Гмурман В. Е. Теория вероятности и математическая статистика. М., «Высшая школа», 2001.

3. Иванова В. М. и др. Математическая статистика. М., «Высшая школа», 1975.

4. Колемаев В. А. и др. Теория вероятностей и математическая статистика. М., «Высшая школа», 1991.

5. Кремер Н. Ш. Теория вероятностей и математическая статистика. М., «Юнити», 2002.

6. Румшиский Л. З. Элементы теории вероятностей. М., «Высшая школа», 1976.

Учебное издание
Комогорцев