Федеральное агентство по образованию
Государственное образовательное учреждение
высшего профессионального образования
“Тверской государственный университет“

А.М. Тапилин

ОСНОВЫ ТЕОРИИ ВЕРОЯТНОСТЕЙ

С ЭЛЕМЕНТАМИ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Издание второе, дополненное
Учебно-методическое пособие

Тверь 2006

Книга содержит в основном весь материал программы по теории вероятностей и элементы математической статистики, иллюстрирующие практическое применение вероятностных закономерностей в геоэкологии и смежных с нею науках. Она отличается нетрадиционной структурой изложения. Аналогичные понятия и теоремы сгруппированы с учетом специфики усвоения курса лицами с преимущественно предметным мышлением.

Предназначено для проведения семинаров, аудиторной и самостоятельной работы студентов вузов.

Печатается по решению научно-методического совета факультета географии и геоэкологии Тверского государственного университета
(протокол N4 от 8.04.2006 г.)

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

Настоящее учебное пособие составлено с учетом следующих обстоятельств. На специальности, где математические методы раньше мало использовались, как правило, идут лица с относительно слабой математической подготовкой. Значительная часть студентов, обладая преимущественно предметным мышлением, испытывает затруднения с усвоением сути базовых понятий теории вероятностей и взаимосвязей между ними. Не разобравшись в сути, они без особого успеха пытаются запомнить очередной материал, весьма нечетко представляя его связь с предыдущим. Поэтому каждый последующий материал воспринимается ими все хуже и хуже. При сложившихся обстоятельствах автор посчитал возможным в определенной степени отойти от традиционной структуры изложения дисциплины.

Во многих случаях может оказаться достаточным глубокое осмысление студентами только ключевых понятий и закономерностей теории вероятностей как основы математической статистики (по принципу: “лучше меньше, да лучше”) и способности воспользоваться компьютерными программными средствами (в частности, Excel Windows) для решения вероятностных задач по своей специальности. С одной стороны, это уменьшит нагрузку на память, с другой стороны, позволит сосредоточиться на ключевых моментах.

Соответствующее пособие должно достаточно полно освещать основные положения теории вероятностей, быть самодостаточным и ориентировать студентов на активное и глубокое усвоение материала. Оно должно учитывать специфику освоения курса студентами, в разной степени склонными к аналитическому мышлению. Прежде всего, это касается студентов с преимущественно предметным мышлением.

Оно должно быть предельно сжатым, акцентируя внимание на базовых понятиях и основных теоремах теории вероятностей. Последние должны быть выделены (курсивом или жирным шрифтом). Структура изложения должна как можно лучше отражать взаимосвязи между ними.

Например, теоремы умножения вероятностей представляются как следствия из определений условной вероятности и условия независимости событий. Последовательно излагаются теоремы сложения вероятностей несовместных и совместных событий, а также – независимых в совокупности событий. В одной связке дается определение математического ожидания для дискреных и непрерывных случайных величин, математическое ожидание функции случайного аргумента и условное математическое ожидание, а также – его статистической оценки.

Приводятся примеры взаимосвязанного применения вероятностых закономерностей при решении практических задач, в том числе из области геоэкологии и связанных с нею наук. К ним относится пример совместного использования основных теорем при оценивании вероятностей событий, связанных с риском принятия тех или иных решений. Введение элементов математической статистики дает возможность связать теорию с практикой и, тем самым, более глубоко понять суть и роль изучаемой дисциплины. Это “опредмечивает” базовые понятия теории вероятностей, дает статистическую оценку основных числовых характеристик случайных величин и взаимосвязей между ними. С одной стороны, это упрощает осмысление вероятностных закономерностей лицами, склонными к предметному мышлению, с другой стороны, это способствует восприятию теории вероятностей и математической статистики как единого целого. Последнее благоприятно отразится на последующем изучении методов математической статистики.

Приводятся определения базовых понятий, примеры, задачи, задания и вопросы. Этого достаточно для самостоятельного усвоения основ теории вероятностей. Задачи сопровождаются методическими указаниями по их решению, задания – изложением узловых моментов их выполнения. Для студентов, заинтересованных в более глубоком усвоении теории, приводятся задания повышенной сложности. Студент может не справиться с заданием, то он должен обязательно понять хотя бы смысл соответствующей формулы и знать определения ее элементов. Последнее определяет минимум требований к уровню подготовки студентов. В зависимости от того, в какой степени студенты будут самостоятельно выполнять задания, находится содержание аудиторных занятий.

Настоящее учебно-методическое пособие ориентировано на освоение вероятностной основы статистической обработки результатов наблюдений и статистического выражения числовых характеристик случайных величин, на наиболее рациональную подготовку к экзаменам. Оно отличается нетрадиционной структурой изложения. Прежде всего, оно акцентирует внимание на базовых понятиях и нацелено на составление целостного представления о дисциплине. Его структура составлена, исходя из того, чтобы студентам было легче выделить общее и особенное, а в последующем – упростить изучение методов математической статистики. Вероятностные закономерности иллюстрируются примерами из области геоэкологии и близких к ней наук (геохимии, техногенных систем и экологического риска).

Данное пособие составлено на основе книги: Гмурман В.Е. Теория вероятностей и математическая статистика: Учеб. пособие для вузов. 7-е изд., стер. – М.:Высшая школа, 1999.- 479 с.

А. Тапилин

1. СЛУЧАЙНЫЕ СОБЫТИЯ

События А, В, С,…, которые могут произойти или не произойти при осуществлении тех или иных определенных условий (испытании) – случайные. Если событие обязательно произойдет, то оно достоверное (Ω), если заведомо не произойдет – невозможное (Θ). Если в урне 10 белых шаров, то возможность извлечения наудачу одного белого шара – достоверное событие, иного цвета – невозможное. Мерой возможности наступления события при осуществлении одних и тех же условий, которые могут повторяться неограниченное число раз, является вероятность. Если вероятность события достаточно велика, то его считают практически достоверным. Аналогично, если она незначительна, то в единичном испытании это событие практически не наступит, т.е. оно практически невозможно. *Изучение вероятностных закономерностей массовых однородных случайных событий, т.е. многократно повторяющихся при осуществлении одних и тех же условий – предмет теории вероятностей.*

Если каждому исходу испытаний поставлено в соответствие число, то это означает, что задана случайная величина. В этом случае случайное событие состоит в появлении того или иного числа. При бросании игральной кости может появиться любое из следующих чисел: 1, 2, 3, 4, 5, 6. Появление каждого из них – случайное событие. Каждое из них – значение случайной величины (числа очков), вероятность появления каждого – 1/6. *Соотношение между всеми возможными значениями случайной величины и их вероятностями отражает закон ее распределения.*

Для описания общих свойств распределения вероятностей случайной величины используют числовые характеристики, наиболее употребительными из которых являются математическое ожидание и дисперсия. *Математическое ожидание M(X) характеризует центр распределения случайной величины X, дисперсия D(X) – рассеяние значений случайной величины X относительно ее математического ожидания. Мерой связи случайных величин между собой является коэффициент корреляции.*

Теория вероятностей изучает закономерности, отражающие взаимодействие множества случайных, мало связанных друг с другом факторов. Вероятностным закономерностям могут подчиняться как сами изучаемые явления, так и приемы их исследования, независимо от природы изучаемых объектов. Вероятности событий всегда удовлетворяют некоторым простым соотношениям. Теория вероятностей позволяет по вероятностям одних случайных событий находить вероятности других, в определенном смысле связанных с первыми. Она является теоретической основой математической статистики – систематизации, обработки и использованию статистических данных (о числе объектов в какой-либо их совокупности, обладающих теми или иными признаками, важными для решения поставленной задачи). Математическая статистика дает возможность принимать решения в условиях неопределенности.

Наличие у события определенной вероятности проявляется в том, что почти в каждой достаточно большой серии испытаний относительная частота этого события приблизительно равна его вероятности. Среднее арифметическое наблюдаемых значений случайной величины (выборочное среднее) является статистической оценкой ее математического ожидания. При многократных измерениях физических величин оценка математического ожидания характеризует истинное значение Θ каждой из них, а оценка дисперсии – случайную погрешность измерения.

Вероятностные закономерности получают статистическое выражение в силу закона больших чисел.

1.1. Вероятность и относительная частота

1.1.1. Аксиоматическое определение вероятности.

Современные курсы теории вероятностей построены на аксиоматическом определении случайного события и его вероятности.

Пусть при испытании появляется одно и только одно из событий ω_i $(i = 1, 2, \ldots, n)$. Эти события – элементарные. *Каждый из элементарных исходов, в котором событие A наступает, называют благоприятствующим событию A.* Множество всех элементарных событий, которые могут появиться в испытании, называют пространством элементарных событий Ω. События представляют как подмножества этого пространства. В общем случае пространство элементарных событий может быть любой природы: конечным или бесконечным, дискретным или непрерывным. Подмножество, элементы которого олицетворяют элементарные исходы, благоприятствующие событию A, представляет событие A.

Отношения между событиями интерпретируют как соотношения между подмножествами. В частности, несовместные события – это те, которые не содержат общих элементов (т.е. элементарные события являются попарно несовместными); события, противоположные событию A, включают все элементарные события, не входящие в A. Сумма событий $(A + B)$ эквивалентна такому множеству элементарных событий, которые входят или в A или в B (наступает или A или B или оба этих события вместе, т.е. наступает хотя бы одно из этих событий). Произведение (AB) эквивалентно множеству элементарных событий, которые являются общими для A и B (наступают совместно и событие A и событие B). Событие, включающее все элементарные события – достоверное, не содержащее ни одного элемента – пустое.

Вероятность событий (по Колмогорову А.Н.) определяют аксиомы:

каждому событию A поставлено в соответствие неотрицательное действительное число $P(A)$ – его вероятность;

вероятность достоверного события $P(\Omega) = 1$;

вероятность наступления хотя бы одного из попарно несовместных событий равна сумме вероятностей этих событий.

Свойства вероятностей и зависимостей между ними выводят, исходя из этих аксиом. Последние постулируют существование вероятностей и задают определенные правила действий с ними.

Пример. Пусть исход испытания состоит в появлении одного из попарно несовместных событий: A или B или C. Каждому из них поставлена в соответствие вероятность (первая аксиома): $P(A) = 0,5$, $P(B) = 0,3$, $P(C) = 0,2$. Поскольку сумма событий ($A + B + C$) является достоверным событием, то согласно второй и третьей аксиомам его вероятность

$$P(A + B + C) = P(A) + P(B) + P(C) = 0,5 + 0,3 + 0,2 = 1.$$

1.1.2. Экспериментальное определение вероятности любого события, например i-го значения признака X некоторой совокупности объектов, может быть осуществлено в результате испытаний, выполняемых при одном и том же комплексе условий. Различают выборочную и генеральную совокупности. *Под генеральной совокупностью понимают множество абсолютно всех объектов, обладающих теми или иными признаками, важными для решения поставленной задачи. Случайно отобранную часть генеральной совокупности называют выборочной совокупностью.* Выборочная совокупность должна быть достаточно представительной, чтобы отражать исследуемые свойства генеральной совокупности. Изучение свойств генеральной совокупности по ее части лежит в основе математической статистики.

Относительная частота признака генеральной совокупности – это его вероятность. Относительная частота является базовым понятием не только теории вероятностей, но и математической статистики. Числа наблюдений i-х значений признака X называют частотами n_i. Их отношения к объему совокупности n называют относительными частотами (частостями):

$$\mathbf{W_i = n_i / n}, \quad i = 1, 2, ..., k.. \qquad (1.1а)$$

где

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

Относительная частота – это отношение числа испытаний, в которых событие появилось, к числу всех фактически проведенных испытаний. Ее или число, близкое к ней, определяют как статистическую вероятность.

Недостаток статистического определения вероятности – в ее неоднозначности, колебании ее значений от серии к серии испытаний около фактического значения вероятности. С увеличением числа испытаний (объема выборочной совокупности) можно добиться того, что отклонение относительной частоты от вероятности события будет меньше любого сколь угодно малого положительного числа (она будет сходиться по вероятности к вероятности). Относительная частота – это *статистическая оценка* вероятности, вычисляемая по результатам фактически проведенных испытаний, т.е. *такая функция значений выборочной совокупности, которая в определенном смысле близка к истинному значению оцениваемого параметра*. Относительная частота является хорошей статистической оценкой вероятности.

1.1.3. Классическое определение вероятности

Для конечного числа равновозможных несовместных событий вероятность события А – это отношение числа m исходов, благоприятствующих событию А, к числу n всех возможных элементарных исходов испытания:

$$\mathbf{P(A) = m/n} \qquad (1.1б)$$

Несовместные события – это события, появление любого из которых исключает появление других в том же испытании. Исходы, благоприятствующие событию A – те, в которых событие A наступает.

Пример. В урне 1 белый шар, 2 синих, 3 красных. Извлечение шара наудачу – испытание. Всего – 6 равновозможных несовместных исходов. Пусть событие A состоит в извлечении цветного шара. Ему благоприятствуют 5 исходов, т.е. $P(A) = 5/6$.

Задание. Показать, что вероятность достоверного события $P(\Omega) = 1$, а вероятность невозможного события $P(\Theta) = 0$.

Вопрос. В чем различие между классическим и аксиоматическим определениями вероятности?

1.1.4. Классическое определение вероятности неприменимо к испытаниям с бесконечным числом исходов. Для преодоления этого недостатка введено понятие геометрической вероятности – вероятности попадания точки в область, например, левее любой точки на отрезке прямой, или левее и ниже любой точки – вершины бесконечного квадранта. В частности, вероятность попадания случайной точки на отрезок l, составляющий часть отрезка L, равна отношению длин этих отрезков: l / L.

1.1.5. При вычислении вероятностей часто используют формулы числа комбинаций элементов, связанных определенными условиями:

Число всех возможных *перестановок* – комбинаций из одних и тех же n элементов, отличающихся только порядком расположения последних,

$$P_n = n!,$$

где $n! = 1\cdot 2\cdot \ldots \cdot n$. По определению, $0! = 1$.

Число всех возможных *сочетаний* – комбинаций, составленных из n различных элементов по m элементов, которые отличаются хотя бы одним элементом,

$$C_n^m = n!/(m!(n-m)!).$$

Число всех возможных *размещений* – комбинаций, составленных из n различных элементов по m элементов, которые отличаются составом или порядком элементов,

$$A_n^m = n(n-1)\ldots(n-m+1).$$

Задача. В урне 5 одинаковых шаров. На каждом шаре написана одна из следующих букв: с, п, у, х, е. Шары извлекают наудачу по одному и располагают в порядке извлечения. Найти вероятность того, что составится слово "успех".

Ответ: 1/120 (перестановки).

Задача. В урне 6 одинаковых шаров. На каждом шаре написана одна из следующих букв: с, п, у, к, р, е. По одному извлекают наудачу 4 шары и располагают их в порядке извлечения. Найти вероятность того, что при этом составится слово "курс".

Ответ: 1/360 (размещения).

Задача. Из 10 элементов – 7 стандартных. Найти вероятность появления 4 стандартных элементов среди 6, взятых наугад.

Решение задачи предполагает ответы на следующие вопросы:

в чем состоит испытание (извлечение 6 элементов из 10);

являются ли исходы равновозможными (да);

чему равно число всех несовместных исходов (числу сочетаний из 10 элементов по 6: $C_{10}^6 = 10!/(6!\cdot 4!) = 210$);

в чем состоит событие A (появляются 4 стандартных и 2 нестандартных элемента);

чему равно число исходов, благоприятствующих событию A (произведению числа сочетаний из 7 стандартных элементов по 4 на число сочетаний из 3 нестандартных элементов по 2, т.е. $C_7^4 C_3^2 = 105$).

Ответ. $P(A) = 105/210 = 0{,}5$.

1.1.6. Классическое определение положено в основу элементарной теории вероятностей. Последняя недостаточна уже для самых простых случаев, не вписывающихся в указанные выше ограничения. Однако, в ней наиболее просто определяются основные понятия теории вероятностей. Многие определения и теоремы с незначительными по существу изменениями можно использовать и в общем случае. Наиболее серьезное изменение связано с тем, что случайные события рассматриваются как объединения не конечного, а бесконечного числа исходов, вероятность каждого из которых может быть равной нулю. Поэтому свойство, выраженное в элементарной теории теоремой сложения вероятностей, не выводится из классического определения вероятности (пункт 1.3.1), а само включается в определение (последняя аксиома пункта 1.1.1.).

1.2. Условная вероятность и произведение событий

1.2.1. *Вероятность появления события B при условии, что событие A уже наступило, называют условной вероятностью* $P_A(B)$. По определению, оно равно отношению вероятности $P(AB)$ совместного наступления событий A и B (произведения событий) к вероятности события A:

$$\boldsymbol{P_A(B) = P(AB)/P(A).} \qquad (1.2)$$

Отсюда вытекает теорема умножения вероятностей: *вероятность совместного наступления событий равна произведению вероятности одного из них на условную вероятность другого, вычисленную в предположении, что первое событие уже наступило*:

$$\boldsymbol{P(AB) = P(A)\, P_A(B).} \qquad (1.3)$$

Аналогично,

$$P(BA) = P(B)\, P_B(A),$$

т.е.

$$\boldsymbol{P(A)\, P_A(B) = P(B)\, P_B(A).} \qquad (1.4)$$

Задание. Пусть r – число исходов, благоприятствующих событию B при условии, что событие A уже наступило, $к$ – число исходов, благоприятствующих событию A, n – общее число всех возможных элементарных исходов испытания. Учитывая, что $P(B/A) = r/к$, получить формулу (1.2).

Пример. В урне 3 белых и 7 черных шаров. Вероятность достать сначала белый шар $P(A)$ =3/10, затем черный – $P_A(B)$ = 7/9, а вероятность достать сначала белый, затем черный шары $P(AB)$ =3/10·7/9 = 7/30. Аналогичный результат получится, если использовать формулу (1.2): событию (AB) благоприятствует 21 исход (3·7) из 90 размещений (A_{10}^{2}=10·9=90) – общего числа возможных исходов. Следовательно, $P_A(B)$ = (21/90)/(3/10)

= 7/9.

1.2.2. *Событие В не зависит от события А, если вероятность события В не зависит от того, наступило ли событие А,* т.е.

$$\boldsymbol{P(B) = P_A(B)}. \qquad (1.5)$$

Задание. Показать, что для независимых событий

$$\boldsymbol{P(AB) = P(A)\,P(B)}, \qquad (1.6)$$

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

1.2.3. Несколько событий попарно независимы, если каждые 2 из них независимы. Они независимы (в совокупности), если они и попарно независимы, и независимы каждое из них и возможные произведения остальных. Для таких событий

$$P(A_1 A_2 \ldots A_n) = P(A_1)P(A_2) \ldots P(A_n). \qquad (1.7)$$

Пример. В каждой из 3 урн по 10 шаров. Из них белых шаров в урне 1 – 8, урне 2 – 7, в урне 3 – 9. Из каждой урны вынимают наугад по одному шару. События A_1, A_2 и A_3 – извлечения белого шара из урн 1, 2 и 3, соответственно. Найти вероятность совместного наступления этих событий $P(A_1 A_2 A_3)$.

Решение:

определить, являются ли события A_1, A_2 и A_3 независимыми (да);

выбрать соответствующую независимым событиям формулу (1.7);

найти $P(A_1)$ = 8/10, $P(A_2)$ = 7/10, $P(A_3)$ = 9/10;

найти $P(A_1 A_2 A_3)$ = 0,8·0,7·0,9 = 0,504.

1.3. Вероятность появления хотя бы одного из событий $A_1 A_2 \ldots A_n$

Появление хотя бы одного из событий $A_1 A_2 \ldots A_n$ – это сумма событий $(A_1 + A_2 + \ldots + A_n)$. Сумма событий $(A+B)$ состоит в том, что появляется или событие A, или событие B, или оба эти события вместе (AB). *Несколько событий образуют полную группу, если в результате испытания появится хотя бы одно из них.*

1.3.1. *Для суммы несовместных событий*

$$\boldsymbol{P(A_1 + A_2 + \ldots + A_n) = P(A_1) + P(A_2) + \ldots + P(A_n)}, \qquad (1.8)$$

т.е. вероятность суммы несовместных событий равна сумме вероятностей этих событий. В частности, $P(A+B) = P(A) + P(B)$. Действительно, если n – число всех равновозможных несовместных исходов испытания, m_A и m_B – число исходов, благоприятствующих событиям A и B, то

$$P(A+B) = (m_A + m_B) / n = m_A / n + m_B / n = P(A) + P(B).$$

Пример. В урне 3 белых, 10 красных и 7 черных шаров. Всего – 20 шаров. Вероятность случайно вынуть белый шар – 3/20, черный – 7/20, а не цветной шар (черный или белый) составляет (3/20 + 7/20) = 0,5.

Задание. Сравнить формулу (1.8) с последней аксиомой определения вероятности по Колмогорову А.Н.

Для несовместных событий, образующих полную группу:

$$P(A_1)+P(A_2)+ \ldots +P(A_n) = 1. \tag{1.9}$$

В частности, *если полную группу образуют 2 события, то их называют противоположными (A и $\overline{A}$).* Если $P(A) = p$, $P(\overline{A}) = q$, то

$$\boldsymbol{p + q = 1}. \tag{1.10}$$

Пример. Если при бросании игральной кости выпадет 1 очко, то это событие A. Вероятность противоположного события (или 2, или 3, или 4, или 5, или 6 очков) равна (1 – 1/6) = 5/6.

1.3.2. *Для суммы совместных событий*

$$\boldsymbol{P(A+B) = P(A) + P(B) - P(AB)}, \tag{1.11}$$

т.е. ее вероятность равна сумме вероятностей этих событий без вероятности их совместного появления.

Задание. Подтвердить это, учитывая, что событию $(A+B)$ благоприятствует 3 несовместных события: AB, $A\overline{B}$ и $\overline{A}B$; событию A: AB, $A\overline{B}$; событию B: AB, $\overline{A}B$. Отсюда следует:

$$P(A+B) = P(AB) + P(A\overline{B}) + P(\overline{A}B);$$
$$P(A) = P(AB) + P(A\overline{B}), \text{ откуда } P(A\overline{B}) = P(A) - P(AB);$$
$$P(B) = P(AB) + P(\overline{A}B), \text{ откуда } P(\overline{A}B) = P(B) - P(AB).$$

Пример. Вероятность поражения цели одним стрелком – 0,7, другим – 0,8. Найти вероятность поражения цели хотя бы одним из стрелков.

Решение: по формуле (1.11): $P(A+B) = 0{,}7 + 0{,}8 - P(AB)$, где по формуле (1.6) $P(AB) = 0{,}7 \cdot 0{,}8 = 0{,}56$, так что $P(A+B) = 0{,}94$.

1.3.3. *Для группы независимых в совокупности событий* $A_1, A_2, \ldots, A_n$ вероятность появления хотя бы одного из них равна разности между единицей и произведением вероятностей противоположных событий

$$P(A) = 1 - q_1 q_2 \ldots q_n \qquad (1.12)$$

где события A и $(\overline{A}_1 \overline{A}_2 \ldots \overline{A}_n)$ – противоположны, т.е.

$$P(A) = 1 - P(\overline{A}_1 . \overline{A}_2 \ldots \overline{A}_n) = 1 - P(\overline{A}_1) P(\overline{A}_2) \ldots P(\overline{A}_n).$$

Пример. Вероятность поражения цели одним стрелком – 0,7, другим – 0,8. Найти вероятность поражения цели хотя бы одним из стрелков.

Решение: по формуле (1.12): $P(A+B)$ = 1- (1- 0,7) (1- 0,8) = 0,94.

Задание. Найти вероятность появления только одного из 3 независимых событий A_1, A_2 или A_3, если их вероятности равны p_1, p_2 и p_3 .

Методические указания:

появление только события A_1 равносильно появлению события $B_1 = A_1 \overline{A}_2 \overline{A}_3$ (аналогично определить события B_2 и B_3);

использовать формулу (1.7)) для определения вероятностей каждого из событий B_1, B_2 и B_3 – совместного наступления независимых событий;

искомая вероятность равна вероятности появления любого из несовместных событий B_1, B_2 или B_3 (использовать формулу (1.8)).

1.4. Случаи, когда событие A может наступить лишь при условии появления одного из несовместных событий $B_1, B_2, \ldots, B_n$, образующих полную группу.

1.4.1. *Формула полной вероятности*: вероятность события A, которое может наступить лишь при условии появления одного из образующих полную группу несовместных событий $B_1, B_2, \ldots, B_n$, равна сумме произведений вероятностей каждого из этих событий на соответствующую условную вероятность события A:

$$\boldsymbol{P(A) = P(B_1)P_{B1}(A) + P(B_2)P_{B2}(A) + \ldots + P(B_n)P_{Bn}(A)} \qquad (1.13)$$

Задание. Подтвердить формулу (1.13), учитывая, что событие А состоит в появлении хотя бы одного из несовместных событий $AB_1, AB_2, \ldots, AB_n$:

$P(A) = P(AB_1 + AB_2 + \ldots + AB_n)$, где $P(AB_1) = P(B_1)P_{B1}(A)$ и т.п.

Пример. Вероятность извлечения белого шара из одной урны – 0,8, из другой – 0,9. Найти вероятность извлечения белого шара при случайном выборе урны.

Решение. События B_1 и B_2 – отбор из урн 1 и 2, т.ч. $P(B_1) = P(B_2) = 0{,}5$, а

$P(A)$ = 0,5·0,8 + 0,5·0,9 = 0,85.

1.4.2. *Вероятность гипотез (формулы Бейеса*, позволяющие переоценить вероятности гипотез при появлении события *A)*: условная вероятность одного из образующих полную группу несовместных событий $B_1, B_2,...,B_n$, , при условии, что событие *A*, которое может наступить лишь при появлении одного из этих событий B_i, уже наступило, равна отношению к вероятности события *A* произведения вероятности события B_i на соответствующую условную вероятность события *A*.

$$\boldsymbol{P_A(B_i) = P(B_i)P_{B\,i}(A)/P(A)},\ (i = 1,\ 2,\ \dots,\ n). \qquad (1.14)$$

Вопрос. Вытекает ли формула (1.14) из равенства типа (1.4):

$$P(A)\ P_A(B) = P(B)\ P_B(A)\ ?$$

Пример. Вероятность извлечения белого шара из одной урны – 0,8, из другой – 0,9. Извлеченный шар оказался белым. Найти вероятность того, он извлечен урны 1.

Решение. События B_1 и B_2 – отбор из урн 1 и 2, т.ч. $P(B_1) = P(B_2)$ = 0,5, а $P(A)$ = 0,5·0,8 + 0,5·0,9 = 0,85. Если извлеченный шар оказался белым, то вероятность того, что он извлечен из урны 1: $P_A(B_1)$ = 0,5·0,8/0,85 = 0,47.

1.5. Вероятность того, что в *n* независимых испытаниях событие *A* наступит *к* раз

Испытания называют независимыми относительно события *A*, если вероятность его наступления в каждом испытании не зависит от исходов других испытаний. Событие, состоящее в том, что в *n* независимых испытаниях событие *A* наступит *к* раз – сложное, совмещающее *n* простых событий. Сложные события являются несовместными, если отличаются хотя бы одним из исходов испытания.

Пример. В трех испытаниях возможны следующие сложные события: $AA\bar{A},\ A\bar{A}A,\ \bar{A}AA$. Это несовместные события, так как например, в первом из этих сложных событий появление в третьем исходе испытания простого события, противоположного *A,* исключает возможность появления в этом же исходе испытания события *A,* что имеет место в остальных двух сложных событиях.

1.5.1. Формула Бернулли. Вероятность того, что в *n* независимых испытаниях событие *A* наступит *к* раз

$$P_n(k) = C_n^k\, p^k q^{\,n-k}, \qquad (1.15)$$

где *p* – одинаковая для всех испытаний вероятность события;

$p^{k}q^{n-k}$ – вероятность несовместных сложных событий, в каждом из которых событие A наступает k раз и не наступает $n - k$ раз в серии из n испытаний (совместное наступление независимых событий A и $\overline{A}$);

C_n^k – число сочетаний из n элементов по k, равное числу комбинаций из n различных испытаний.

Задание. Подтвердить, что, если вероятность p превышения среднесуточной нормы значения показателя равна 0,2, то вероятность того, что 2 суток из 4 норма окажется превышенной, составит $P_4(2)$ = 0,054.

1.5.2. Локальная теорема Лапласа (асимптотическое приближение формулы (1.15) для больших n)

$$P_n(k) \approx \varphi(x) / \sqrt{npq}, \qquad (1.16)$$

где p – одинаковая для всех испытаний вероятность события, отличная от нуля и единицы,

$$x = (k - np) / \sqrt{npq},$$

$$\varphi(x) = exp(-x^2/2) / \sqrt{2\pi}.$$

Функция φ(x) четная, т.е. $\varphi(x) = \varphi(-x)$. Она табулирована.
Использование формулы Бернулли связано со сложными вычислениями для большого числа испытаний. Использование формулы (1.16) и таблицы значений функции $\varphi(x)$ значительно упрощает вычисления.

Функция $\varphi(x)$ описывает закон распределения нормированной нормальной случайной величины. Ее значения x вычисляют как отношение отклонения k от np (среднего числа появлений события) к $\sqrt{npq}$ (квадратному корню из дисперсии – меры рассеяния случайной величины относительно среднего значения). Это обстоятельство важно для установления взаимосвязи с последующим изложением курса.

Пример. Вероятность превышения среднесуточной нормы значения показателя – 0,2. Найти вероятность того, что в течение 80 суток из 400 норма окажется превышенной.

Решение. x = (80 – 400·0,2)/$\sqrt{400 \cdot 0{,}2 \cdot 0{,}8}$ = 0. По таблице находим φ(0) = 0,3989. По формуле (1.16) имеем: P400(800) ≈ 0,3989/$\sqrt{400 \cdot 0{,}2 \cdot 0{,}8}$ = 0,0498.

Для приведенного выше примера результаты расчета по формулам (1.15) и (1.16) отличаются только четвертой значащей цифрой. Естественно, при малом числе испытаний использование формулы (1.16) неправомерно.

1.5.3. Формула Пуассона (для редких событий: малых p и больших n)

$$P_n(k) \approx \lambda^{k} e^{-\lambda} / k!, \qquad (1.17)$$

где λ - *среднее число появлений события А в различных сериях испытаний* (*постоянная* $\lambda = np$)

Пример. Вероятность превышения среднесуточной нормы значения показателя – 0,01. Найти вероятность того, что в течение 100 дней норма будет превышена 2 раза.

Решение. $\lambda = np = 100 \cdot 0{,}01 = 1$. $P_{100}(2) \approx \lambda^2 e^{-1} / 2! = 0{,}18$.

1.6. Вероятность наступления события A в n независимых испытаниях от k_1 до k_2 раз (интегральная теорема Лапласа)

$$P_n(k_1, k_2) \approx \Phi(x'') - \Phi(x'), \tag{1.18}$$

где $x' = (k_1 - np) / \sqrt{npq}$, $x'' = (k_2 - np) / \sqrt{npq}$;

p – вероятность наступления события в каждом испытании;

$\Phi(x)$ – функция Лапласа (табулирована для $x > 0$):

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-z^2/2} dz. \tag{1.19}$$

Вопрос. Является ли $\Phi(x)$ четной функцией?

Задание. Подтведить формулу (1.18), учитывая, что

$$P_n(k_1, k_2) = \frac{1}{\sqrt{2\pi}} \int_{x'}^{x''} e^{-z^2/2} dz.$$

Пример. Вероятность превышения среднесуточной нормы значения показателя – 0,2. Найти вероятность того, что из 400 суток норма будет превышена от 70 до 100 суток.

Решение:

$x' = (70 - 400 \cdot 0{,}2) / \sqrt{400 \cdot 0{,}2 \cdot 0{,}8} = -1{,}25$;

$x'' = (100 - 400 \cdot 0{,}2) / \sqrt{400 \cdot 0{,}2 \cdot 0{,}8} = 2{,}5$;

$\Phi(2{,}5) = 0{,}4938$; $\Phi(1{,}25) = 0{,}3944$, $P_{400}(70, 100) = 0{,}4938 + 0{,}3944 = 0{,}8882$.

1.7. Вероятность отклонения относительной частоты m/n от постоянной вероятности p события

Она приближенно равна удвоенной функции Лапласа при $x = \varepsilon\sqrt{\frac{n}{pq}}$:

$$P(|m/n - p| \le \varepsilon) \approx 2\Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right), \tag{1.20}$$

где m – число появлений события, n – число независимых испытаний, ε - заданное положительное число.

Пример. Вероятность события – 0,75. Вероятность того, что относительная частота его появления в 10000 независимых испытаниях отклоняется от его вероятности по абсолютной величине не более, чем на 0,001, равна $\approx 2\Phi(0{,}001\cdot\sqrt{10000/0{,}75/0{,}25}$ = 0,182.

Задание. Преобразовать неравенство $(|m/n - p| \le \varepsilon)$ в равносильное ему двойное неравенство

$$-\varepsilon\sqrt{\frac{n}{pq}} \le (m-np)/(\varepsilon\sqrt{\frac{n}{pq}}) \le \varepsilon\sqrt{\frac{n}{pq}}$$

и получить формулу (1.20), используя формулы (1.18) и (1.19).

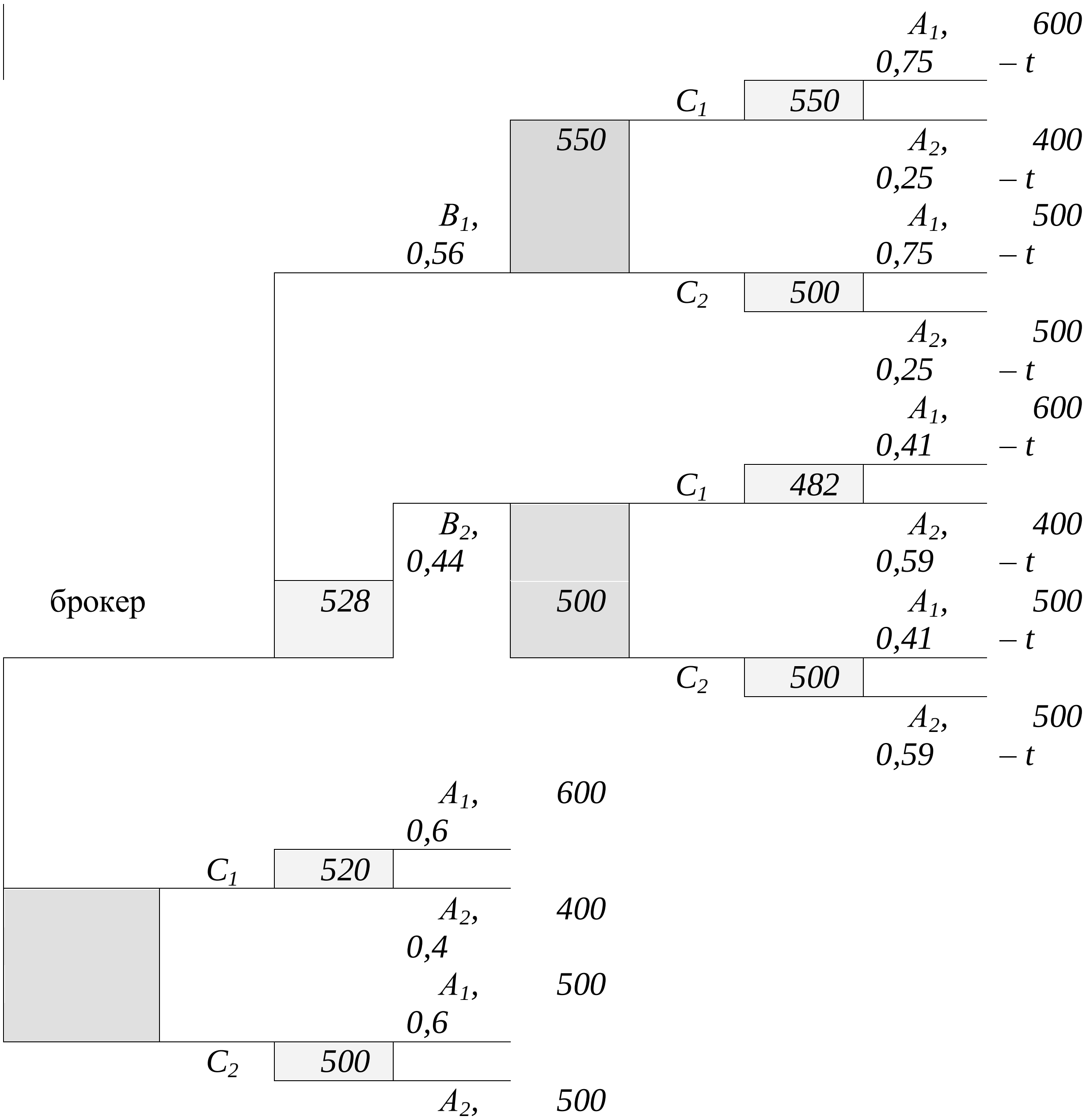
1.8 Метод деревьев

Принятие решений связано с опасностью потерь, обусловленных как расходами на их реализацию, так и последствиями реализации этих решений. Для оценки риска используется метод деревьев. Этот метод отличается удобством и наглядностью графического представления, простотой расчетов на компьютере. Он особенно эффективен, когда сложная проблема может быть расчленена на части, каждая из которых решается отдельно, и на основе этих относительно простых решений можно синтезировать сложное решение. Для пояснения этого метода приведем следующий пример.

Рассматривается вероятность правильного решения о покупке акций, которые могут подорожать или подешеветь, что может привести к выигрышу или потерям при их последующей продаже. Бизнесмен принимает решение сам, но может обратиться за советом к брокеру. Пусть события A_1 и A_2 – возможное повышение или понижение цены акций, соответственно. Вероятности этих событий: $P(A_1)$ = 0.6 и $P(A_2)$ = 0,4. События B_1 и B_2 – совет покупать или не покупать акции, соответственно. Бизнесмен считает, что вероятность правильного совета покупать акции в случае повышения цены $P_{A1}(B_1)$ = 0,70, а вероятность правильного совета не покупать акции в случае понижения цены $P_{A2}(B_2)$ = 0,65. Эти вероятности условные, т.е. бизнесмен считает брокера более способным к выигрышу. События C_1 и C_2 – возможные решения бизнесмена снимать или не снимать деньги на покупку акций в банке. Бизнесмен решает, покупать ли совет брокера и покупать ли акции.

Дерево решений содержит ствол и ветви, в совокупности составляющие единый организм. Дерево решений растет горизонтально: ствол слева, ветви направлены направо (рис. 1.1). Какой бы из ветвей ни следовать, каждый раз надо выбирать между событиями C_1 и C_2. Каждая из новых ветвей дает два отвода: A_1 и A_2. Каждая ветвь отображает или принятие решения, или результат. Нормальный порядок событий – от прошлого к будущему. Точки (на рис. 1.1 – прямоугольники), в которых ветви расщепляются, называют узлами. Выбор одной из ветвей (слева) зависит от лица, принимающего решение после рассмотрения всех возможностей. Соответствующие узлы (решений) обозначены более широкими прямоугольниками.

На пути с ветвями B_1 и B_2 лицо, принимающее решение, не контролирует ситуацию, отображаемую этим узлом (решение принимает брокер). Оно не может выбирать ветвь, но способно приписывать соответствующим событиям некоторые вероятности. Такие узлы называют узлами случаев (они обозначены менее широкими прямоугольниками). При анализе дерева решений с количественными оценками последствий привлекаются два вида величин: вероятности событий и оценки последствий принятых решений. Оценку последствий мы осуществим после, рассматривая понятие математического ожидания случайных величин.



$0{,}4$

Рис. 1.1. Дерево решений

Отходящие от левого узла решений бизнесмена ветви C_1 и C_2, отражают, что обращение к брокеру отвергнуто. Вероятности, характеризующие эти ветви, равны начальным (априорным) значениям. Ветвь "брокер" после узла случая разделяется на две, в соответствии с двумя возможностями: B_1 и B_2 (покупать или не покупать). В этом узле еще не известно, какое из двух событий произойдет (A_1 или A_2). Это не позволяет использовать условные вероятности типа $P_{A1}(B_1)$ или $P_{A2}(B_2)$. События B_1 и B_2 образуют полную группу, т.е. сумма вероятностей ветвей, отходящих от узла случая равна единице: $P(B_1) + P(B_2) = 1$. Согласно формуле полной вероятности в случае повышения цены акций вероятность правильного совета "покупать" равна

$$P(B_1) = P_{A1}(B_1)\, P(A_1) + P_{A2}(B_1)\, P(A_2) = 0{,}7 \cdot 0{,}6 + 0{,}3 \cdot 0{,}4 = 0{,}56,$$

а вероятность правильного совета "не покупать" равна

$$P(B_2) = 1 - P(B_1) = 0{,}44.$$

Продвигаясь по ветвям, отходящим от узла случая, подходят к узлам решений, после каждого из которых разветвляются в соответствии с альтернативой решений C_1 или C_2. Если совет B_1 (покупать) принят, то нужно двигаться по ветви C_1 Она приводит к узлу случая, после которого разветвляется на A_1 и A_2, условные вероятности которых $P_{B1}(A_1)$ и $P_{B1}(A_2)$. Последние отражают вероятности, соответственно, повышения или понижения цены, с учетом того, что принят совет купить акции. Это апостериорные вероятности, т.к. они представляют собой переоценки первичных (априорных) вероятностей событий A_1 и A_2 после того, как произошло событие B_1 (получен и принят совет "покупать").

По формуле Бейеса

$$P_{B1}(A_1) = P(A_1)\, P_{A1}(B_1) / P(B_1) = 0{,}6 \cdot 0{,}7 / 0{,}56 = 0{,}75.$$

Отметим, что условные вероятности $P_A(B_1)$ и $P_A(B_2)$ обладают следующими свойствами:

$$P_{A1}(B_1) + P_{A1}(B_2) = 1,$$
$$P_{A2}(B_1) + P_{A2}(B_2) = 1.$$

Учитывая это, получим:

$$P_{B1}(A_2) = [1 - P_{A2}(B_2)]\, P(A_2) / P(B_1) = 0{,}35 \cdot 0{,}4 / 0{,}56 = 0{,}25.$$

Таким образом, найдены апостериорные вероятности, которыми характеризуются ветви, отходящие от узла случая и соответствующие событиям A_1 и A_2.

Ко второму узлу случая подходит ветвь C_2. После разветвления в узле имеем 2 события A_1 и A_2, апостериорные вероятности которых равны уже вычисленным величинам: 0,75 и 0,25. Полученные данные показывают, что оценка вероятности повышения цены возросла с 0,60 до 0,75 (результат обращения к брокеру и получения совета “покупать”).

Ветвь B_2 (не покупать) после узла решения дает ветви C_1 и C_2. последние подходят к узлам случаев и разветвляются на A_1 и A_2, для которых надо оценить апостериорные вероятности $P_{B2}(A_1)$ и $P_{B2}(A_2)$. По формуле Бейеса:

$$P_{B\,2}(A_1) = P(A_1)\,P_{A1}(B_2)\,/\,P(B_2) = 0{,}41.$$

$$P_{B\,2}(A_2) = P(A_2)\,P_{A2}(B_2)\,/\,P(B_2) = 0{,}59.$$

Таким образом, всем ветвям приписаны соответствующие вероятности.

Следующий этап – *расчет оценок последствий решений* (выгоды, получаемой при следовании по ветвям). Этот этап целесообразно разобрать после освоения понятия математического ожидания. Оценки (в виде денежных сумм) характеризуют конечные точки (терминалы рассматриваемого дерева) и все имеющиеся узлы. На рис. 1.1 приведены оценки полезности последствий решений, причем на узлах и ветвях дерева, образовавшихся после обращения к брокеру, к каждой оценке добавлена стоимость услуги брокера в виде отметки “минус *t*”. Расчет показателей полезности последствий начинается с терминалов дерева по направлению к стволу (справа налево). Для каждого узла случаев вычисляют математическое ожидание показателя, а при подходе к узлу решений проводят максимизацию приписанных к соответствующим ветвям значений показателя (выбирают наибольшие значения).

Математическое ожидание показателя для первого сверху правого узла случаев равно 0,75·600 +0,25·400 = 550, для второго – 0,75·500 + 0,25·500 = 500, для третьего – 0,41·600 + 0,59·400 = 482, для четвертого – 500.

Пара верхних узлов случаев соединяются ветвями C_1 и C_2 с узлом решений, т.е. нужно провести максимизацию. Выбирают большее значение (550, а не 500) и присваивают его данному узлу решений. Ветвь C_2 уже не представляет интереса. Аналогично, показатель второго узла решений – 500 (а не 482). Следование по ветви C_1 нецелесообразно. Математическое ожидание показателя для левого узла случаев (связанного с обращением к брокеру) равно 0,56·550 + 0,44·500 = 528.

Аналогично определяется математическое ожидание показателя для узлов случаев, не связанных с обращением к брокеру. Для верхнего из них оно равно 520, для нижнего – 500. В этом случае показатель узла решений равен 520. Следование по ветви C_2 нецелесообразно. Выбор одной из величин (528 или 520) зависит от стоимости услуги брокера. Если она меньше, чем 528 – 520 = 8, то обращение к брокеру оправдано.

Анализ дерева решений позволяет найти путь снижения риска потери и количественно оценить, насколько он снизится и каков предел дополнительных вложений, которые обеспечат указанное снижение риска. То, что анализ ведут от терминалов к стволу (сначала рассматривают события, которые происходят последними) приводит к фундаментальному выводу: “неоправданно принятие того или иного решения, пока не оценены все возможные последствия этих решений”.

Аналогичный подход можно использовать при управлении не только финансовым, но и другими видами риска, включая экологический.

2. СЛУЧАЙНЫЕ ВЕЛИЧИНЫ И СПОСОБЫ ИХ ЗАДАНИЯ

Случайное событие может состоять в появлении численного значения некоторой величины. *Величину, которая в результате испытания примет одно и только одно из возможных значений, заранее неизвестное, называют случайной величиной.* Случайная величина задана, если каждому исходу испытания поставлено в соответствие число. Дискретная случайная величина принимает отдельные значения, непрерывная – любые, из некоторого промежутка.

Примером случайной величины могут служить те или иные свойства территориальных комплексов, пространственно-временные связи между ними. Эти свойства формируются под влиянием множества разнородных факторов, совокупное влияние которых учесть трудно или практически невозможно. Их модели (статистические) составляют на вероятностной основе. Условие использования наблюдаемой величины в качестве статистической модели – массовость и случайность. В частности, при геохимических съемках условиями наблюдений являются замеры в случайно выбранных точках территориального комплекса. Результаты наблюдений – значения случайной величины. Число возможных замеров практически неограниченно. Число появлений события (значения свойства) в серии наблюдений – его частота. Отношение последней к общему числу наблюдений определяет статистическую вероятность.

Если каждому возможному значению случайной величины X соответствует только одно возможное значение случайной величины Y, то $Y = \varphi(X)$ – *функция случайного аргумента* X. Если каждой паре возможных значений X и Y соответствует одно возможное значение случайной величины Z, то $Z = \varphi(X,Y)$ – *функция двух случайных аргументов* X *и* Y.

Закон распределения случайной величины как соотношение между всеми ее возможными значениями и их вероятностями можно отобразить разными способами: в виде таблицы, графика или аналитически. В виде таблицы можно задать только дискретную случайную величину. *Функция распределения* вероятностей (интегральная), позволяет задать как дискретную, так и непрерывную случайную величину, *плотность* *распределения* вероятностей (интегральная) – только непрерывную случайную величину.

2.1. Дискретная случайная величина как совокупность всех ее возможных значений и их вероятностей

2.1.1. Возможное значение одномерной дискретной случайной величины определяется одним числом. *При табличной форме* задания приводят все ее возможные значения в первой строке таблицы, а во второй – соответственно, их вероятности:

$$\begin{array}{lllll} \boldsymbol{X} & \boldsymbol{x_1} & \boldsymbol{x_2} & \ldots & \boldsymbol{x_n} \\ \boldsymbol{p} & \boldsymbol{p_1} & \boldsymbol{p_2} & \ldots & .\boldsymbol{p_n}. \end{array}$$

На графике точки (x_i, p_i) соединяют отрезками прямых в порядке увеличения значений случайной величины, получая многоугольник распределения.

Перечень наблюдаемых значений признака выборочной совокупности и соответствующих им частот или последовательность интервалов значений и соответствующих им частот (сумм частот, попавших в соответствующий интервал) представляет собой *эмпирическое распределение* (выборки). Последнее отражает собой теоретическое *распределение* (признака генеральной совокупности как случайной величины, вероятности появления каждого значения которой равны относительному числу соответствующих объектов в генеральной совокупности).

Наблюдаемые значения x_i случайной величины X называют *вариантами*. Последовательность вариант в порядке их возрастания образует *вариационный ряд*. Наглядное представление о распределении случайной величины дает *полигон частот* – ломаная, отрезки которой последовательно соединяют точки (x_i; n_i) вариационного ряда (для полигона относительных частот – (x_i; W_i));

Пример полигона частот (рис. 2.1), построенного по вариационному ряду:

x_i	1	2	3	4
n_i	2	4	5	3

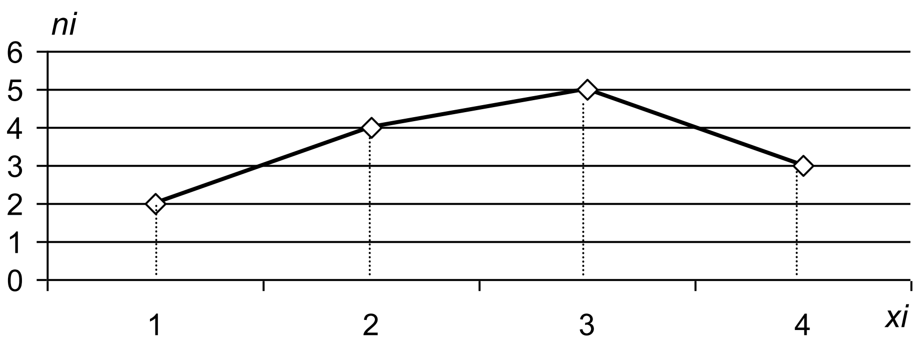


Рис. 2.1

Формулой Бернулли (1.14) *аналитически* описывается биномиальное распределение, формулой (1.16) – распределение Пуассона. Вероятность того, что событие A наступит только в k-ом из независимых испытаний отражает *геометрическое распределение*

$$P(X = к) = q^{k-1} p. \qquad (2.1.)$$

Задание. Используя формулу (1.3) умножения вероятностей независимых событий, вывести закон геометрического распределения.

Вопрос. Что вытекает из того, что события $X = x_1$, $X = x_2$, $,.X =.x_n$ образуют полную группу ?

Произведение постоянной величины C на дискретную случайную величину X – дискретная случайная величина CX, возможные значения которой равны произведениям C на возможные значения X, а вероятности ее возможных значений – вероятностям соответствующих значений X.

Пример задачи. Найти закон распределения случайной величины CX, если C =2, а X имеет следующий закон распределения:

X	2	5	9
p	0,3	0,5	0,2.

Поскольку дискретная случайная величина CX есть произведение постоянной величины C на дискретную случайную величину X, возможные значения которой равны произведениям C на возможные значения X, а вероятности ее возможных значений – вероятностям соответствующих значений X, то

CX	4	10	18
p	0,3	0,5	0,2.

Произведение независимых случайных величин X и Y – случайная величина XY – возможные значения которой равны произведениям каждого возможного значения X на каждое возможное значение Y, а вероятности возможных значений XY – произведениям вероятностей возможных значений сомножителей. Две случайные величины независимы, если закон распределения одной из них не зависит от того, какие возможные значения приняла другая величина. Несколько случайных величин взаимно независимы, если закон распределения любого числа из них не зависит от того, какие возможные значения приняли остальные случайные величины.

Пример задачи. Найти законы распределения функции $Z = XY$, если

X	1	2
p	0,6	0,4

Y	3	4
p	0,8	0,2.

Поскольку случайная величина XY есть произведение независимых случайных величин X и Y, возможные значения которой равны произведениям каждого возможного значения X на каждое возможное значение Y, а вероятности возможных значений XY – произведениям вероятностей возможных значений сомножителей, то получим:

$$\begin{array}{lcccc} XY & 1\cdot 3 & 2\cdot 3 & 1\cdot 4 & 2\cdot 4 \\ p & 0{,}6\cdot 0{,}8 & 0{,}4\cdot 0{,}8 & 0{,}6\cdot 0{,}2 & 0{,}4\cdot 0{,}2. \end{array}$$

Сумма случайных величин X и Y – случайная величина $(X+Y)$, возможные значения которой равны суммам каждого возможного значения X и каждого возможного значения Y, а их вероятности для независимых X и Y равны произведениям вероятностей слагаемых, а для зависимых X и Y – произведениям вероятности одного слагаемого на условную вероятность другого.

Пример задачи. Найти законы распределения функции $Z = X + Y$, если

$$\begin{array}{lcc} X & 1 & 2 \\ p & 0{,}6 & 0{,}4 \end{array} \qquad\qquad \begin{array}{lcc} Y & 3 & 4 \\ p & 0{,}8 & 0{,}2. \end{array}$$

Поскольку случайная величина $X+Y$ есть сумма независимых случайных величин X и Y, возможные значения которой равны суммам каждого возможного значения X и каждого возможного значения Y, а вероятности возможных значений XY – произведениям вероятностей возможных значений сомножителей, то получим:

$$\begin{array}{lcccc} XY & 1+3 & 2+3 & 1+4 & 2+4 \\ p & 0{,}6\cdot 0{,}8 & 0{,}4\cdot 0{,}8 & 0{,}6\cdot 0{,}2 & 0{,}4\cdot 0{,}2. \end{array}$$

2.1.2. *Если возможные значения случайной величины определяются n числами, то это n-мерная случайная величина*. В частности, (X,Y) – двумерная случайная величина.

Совокупность всех возможных значений (x_i, y_j) $(i = 1, 2, \dots, n;\ j = 1, 2, \dots, m)$ дискретной двумерной случайной величины и их вероятностей $p(x_i, y_j)$ называют законом ее распределения.

В соответствующей таблице первая строка содержит все возможные значения X, первый столбец – все возможные значения Y, на пересечении строк и столбцов – вероятности $p(x_i, y_j)$ того, что случайная величина примет значение (x_i, y_j).

Таблица 2.1

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

Y_j	$p(x_1, y_j)$	$p(x_2, y_j)$		$p(x_i, y_j)$		$p(x_n, y_j)$
…						
y_2	$p(x_1, y_m)$	$p(x_2, y_m)$		$p(x_i, y_m)$		$p(x_n, y_m)$

Сумма вероятностей, помещенных во всех клетках таблицы, равна единице. Зная закон распределения двумерной случайной величины, можно найти закон распределения каждой из составляющих. По теореме сложения вероятностей несовместных событий:

$$P(x_i) = p(x_i, y_1)+ p(x_i, y_2)+\ldots + p(x_i, y_m), \qquad (2.1.)$$

т.е. *вероятность того, что X примет значение* x_i, *равна сумме вероятностей "столбца* x_i*".*

Задание. Найти законы распределения составляющих двумерной случайной величины *(X,Y)*, заданной следующим законом распределения:

Y	X		
	x_1	x_2	x_3
y_1	0,1	0,2	0,3
y_2	0,06	0,16	0,18

Проконтролировать правильность выполнения задания.

2.1.3. Вероятность того, что *X* примет значение x_i при условии, что $Y = y_j$ определяет *условный закон распределения составляющей X*

$$\mathbf{p(x_i|y_j) = p(x_i,\ y_j) \,/\, p(y_j)}. \qquad (2.2)$$

Аналогично, условный закон распределения составляющей *Y* при условии, что $X= x_i$ находят исходя из того, что

$$p(y_j|x_i) = p(x_i,,y_j) \,/\, p(x_i). \qquad (2.3)$$

Задание. Провести аналогию между условным законом распределения составляющей *X* и условной вероятностью события *A* при условии, что событие *B* уже наступило.

Задание. Учитывая формулу (2.1), доказать, что *сумма вероятностей условного распределения равна единице.*

Задание. Найти условный закон распределения составляющей *X* двумерной случайной величины *(X,Y)*, если составляющая *Y* приняла значение y_1, а закон распределения (*X,Y)* задан таблицей 2.2:

Таблица 2.2

Y	X		
	$x_1=1$	$x_2=2$	$x_3=3$

y_1=4	0,1	0,2	0,3
y_2=5	0,06	0,16	0,18

2.2. Функция распределения вероятностей случайной величины

2.2.1. Эта функция *определяет для каждого значения x случайной величины X вероятность того, что X примет значение, меньшее x:*

$$\boldsymbol{F(x) = P(X < x)} \qquad (2.4)$$

Обратная задача заключается в том, чтобы найти для заданной вероятности соответствующее значение случайной величины X (квантиль, отвечающую заданному уровню вероятности).

Генеральную совокупность можно рассматривать как случайную величину.

Функция распределения генеральной совокупности фактически является функцией распределения случайной величины. Ее называют *теоретической функцией распределения. Для каждого значения x признака X она определяет вероятность события X < x.* В отличие от нее *эмпирическая функция распределения определяет относительную частоту этого события:*

$$\boldsymbol{F^*(x) = n_x / n,} \qquad (2.4a)$$

где n_x – число вариант, меньших x; n – объем совокупности. *Эмпирическая функция распределения служит для оценки теоретической функции распределения.*

Примером выборочной совокупности может служить случайно отобранная часть студентов 2-го курса (генеральной совокупности), признаком – их успеваемость по математической статистике. По данным выборки можно судить о генеральной совокупности (о доле студентов курса, имеющих оценки менее указанной).

Пример. Дискретная случайная величина

X	2	5	9
p	0,3	0,5	0,2

имеет следующую функцию распределения

$$F(x) = \begin{cases} 0 & \text{при} \quad x \le 2 \\ 0{,}3 & \text{при} \quad 2 < x \le 5 \\ 0{,}8 & \text{при} \quad 5 < x \le 9 \\ 1 & \text{при} \quad x > 9. \end{cases}$$

Действительно, событие $X < 2$ – невозможно, т.е. его вероятность равна нулю. Меньше любого числа из диапазона $2< x\le 5$ только одно значение X, равное 2. Его вероятность – 0,3. Меньше любого числа из диапазона $5< x\le 9$ случайная величина X может принять два значения: или 2 с вероятностью 0,3 или 5 с вероятностью 0,5. Эти события несовместны. Вероятность их суммы равна сумме их вероятностей – 0,8. То, что все значения X меньше любого из диапазона $x > 9$ – событие достоверное, т.е. его вероятность равна 1. Наглядное представление о распределении дискретной случайной величины дает график ее фукции распределения (рис. 2.2).

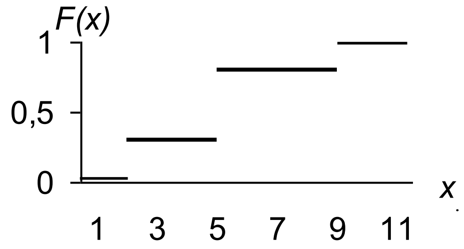


Рис. 2.2.

Задание. Построить теоретическую функцию распределения:

варианты x_i	1	3	8	10
частоты n_i	6	8	14	1

Эмпирическая функция распределения относительных частот успешных исходов испытаний студентов одной из групп приведена на рис. 2.3.

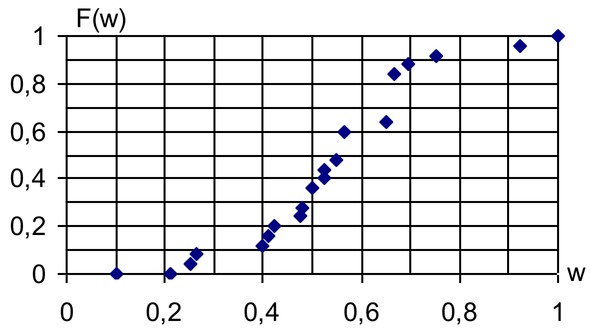


Рис. 2.3

Задание. Построить эмпирическую функцию распределения относительных частот успешных исходов испытаний студентов той или иной группы по фактическим данным преподавателя.

2.2.2. Задание. Подтвердить свойства функции распределения.

а) Значения функции распределения удовлетворяют двойному неравенству:

$$\mathbf{0 \le \mathit{F(x)} \le 1}. \qquad (2.5)$$

б) $F(x)$ – неубывающая функция, т.е.

$$\boldsymbol{F(x_2) \geq F(x_1)}, \text{ если } \boldsymbol{x_2 > x_1} \qquad (2.6)$$

(представить событие ($X < x_2$) как ($X < x_1$) + ($x_1 \leq X < x_2$), чтобы получить: $F(x_2) - F(x_1) = P(x_1 \leq X < x_2)$.)

Следствие 1. *Вероятность того, что случайная величина примет значение, заключенное в интервале (a,b), равна приращению функции распределения на этом интервале*, т.е.

$$\boldsymbol{P\ (a \leq X < b) = F(b) - F(a).} \qquad (2.6а)$$

Следствие 2. Вероятность того, что непрерывная случайная величина примет одно определенное значение, равна нулю, *т.е.*

$$P\ (a \leq X < b) = P\ (a < X < b) = P\ (a < X \leq b) = P\ (a \leq X \leq b).$$

(представить $P\ (x_1 \leq X < x_1 + \Delta x)$ как $F(x_1 + \Delta x) - F(x_1)$, где $\Delta x \to 0$.)

в) Если возможные значения случайной величины принадлежат интервалу (a,b), то

$$F(x) = 0 \text{ при } x \leq a,\ F(x) = 1 \text{ при } x \geq b\ , \qquad (2.7)$$

Следствие. Если возможные значения непрерывной случайной величины расположены на всей оси x, то имеют место следующие предельные соотношения:

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

Задача. Построить график функции распределения случайной величины

X	1	3	5
P	0,2	0,5	0,3 ,

предварительно заполнив следующую форму:

$$F(x) = \begin{array}{lll} 0 & \text{при} & x \leq 1 \\ \ldots.. & \text{при} & 1 < x \leq 3 \\ \ldots.. & \text{при} & 3 < x \leq 5 \\ \ldots.. & \text{при} & x > 5. \end{array}$$

2.2.3. *Функция распределения двумерной случайной величины (X, Y) – это функция F(x, y), определяющая для каждой пары чисел (x, y) вероятность того, что X примет значение, меньшее x, и при этом Y примет значение, меньшее y*:

$$\boldsymbol{F(x,y) = P(X < x,\ Y < y).} \qquad (2.8)$$

Она определяет вероятность попадания случайной точки (X,Y) в бесконечный квадрант, расположенный левее и ниже его вершины (x, y).

2.2.4. Задание. Подтвердить свойства функции распределения двумерной случайной величины.

а) значения функции распределения удовлетворяют двойному неравенству:

$$0 \le F(x,y) \le 1; \tag{2.9}$$

б) $F(x,y)$ – неубывающая функция по каждому аргументу, т.е.

$$F(x_2, y) \ge F(x_1, y), \text{ если } x_2 > x_1, \qquad (2.10)$$
$$F(x, y_2) \ge F(x, y_1), \text{ если } y_2 > y_1;$$

в) имеют место следующие предельные соотношения:

$$F(-\infty, y) = 0, \qquad F(x, -\infty) = 0, \qquad (2.11)$$
$$F(-\infty, -\infty) = 0, \qquad F(\infty, \infty) = 1;$$

г) при $y = \infty$ функция распределения двумерной случайной величины становится функцией распределения составляющей X:

$$\boldsymbol{F(x, \infty) = F_1(x)}. \tag{2.12}$$

Аналогично

$$\boldsymbol{F(\infty, y) = F_2(y)}. \tag{2.13}$$

(Так как событие $X < \infty$ достоверно, то $F(\infty, y)$ определяет вероятность события $Y < y$, т.е. является функцией распределения составляющей Y.)

2.2.5. Задание. Показать, что вероятность попадания случайной точки в полуполосу

$$P\,(x_1 \le X < x_2,\ Y < y) = F(x_2, y) - F(x_1, y), \tag{2.14}$$

Учесть, что геометрически полуполоса представляет собой ту часть квадранта с вершиной (x_2,y), на которую не накладывается квадрант с вершиной (x_1, y).

2.2.6. Задание. Показать, что вероятность попадания случайной точки в прямоугольник

$$P\,(x_1 \le X < x_2,\ y_1 \le Y < y_2) = [F(x_2, y_2) - F(x_1, y_2)] - [F(x_2, y_1) - F(x_1, y_1)]$$

2.2.7. Задание. Сопоставить более строгие определения дискретной и непрерывной случайных величин.

Случайная величина дискретна, если существует конечное или счетное *множество таких чисел $x_1, x_2, \ldots, x_n$, что*

$$P(X = x_i) = p_i \geq 0,\ i = 1,2,\ldots \text{ и } p_1 + p_2 + \ldots = 1.$$

Случайная величина непрерывна, если существует такая неотрицательная функция $f(x)$, что при любых x функцию распределения можно представить в виде

$$F(x) = \int_{-\infty}^{x} f(y)dy$$

(переменную интегрирования обозначают через y, чтобы не путать с пределом интегрирования).

2.3. Плотность распределения вероятностей непрерывной случайной величины

2.3.1. Для непрерывной случайной величины $P(X = x) = 0$, но это не означает, что событие $X = x$ невозможно. Имеет смысл рассматривать вероятность попадания непрерывной случайной величины в сколь угодно малый интервал, задавая *плотность распределения* её вероятностей (дифференциальную функцию):

$$\boldsymbol{f(x) = F'(x)}. \qquad (2.15)$$

Задание. Используя формулу (2.6а), показать, что вероятность попадания непрерывной случайной величины в заданный интервал (a,b):

$$\boldsymbol{P(a < X < b) = \int_a^b f(x)dx}. \qquad (2.16)$$

Пример. Если $f(x) = 2x$ при $0{,}5 < x \leq 1$, то

$$P(0.5 < x < 1) = 2\int_{0,5}^{1} x dx = x^2\Big|_{0,5}^{1} = 0{,}75.$$

Задание. Используя формулу (2.16), показать, что

$$\boldsymbol{F(x)} = \int_{-\infty}^{x} f(x)dx\,. \qquad (2.17)$$

Задание. Подтвердить, что

$$f(x) \geq 0 \text{ и } \int_{-\infty}^{\infty} f(x)dx = 1 \qquad (2.18)$$

(аналогичные зависимости справедливы для всех плотностей распределения).

Задание. Используя определение плотности распределения

$$f(x) = \lim_{\Delta x \to 0} \frac{F(x+\Delta x) - F(x)}{\Delta x}, \qquad (2.19)$$

объяснить, чему равно её значение в точке x. Покажите, что

$$\boldsymbol{F(x+\Delta x) - F(x) \approx f(x)\Delta x}, \qquad (2.20)$$

т.е. *вероятность того, что случайная величина X примет значение, принадлежащее интервалу $(x,\ x+\Delta x)$, приблизительно равна произведению плотности вероятности в точке x на длину интервала Δx.*

Наглядное представление о распределении случайной величины дает *гистограмма частот* – последовательность примыкающих друг к другу прямоугольников, ширина которых равна ширине интервалов группировки (частичных интервалов) h по оси x, а высота определяется отношением n_i / h числа наблюдений n_i, попавших в соответствующий интервал, к ширине частичных интервалов h (для гистограмм относительных частот – отношением W_i / h). *Площадь гистограммы частот равна их сумме, т.е. объему выборки, площадь гистограммы относительных частот равна единице.*

Пример гистограммы относительных частот (рис. 2.4) для непрерывного признака, интервал значений которого (1; 15) разбит на 7 частичных интервалов длиною $h = 2$ (W_i / h – высоты прямоугольников, h – основание):

Частичный интервал	1-3	3-5	5-7	7-9	9-11	11-13	13-15
сумма частот	4	8	14	20	16	6	2

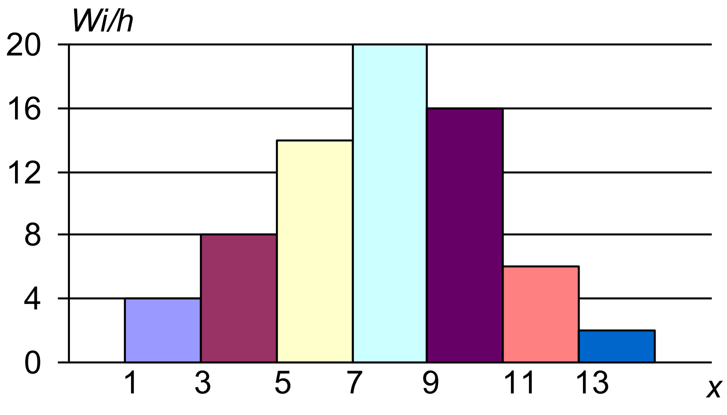


Рис. 2.4.

Обычная группировка включает 10-20 интервалов, в каждый из которых попадает не более 15-20% значений x_i. Этого достаточно для выявления существенных свойств распределения и надежного вычисления его основных характеристик. Группировка с маленькими интервалами дает многовершинную гистограмму, не отражающую наглядно свойств распределения. Группировка с большими интервалами может привести к искаженному представлению о характере распределения.

2.3.2. Задание. По данной $f(x)$ найти $F(x)$ и построить её график:

$$f(x) = \begin{cases} 0 & \text{при } x \le a, \\ 1/(b-a) & \text{при } a < x \le b, \\ 0 & \text{при } x > b \end{cases} \qquad (2.21)$$

(учесть, что $(-\infty, x) = (-\infty, a) + (a, x)$ и $(x > b) = (-\infty, b) + (b, x)$).

Показать, что формула (2.21) задает *закон равномерного распределения* (для него плотность распределения сохраняет постоянное значение C на всем интервале (a, b), которому принадлежат все возможные значения случайной величины, т.е. $\int_a^b C dx = 1$).

2.3.3. Теорема. Если $Y = \varphi(X)$ и $y = \varphi(x)$ –дифференцируемая строго возрастающая или строго убывающая функция, обратная функция которой $\varphi^{-1}(y) = x = \psi(y)$, то плотность распределения случайной величины Y:

$$g(y) = f[\psi(y)]\,|\psi'(y)|. \qquad (2.22)$$

Пример. Найти $g(y)$,если $Y = X^2$, $f(x) = \exp(-x^2/2)/\sqrt{2\pi}$.

Решение. $\psi(y) = x = y^{0,5}$, $f(\psi(y)) = \exp(-y/2)/\sqrt{2\pi}$, $\psi'(y) = y^{-0,5}/2$, т.е. $g(y) = \exp(-y/2)/\sqrt{8\pi y}$.

2.3.4. Аналогичны формулам (2.15) и (2.17) формулы, определяющие *плотность $f(x,y)$ и функцию $F(x,y)$ совместного распределения вероятностей двумерной случайной величины*

$$\mathbf{f(x, y) = \partial^2 F(x, y) / \partial x \partial y} \text{ и } \mathbf{F(x, y)} = \int_{-\infty}^{y}\int_{-\infty}^{x} \mathbf{f(x, y)\, dxdy}. \qquad (2.23)$$

Функцию $f(x,y)$ можно рассматривать как предел отношения вероятности попадания случайной точки в прямоугольник со сторонами Δx и Δy к площади этого прямоугольника, когда обе стороны последнего стремятся к нулю. Тогда вероятность попадания случайной точки в произвольную область D

$$P((X, Y) \subset D) = \iint_D f(x, y)\, dxdy. \qquad (2.24)$$

Аналогично формулам (2.18) имеем: $f(x, y) \ge 0$, $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\, dxdy = 1$.

2.3.5. Согласно формулам (2.25) и (2.25а) *плотность распределения одной из составляющих двумерной случайной величины равна* *несобственному интегралу с бесконечными пределами от плотности совместного распределения, причем переменная интегрирования соответствует другой составляющей*.

Плотность распределения составляющей X

$$\boldsymbol{f_1(x)} = dF_1(x)/dx = dF(x,\infty)/dx = \frac{d}{dx}\int_{-\infty}^{x}\int_{-\infty}^{\infty} f(x,y)dxdy = \int_{-\infty}^{\infty} \boldsymbol{f(x,y)dy}. \qquad (2.25)$$

Задание. Подтвердить, что

$$\boldsymbol{f_2(y)} = \int_{-\infty}^{\infty} \boldsymbol{f(x,y)\,dx}. \qquad (2.25а)$$

2.3.6. Условия независимости случайных величин

Для того, чтобы случайные величины X и Y были независимыми необходимо и достаточно, чтобы функция распределения системы (X,Y) была равна произведению функций распределения составляющих:

$$\boldsymbol{F(x,y)=F_1(x)\,F_2(y)}. \qquad (2.26)$$

Это необходимо, т.к. если X и Y – независимые случайные величины, то события $(X < x)$ и $(Y < y)$ независимы, т.е.

$$P(X < x, Y < y) = P(X < x)\,P(Y < y) \text{ или } F(x,y)=F_1(x)\,F_2(y).$$

Этого достаточно, т.к. из равенства $F(x,y)=F_1(x)\,F_2(y)$ вытекает, что

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

Задание. Подтвердить как следствие из формулы (2.26), что *для того, чтобы X и Y были независимыми, необходимо и достаточно следующее*:

$$\boldsymbol{f(x,y) = f_1(x)\,f_2(y)}. \qquad (2.26а)$$

(Для доказательства необходимости этого условия надо взять производные от функции (2.26) сначала по x, затем по y. Для доказательства достаточности надо проинтегрировать функцию (2.26а) по x и y.)

2.3.7. Теорема. Если $Z = X + Y$, где X и Y – непрерывные независимые случайные величины, то

$$g(z) = \int_{-\infty}^{\infty} f_1(x)\,f_2(z-x)dx \quad \text{или} \quad g(z) = \int_{-\infty}^{\infty} f_1(z-y)\,f_2(y)dy, \qquad (2.27)$$

где f_1 и f_2 – дифференцируемые функции аргументов, из которых хотя бы одна задана на интервале $(-\infty, \infty)$ одной формулой.

2.3.8. Дифференциальную функцию суммы независимых случайных величин называют композицией. Её закон распределения называют устойчивым, если он тот же, что и у составляющих сумму.

Пример. Найти композицию $Z = X+Y$, если $f_1(x)=e^{-x/3}/3$ $(0 \le x < \infty)$; $f_2(y)=e^{-y/4}/4$ $(0<=y<\infty)$.

Решение. $G(z) = \int_0^\infty f_1(x)f_2(z-x)dx = \frac{1}{12}\int_0^\infty e^{-x/3-(z-x)/4}dx = \int_0^\infty e^{-z/4}e^{-x/12}d(x/12) = e^{-z/4}$

(убедитесь, что $\int_0^\infty g(z)dz = 1$).

2.3.9. *Условной плотностью распределения составляющей X при данном значении Y = y называют отношение совместного распределения f(x, y) непрерывной двумерной случайной величины (X,Y) к плотности распределения составляющей Y*

$$\varphi(x \mid y) = f(x, y) / f_2(y); \tag{2.28}$$

Аналогично:

$$\Psi(y \mid x) = f(x, y) / f_1(x). \tag{2.29}$$

Задание. Проведите аналогию между условной вероятностью события A при условии, что событие B уже наступило:

$$P_A(B) = P(AB) / P(A),$$

и условной плотностью распределения составляющей X при данном значении $Y = y$ (формула 2.28).

Задание. Проведите аналогию между формулой (2.28), определяющей условный закон распределения составляющей X двумерной случайной величины (X,Y), и формулой (2.2), определяющей вероятность того, что X примет значение x_i при условии, что $Y = y_j$:

$$p(x_i \mid y_j) = p(x_i, y_j) / p(y_j).$$

3. ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ СЛУЧАЙНЫХ ВЕЛИЧИН И ИХ СТАТИСТИЧЕСКИЕ ОЦЕНКИ

Случайная величина может быть представлена как законом ее распределения, так и с помощью ее числовых характеристик (последнее особенно важно, если неизвестен закон распределения). Числовые характеристики описывают случайную величину суммарно, что бывает достаточно для решения многих задач. Выделяют числовые характеристики положения и рассеяния. Наиболее важной характеристикой положения является математическое ожидание. Оно равно среднему значению случайной величины. Рассеяние значений случайной величины относительно ее математического ожидания характеризует, прежде всего, дисперсия. Для оценки взаимосвязи между случайными величинами обычно используют корреляционные моменты, коэффициенты корреляции и уравнения регрессии.

Признаки генеральной совокупности можно рассматривать как случайные величины. Числовые характеристики признаков – это числовые характеристики соответствующих случайных величин. Их оценки находят по выборке – случайно отобранной части генеральной совокупности. Оценки числовых параметров – это функции выборочных значений случайной величины, которые в определенном статистическом смысле близки к истинному значению оцениваемых параметров.

Может быть оценка:

смещенная и *несмещенная* (если при любом объеме выборки ее математическое ожидание равно истинному значению параметра: $M\theta_n^* = \theta$);

несостоятельная и *состоятельная* (при $n \to \infty$ для любого $\xi > 0$ сходиться по вероятности к истинному значению параметра: $P(|\theta_n^* - \theta| \le \xi) \to 1$);

неэффективная и *эффективная* (иметь минимальную дисперсию в определенном классе оценок).

Относительная частота является несмещенной, состоятельной и эффективной оценкой вероятности события. Подставляя относительную частоту в определение (формулу) математического ожидания, получают статистическую оценку последнего (*выборочную среднюю*). Подставляя относительную частоту и выборочную среднюю в определение дисперсии, получают ее статистическую оценку – *выборочную дисперсию*. Подставляя выброчную дисперсию и выборочную среднюю в определения коэффициентов корреляции или регрессии, получают их статистические оценки (*выборочный коэффициент корреляции* или *выборочный коэффициент регрессии*).

3.1. Характеристики положения

3.1.1. *Математическое ожидание M(X)* является центром распределения случайной величины *X. Для дискретной случайной величины*

$$\boldsymbol{M(X)} = \sum_{i=1}^{n} x_i p_i , \qquad (3.1)$$

где n-число всех ее возможных значений x_i, а p_i- вероятности последних.

Математическое ожидание – это *начальный теоретический момент порядка к:*

$$\nu_k = M(X^{\kappa}) \qquad \mathbf{(3.2)}$$

при $\kappa = 1$

Задание. Пусть x и p –число и вероятность появления события в одном испытании. Используя табличную форму задания случайной величины, доказать, что

$$M(X) = p \qquad \textbf{(3.3)}$$

(в данном случае событие или происходит (x = 1) или не происходит).

3.1.2. *Условное математическое ожидание дискретной случайной величины Y при X = x (одно из возможных значений X) – это произведение всех возможных значений Y на их условные вероятности:*

$$\mathbf{M(Y/X = x_i) =} \sum_{j=1}^{m} y_j p(y_j / x_i). \qquad (3.4)$$

Условное математическое ожидание M(Y / x) = f(x) называют функцией регрессии Y на X.

Аналогично, $M(X / y) = \varphi(x)$ – функция регрессии X на Y.

Задание. Проведите аналогию между математическим ожиданием и условным математическим ожиданием для дискретной случайной величины (формулы 3.10 и 3.11).

Задача. Найдите условное математическое ожидание составляющей Y дискретной случайной величины (XY) при $X = x_1$, если $p(x_i, y_j)$ заданы таблично:

Y	X			
	$x_1 = 1$	$x_2 = 3$	$x_3 = 4$	$x_4 = 8$
y_1=3	0,15	0,06	0,25	0,04
y_2=6	0,30	0,10	0,03	0,07

Ответ: 5.

3.1.3. Задание. Рассмотреть свойства математического ожидания (формулы 3.5-3.8). Используя табличную форму задания случайной величины, доказать следующее:

математическое ожидание постоянной величины С равно самой постоянной, т.е.

$$\mathbf{M(C) = C;} \qquad (3.5)$$

математическое ожидание произведения CX постоянной на случайную величину равно произведению постоянной на математическое ожидание случайной величины, т.е.

$$\mathbf{M(CX) = C\,M(X)}. \qquad (3.6)$$

Задание. Пусть независимые случайные величины заданы следующими законами распределения:

X	5	2	4		Y	7	9	
P	0,6	0,1	0,3		P	0,8	0,2	

Найти *математические ожидания произведения независимых случайных величин X и Y*

$$\boldsymbol{M(XY) = M(X)\,M(Y)} \tag{3.7}$$

и их суммы

$$\boldsymbol{M(X+Y) = M(X) + M(Y)} \tag{3.8}$$

Задание. Доказать, что математическое ожидание числа появлений события A в n независимых испытаниях

$$M(X) = np \tag{3.9}$$

(учесть, что случайная величина X равна сумме числа появлений события A в каждом из n испытаний, а также учесть формулы (3.3) и (3.8)).

Задание. Подтвердить, что *для среднего арифметического* $\overline{(X)}$

$$\overline{(X)} = (X_1+X_2+\ldots+X_n)/\,n$$

одинаково распределенных, взаимно независимых случайных величин

$$\boldsymbol{M\overline{(X)} = a},$$

где a– математическое ожидание, одинаковое для всех этих величин.

3.1.4. Оценку математического ожидания количественного признака X можно получить, подставив вместо вероятностей их статистические оценки ($W_i = n_i\,/\,n$ вместо p_i) в определение математического ожидания дискретной случайной величины X . Несмещенной, состоятельной и эффективной оценкой вероятности p является относительная частота. В частности,

$$M(W) = M[m\,/\,n] = M(m)\,/\,n = np\,/\,n = p. \tag{3.10}$$

Следовательно,

$$M(X) \approx \frac{1}{n}\sum_{i=1}^{n} x_i n_i = \bar{x}_B, \tag{3.11}$$

т.е. *выборочное среднее* $\bar{x}_B$ *является статистической оценкой математического ожидания*. Выборочное среднее является начальным эмпирическим моментом порядка *к:*

$$M_к \approx \frac{1}{n}\sum_{i=1}^{n} x_i^{\,к} n_i. \tag{3.12}$$

при $к = 1$.

Для генеральной совокупности математическое ожидание значений того или иного признака - случайной величины X равно среднему арифметическому этих значений.

$$M(X) = \frac{1}{N}\sum_{i=1}^{n} x_i N_i = \bar{x}_Г, \quad (3.13)$$

Выборочная средняя – несмещенная и состоятельная оценка генеральной средней. Если все значения количественного признака *X* генеральной или выборочной совокупности разбиты на несколько групп, то среднее арифметическое значений признака, принадлежащих группе, называют групповой средней. *Общая средняя* $\bar{x}$ *равна средней арифметической групповых средних* $\bar{x}_1$ *и* $\bar{x}_2$, *взвешенной по объемам групп* n_1 *и* n_2, т.е.

$$\bar{x} = (\bar{x}_1 n_1 + \bar{x}_2 n_2) / (n_1 + n_2).$$

Задание. Найти общую среднюю по данным таблицы:

Группа	1		2	
Значение признака	1	6	1	5
Частота n_{ji}	10	15	20	30

(Ответ: $\bar{x}$ =3,6).

Задание. Докажите, что сумма произведений отклонений $(x_i - \bar{x})$ на соответствующие частоты равна нулю:

$$\sum n_i (x_i - \bar{x}) = 0$$

3.1.5. *Для непрерывной случайной величины X (a ≤X<b):*

$$\mathbf{M(X)} = \int_a^b x f(x) dx$$

Задание. Провести аналогию с дискретной случайной величиной, разбив интервал $(a,\ b)$ на n частичных интервалов, выбрав в каждом из них произвольное значение x_i $(i=1,\ 2,\dots,\ n)$ и составив сумму произведений возможных значений x_i на вероятности их попадания в частичный интервал Δx_i:

$$\Sigma\, x_i\, f(x_i)\, \Delta x_i.$$

Условное математическое ожидание составляющей Y непрерывной случайной величины (XY) при X = x:

$$\mathbf{M(Y/X = x)} = \int_{-\infty}^{\infty} \boldsymbol{y\ \psi(y/x) dy},$$

где ψ(y/x) – условная плотность распределения вероятностей случайной величины Y при X = x.

Аналогично,

$$M(X/Y = y) = \int_{-\infty}^{\infty} x\,\varphi(x/y)\,dx,$$

где $\varphi(x/y)$ – условная плотность распределения вероятностей случайной величины X при $Y = y$.

Задание. Проведите аналогию между условным математическим ожиданием для дискретной и для непрерывной случайной величины.

3.1.6. Другие характеристики положения случайной величины:

среднее геометрическое

$$G(X) = exp\,(M(lnX));$$

среднее гармоническое

$$H(X) = 1/M(1/X),\ X>0;$$

медиана MeX– квантиль распределения случайной величины X, соответствующая вероятности 0,5;

мода MoX– наиболее вероятное значение для дискретной случайной величины; точка максимума плотности распределения – для непрерывной случайной величины.

3.2. Характеристики рассеяния случайной величины

3.2.1. *Дисперсия D(X) случайной величины X – это математическое ожидание квадрата отклонения случайной величины от ее математического ожидания (центральный теоретический момент порядка к = 2):*

$$\boldsymbol{D(X) = M(X - M(X))^{\kappa = 2} = \mu_{\kappa = 2}}. \tag{3.14}$$

Для дискретной случайной величины

$$\boldsymbol{D(X)} = \sum_{i=1}^{n} (x_i - M(X))^2 p_i. \tag{3.15}$$

Для непрерывной случайной величины

$$D(X) = \int_{-\infty}^{\infty} (x - M(X))^2 f(x)\,dx. \tag{3.16}$$

Задание. Проведите аналогию между формулами (3.15) и (3.16).

3.2.2. Квадратный корень из дисперсии случайной величины – это ее *среднее квадратическое отклонение*

$$\boldsymbol{\sigma_x} = \sqrt{D(X)}, \tag{3.17}$$

а отношение среднего квадратического отклонения к математическому ожиданию случайной величины – ее *коэффициент вариации*:

$$V_x = 100\sigma_x / M(X). \qquad (3.18)$$

3.2.3. Задание. Преобразовать формулу (3.15), возведя отклонение в квадрат, чтобы получить формулу, удобную для вычисления дисперсии:

$$D(X) = M(X^2) - M^2(X). \qquad (3.19)$$

3.2.4. Задание. Используя свойства математического ожидания, подтвердить следующие свойства дисперсии:

$$\boldsymbol{D(C) = 0} \quad \text{и} \quad \boldsymbol{D(CX) = C^2 D(X),} \qquad (3.20)$$

т.е. *дисперсия постоянной равна нулю, а постоянный множитель можно вынести за знак дисперсии, возведя его в квадрат*.

3.2.5. Задание: Подтвердить: если X и Y – независимые случайные величины, то

$$\boldsymbol{D(X \pm Y) = D(X)+D(Y)} \qquad (3.21)$$

(использовать формулу (3.8) и представить Y как + (-1)Y).

3.2.6. Задание. Подтвердить:

дисперсия события A в одном испытании равна pq;

дисперсия события A в n независимых испытаниях, в каждом из которых вероятность p появления события A одинакова, равна npq:

$$D(X) = npq \qquad (3.22)$$

(учесть, что $X = X_1 + X_2 + \ldots + X_n$, где X_i – число появлений события A в i-м испытании; использовать формулу (3.21)).

Очевидно, что дисперсия оценки

$$D(W) = D[m / n] = D(m) / n^2 = npq / n^2 = pq / n. \qquad (3.23)$$

3.2.7. Задание. Подтвердить: *дисперсия среднего арифметического* $\overline{(X)}$ *одинаково распределенных, взаимно независимых* n *случайных величин* X_i

$$\boldsymbol{D\overline{(X)} = D / n,} \qquad (3.24)$$

где $D = D_i$ *– дисперсия, одинаковая для* $i = 1, 2, \ldots, n$.

3.2.8. Задание. Подтвердить: среднее квадратическое отклонение суммы n взаимно независимых случайных величин равно квадратному корню из суммы квадратов средних квадратических отклонений этих величин:

$$\sigma(X_1 + X_2 + \ldots + X_n) = (\sigma^2(X_1) + \sigma^2(X_2) + \ldots + \sigma^2(X_n))^{0,5}. \qquad (3.25)$$

3.2.9. Статистические оценки генеральной дисперсии.

Оценку дисперсии количественного признака X можно получить, подставив в определение дисперсии дискретной случайной величины X (формулу 3.15) вместо вероятностных характеристик (p_i и $M(X)$) их оценки ($W_i = n_i / n$ и $\bar{x}$). Получим:

$$D_в(X) = \frac{1}{n}\sum (x_i - \bar{x})^2 n_i, \qquad (3.26)$$

где $D_в(X)$ – *выборочная дисперсия – среднее арифметическое квадратов отклонения наблюдаемых значений* x_i *признака* X *от их среднего значения* $\bar{x}$ *- выборочной средней.* Выборочная дисперсия является центральным эмпирическим моментом порядка *к:*

$$m_k = \frac{1}{n}\sum (x_i - \bar{x})^k n_i \qquad (3.27)$$

при $к=2$.

Аналогично, генеральной дисперсией $D_г$ называют среднее арифметическое квадратов отклонений значений признака генеральной совокупности от их среднего значения $\bar{x}_Г$

$$D_Г(X) = \frac{1}{N}\sum (x_i - \bar{x}_Г)^2 N_i. \qquad (3.28)$$

Выборочная дисперсия является смещенной оценкой генеральной дисперсии, так как

$$M(D_в(X)) = \frac{n-1}{n} D_Г = \sigma^2(X)\frac{n-1}{n}, \qquad (3.29)$$

Чтобы получить несмещенную оценку генеральной дисперсии, надо обе части равенства (3.33) умножить на $\frac{n}{n-1}$.

Исправленная выборочная оценка дисперсии генеральной средней

$$s_x^2 = \frac{n}{n-1} D_в(X) = \frac{1}{n-1}\sum (x_i - \bar{x})^2 n_i. \qquad (3.30)$$

Квадратный корень из выборочной дисперсии называют выборочным средним квадратическим отклонением $\sigma_в = \sqrt{D_в}$*.Квадратный корень из исправленной дисперсии называют исправленным средним квадратическим отклонением* s.

Задание. Используя определение выборочной дисперсии (формула 3.26) выведите формулу, удобную вычисления дисперсии

$$D = \overline{x^2} - (\bar{x})^2. \qquad (3.31)$$

Задание. Найти выборочную дисперсию признака X:

x	1	2	3	4
n	20	15	10	5

Ответ: 1.

Если все значения количественного признака X выборочной или генеральной совокупности разбиты на несколько групп, то *дисперсию значений признака, принадлежащих группе, относительно групповой средней называют групповой дисперсией*

$$D_j = \frac{1}{n_j} \sum (x_{ij} - \bar{x}_j)^2 n_{ij}, \tag{3.32}$$

где n_j – объем группы j; x_{ij} – i-е значение j-й группы признака; $\bar{x}_j$ - групповая средняя; n_{ij} – частота i-го значения j-й группы признака.

Внутригрупповой дисперсией называют среднюю арифметическую групповых дисперсий, взвешенную по объемам групп n_j

$$D_{вг} = \frac{1}{n} \sum n_j D_j \tag{3.33}$$

Межгрупповой дисперсией называют дисперсию групповых средних относительно общей средней

$$D_{мг} = \frac{1}{n} \sum n_j (\bar{x}_j - \bar{x})^2 \tag{3.34}$$

Общей дисперсией называют дисперсию значений признака всей совокупности относительно общей средней

$$D_{общ} = \frac{1}{n} \sum n_i (x_i - \bar{x})^2, \tag{3.35}$$

где n– объем всей совокупности; n_i- частота значения x_i; $\bar{x}$ - общая средняя.

Если совокупность состоит из нескольких групп, то *общая дисперсия*

$$D_{общ}. = D_{вг} + D_{мг}. \tag{3.36}$$

4. ЗАКОНЫ РАСПРЕДЕЛЕНИЯ СЛУЧАЙНЫХ ВЕЛИЧИН

4.1. Закон больших чисел

Теоремы, определяющие условия, при которых суммарное поведение большого числа случайных величин почти утрачивает случайный характер и становится закономерным, называют законом больших чисел. К ним относятся теоремы Чебышева и Бернулли.

4.1.1. По теореме Чебышева, если X_1 X_2 ....X_n попарно независимые случайные величины и их дисперсия не превышает постоянного числа C, то, как бы ни было мало $\varepsilon \geq 0$,

$$\lim_{n\to\infty} P[\,|(X_1+X_2+\ldots+X_n)/n-(M(X_1)+M(X_2)+\ldots+M(X_n))/n\,|<\varepsilon] =1,$$

т.е. при стремлении числа случайных величин к бесконечности предел вероятности того, что модуль отклонения среднего арифметического случайных величин от среднего арифметического их математических ожиданий окажется меньше сколь угодно малого положительного числа, равен единице.

Задание. Подтвердить: если все случайные величины имеют одно и то же математическое ожидание а, то

$$\lim_{n\to\infty} P[\,|(X_1+X_2+\ldots+X_n) / n - a\,|< \varepsilon] = 1. \qquad (4.1)$$

При указанных условиях среднее арифметическое достаточно большого числа результатов повторных измерений (независимых случайных величин) сколь угодно мало отличается от истинного значения измеряемой величины. Такие условия соблюдаются, если результат каждого измерения не зависит от результатов остальных измерений, если нет систематических ошибок измерений, если обеспечивается определенная точность измерений.

4.1.2. По теореме Бернулли, если в каждом из n независимых испытаний вероятность появления события одинакова, то

$$\lim_{n\to\infty} P[\,|m / n - p\,| < \varepsilon] = 1, \qquad (4.1а)$$

т.е. при стремлении числа испытаний к бесконечности предел вероятности того, что модуль отклонения относительной частоты от постоянной вероятности события окажется меньше сколь угодно малого положительного числа, равен единице.

Нельзя утверждать, что с ростом числа испытаний относительная частота неуклонно стремится к вероятности. Можно говорить лишь о вероятности того, что при достаточно большом числе испытаний относительная частота будет сколь угодно мало отличаться от постоянной вероятности появления события в каждом испытании. Поэтому в отличие от сходимости в смысле обычного анализа вводят понятие сходимости по вероятности. Теорема Бернулли утверждает, что при $n \to \infty$ относительная частота события стремится по вероятности к вероятности этоготсобытия.

Задание. Перейти от формулы (4.1) к формуле (4.1а), учитывая, что случайные величины X_i $(i = 1, 2,\ldots, n)$ – число появлений события в i–м испытании попарно независимы, их дисперсии ограничены ($pq \le 0{,}25$), математическое ожидание числа появлений события в одном испытании $M(X_i) = a = p$, т.е. равно вероятности наступления события.

4.1.3. Выборочный метод

На теореме Чебышева основан выборочный метод статистического исследования всей совокупности объектов, обладающих тем или иным общим признаком (генеральной совокупности), по сравнительно небольшой ее части (выборочной совокупности).

Выборочная совокупность (выборка) – это результат последовательных и независимых наблюдений над случайной величиной, представляющей генеральную совокупность. Цель обработки результатов наблюдений – определение вида распределения или вычисление статистических оценок неизвестных параметров распределения определенного вида. Под статистическими данными подразумевают сведения о числе объектов, принадлежащих какой-либо более или менее обширной совокупности и обладающих теми или иными признаками, важными для решения поставленной задачи. Методы исследования, опирающиеся на рассмотрение статистических данных о тех или иных совокупностях объектов, называют статистическими. Их общие черты: подсчет числа объектов, входящих в те или иные группы; рассмотрение распределения количественных признаков; применение выборочного метода; использование теории вероятностей при определении достаточности числа наблюдений для тех или иных выводов. Эта формальная сторона статистических методов исследования, безразличная к специфической природе изучаемого объекта, составляет предмет математической статистики.

Для решения практических задач нужна не сама генеральная совокупность, а лишь те или иные характеристики, которые ставятся ей в соответствие. Последние являются числовыми или функциональными характеристиками некоторого распределения вероятностей. Конкретный набор выборочных значений $x_1, x_2,..., x_n$ признака генеральной совокупности X рассматривают как реализацию (одну из многих) многомерной случайной величины $\mathbf{X} = (X_1, X_2,..., X_n)$, компоненты которой независимы и имеют одну и ту же функцию распределения $F(x)$, соответствующую генеральной совокупности.

Теорему Чебышева применяют при обработке результатов измерений $(x_1, x_2,..., x_n)$, каждый из которых рассматривают соответственно как реализацию случайных величин $X_1, X_2,... X_n$. Так как каждая случайная величина $X_1, X_2,..., X_n$ имеет одну и ту же функцию распределения (генеральной совокупности), то ее числовые параметры (математическое ожидание a, дисперсия D и пр.) также одинаковы. Выборочный метод используется для определения оценок этих параметров.

В частности, $M(X) = \bar{x}_г = a,$ математическое ожидание средней арифметической $M(\bar{X}_B) = M[(X_1+ X_2 +...+ X_n) / n] = a$, т.е. выборочная средняя $\bar{x}_в$ является несмещенной оценкой генеральной средней $\bar{x}_г$, т.е.

$M(\bar{X}_в) = \bar{x}_г$. Допуская, что случайные величины $X_1, X_2, \ldots, X_n$ имеют ограниченные дисперсии, в силу теоремы Чебышева можно утверждать, что при увеличении n среднее арифметическое $\bar{X}_в$ рассматриваемых величин стремится по вероятности к математическому ожиданию каждой из этих величин, т.е. к $a = \bar{x}_г$. Это доказывает, что выборочная средняя является и состоятельной оценкой генеральной средней.

При выборках малого объема точечные оценки могут привести к грубым ошибкам. В таких случаях необходимы интервальные оценки. Пусть найденная по данным выборки статистическая характеристика Θ^* служит оценкой неизвестного параметра Θ. Если $\delta > 0$ и $|\Theta - \Theta^*| < \delta$, то δ характеризует точность оценки. Вероятность γ, с которой осуществляется это неравенство называют надежностью или доверительной вероятностью. Если

$$P[|\Theta - \Theta^*| < \delta] = \gamma,$$

то

$$P[\Theta^* - \delta < \Theta < \Theta^* + \delta] = \gamma,$$

т.е. вероятность того, что интервал $(\Theta^* - \delta, \ \Theta^* + \delta)$ заключает в себе неизвестный параметр Θ равен γ. Этот интервал называют доверительным.

4.1.4. По теореме Ляпунова (центральная предельная теорема), если случайная величина представляет собой сумму очень большого числа взаимно независимых случайных величин, влияние каждой из которых на всю сумму ничтожно мало, то она имеет распределение, близкое к нормальному.

4.2. Нормальное распределение

Нормальное распределение $N(a,\sigma)$ непрерывной случайной величины определяется двумя параметрами: математическим ожиданием $M(X)=a$ и средним квадратическим отклонением $\sigma = \sqrt{D(X)}$.

4.2.1. Плотность распределения вероятностей нормальной случайной величины

$$f(x) = \exp\left[-\left(\frac{x - M(X)}{\sigma}\right)^2 / 2\right] / \sigma\sqrt{2\pi}, \qquad (4.2)$$

Задание. Определить максимум плотности вероятностей нормального распределения, исследуя функцию на экстремум.

Ответ: при $x = a$ функция $f(x)$ имеет максимум, равный $\dfrac{1}{\sigma\sqrt{2\pi}}$.

4.2.2. Функция распределения вероятностей нормальной случайной величины

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{z-a}{\sigma}\right)^2} dz. \qquad (4.3)$$

4.2.3. Задание. Найти *M(X)* нормальной случайной величины *X*.

Методические указания:

ввести переменную **z** = $\frac{x - a}{\sigma}$ в формулу

$$M(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-\frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2} dx;$$

учесть, что $x - a = dz,\ dx = \sigma dz;$

учесть, что интеграл от нечетной функции равен нулю, если пределы интегрирования симметричны относительно начала координат;

учесть, что

$$\int_{-\infty}^{\infty} e^{-z^2/2} dz = \sqrt{2\pi};$$

4.2.4. Задание. Найти *D(X)* нормальной случайной величины *X*.

Методические указания: проинтегрировать функцию

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

по частям ($\int_{-\infty}^{\infty} u dv = uv\Big|_{-\infty}^{+\infty} - \int_{-\infty}^{\infty} v du$), принимая $u = z,\ dv = ze^{-z^2/2} dz$.

4.2.5. Если *X* – нормальная случайная величина, то

$$Z = \frac{(X - a)}{\sigma} \qquad (4.4)$$

- нормированная нормальная случайная величина. Нормирование случайной величины X изменяет начало отсчета и масштаб измерения. Нормированная случайная величина безразмерна и не зависит от выбора масштаба измерения. Единицей ее измерения становится σ.

4.2.6. Задание. Используя свойства *M(X),* подтвердить, что для нормированной случайной величины *Z*

$$M(Z) = 0,\ D(Z) = 1. \qquad (4.5)$$

4.2.7. Плотность распределения вероятностей нормированной нормальной случайной величины

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \tag{4.6}$$

функция распределения

$$F_0(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-z^2/2} dz. \tag{4.7}$$

На рис. 4.1 представлены результаты расчета функции и плотности распределения вероятностей нормированной нормальной случайной величины

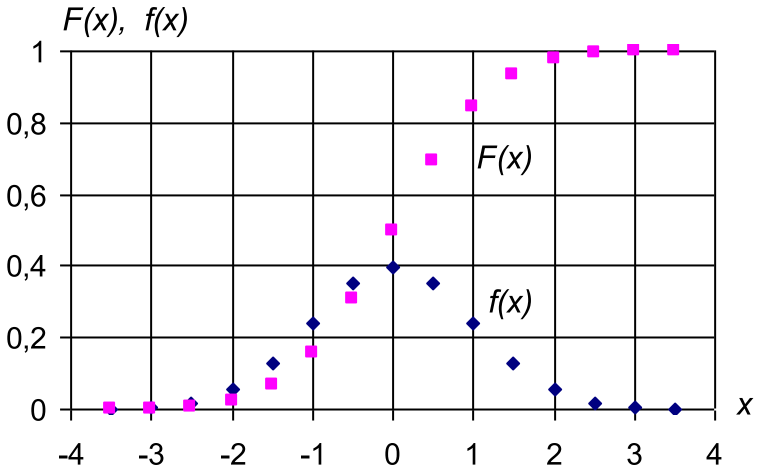


Рис. 4.1.

Задание. Подтвердить формулу (4.6), используя формулы (4.2, 4.4 и 4.5).
Задание. Подтвердить, что

$$\boldsymbol{F(x) = F_0((x - a) / \sigma)}$$

Задание. Подтвердить, что

$$P(0 < X < x) = \Phi(x),$$

где $\Phi(x)$ - функция Лапласа.

Задание. Подтвердить, что $\int_{-\infty}^{0} \varphi(x)dx = 0{,}5$,

т.е.

$$F(x) = 0{,}5 + \Phi(x).$$

4.2.8. Задание. Подтвердить, что вероятность попадания нормальной случайной величины в заданный интервал равна разности функций Лапласа от большего и меньшего нормированных значений аргументов - границ интервала:

$$P(\alpha < x < \beta) = \Phi(\frac{\beta - a}{\sigma}) - \Phi(\frac{\alpha - a}{\sigma}). \tag{4.8}$$

Методические указания:

ввести переменную $z = \frac{x-a}{\sigma}$, получая $x = \sigma z + a$, $dx = dz$,

установить новые пределы интегрирования, в частности, при $x = \alpha$ получая $z = \frac{(\alpha - a)}{\sigma}$;

представить заданный интервал как сумму $(\frac{(\alpha - a)}{\sigma}, 0) + (0, \frac{(\beta - a)}{\sigma})$.

4.2.9: Вероятность того, что модуль отклонения нормально распределенной случайной величины от ее математического ожидания меньше заданного положительного числа *δ*, равна удвоенной функции Лапласа от аргумента δ/σ

$$P(|X - a| < \delta) = 2\Phi(\frac{\delta}{\sigma}), \tag{4.9}$$

Задание. Подтвердить формулу (4.9), заменяя неравенство в скобках на равносильное: $a - \delta < x < a + \delta$ и применяя формулу (4.8).

4.2.10. Для проверки соответствия эмпирического распределения нормальному теоретическому распределению используют специальные характеристики случайной величины: асимметрию и эксцесс. Для нормального распределения они равны нулю.

Асимметрия теоретического распределения

$$\boldsymbol{A_S = \mu_3 /\sigma^3} = M(X - M(X))^3/\sigma^3 = M[(X - M(X)) /\sigma]^3 = \mu_3 (Z). \tag{4.10}$$

Она положительна, если длинная часть кривой – справа от *M(X)*. Она равна нулю для симметричных распределений.

Эксцесс теоретического распределения

$$\boldsymbol{E_к = \mu_4 /\sigma^4 - 3} = M[(X - M(X)) /\sigma]^4 - 3 = \mu_4 (Z) - 3. \tag{4.11}$$

Он положителен или отрицателен в зависимости от того, больше или меньше максимум плотности рассматриваемого распределения по сравнению с нормальным (при тех же математическом ожидании и дисперсии).

4.2.11. Нормальное распределение является устойчивым. Математическое ожидание и дисперсия композиции нормальных законов распределения соответственно равны суммам математических ожиданий и дисперсий слагаемых.

4.2.12. Распределения, связанные с нормальным (*хи-квадрат*, *Стьюдента, Фишера*) обычно используются для оценки доверительных интервалов и проверки статистических гипотез.

4.2.13. Доверительные интервалы для оценки с надежностью γ математического ожидания нормально распределенного признака *X* при известном σ.

Рассматривая выборочные значения $x_1, x_2,..., x_n$ признака X как одинаково распределенные независимые случайные величины $X_1, X_2, ..., X_n$, математическое ожидание каждой из которых равно a, а дисперсия – D, для выборочной средней имеем: $M(\overline{X}) = a,\ \ \sigma(\overline{X}) = \sigma / \sqrt{n}$. Это связано с тем, что выборочная средняя как случайная величина, найденная по независимым наблюдениям за нормально распределенной случайной величиной, также распределена нормально.

В соответствии с неравенством $P[|\ \Theta\ -\ \Theta^*| < \delta] = \gamma$ потребуем выполнение соотношения

$$P[|\ \overline{X} - a\ | < \delta] = \gamma,$$

где γ – заданная надежность, δ – точность. Заменяя X на $\overline{X}$ и σ на $\sigma(\overline{X}) = \sigma / \sqrt{n}$ в формуле (4.9), получим:

$$P(\left|\overline{X} - a\right| < \delta) = 2\Phi(\frac{\delta\sqrt{n}}{\sigma}) = 2\Phi(t),$$

где $t = \delta\sqrt{n}/\sigma$. Так как $\delta = t\ \sigma\sqrt{n}$, то $P(\left|\overline{X} - a\right| < t\sigma / \sqrt{n}) = 2\Phi(t)$. При заданной надежности γ и $\overline{X} = \overline{x}$ (одной из реализаций $\overline{X}$)

$$P\ (\overline{x} - t\sigma / \sqrt{n} < a < \overline{x} + t\sigma / \sqrt{n}) = 2\Phi(t) = \gamma,$$

т.е. доверительный интервал $(\overline{x} - t\sigma / \sqrt{n}, x + t\sigma / \sqrt{n})$ покрывает с надежностью γ неизвестный параметр a. Точность оценки $\delta = t\ \sigma\sqrt{n}$, где t определяется из равенства $2\Phi(t) = \gamma$ по таблице функции Лапласа.

4.3. Проверка статистических гипотез

Гипотезы о параметрах известного распределения или о виде неизвестного распределения называют статистическими.

Проверка статистической гипотезы необходима, например, при выборе метода или средства измерений, обеспечивающих наибольшую точность (т.е. наименьщую дисперсию, наименьшее рассеяние результатов измерений). В этом случае сравнивают исправленные выборочные дисперсии, найденные по независимым выборкам из генеральных совокупностей X и Y. Известно, что закон распределения этих совокупностей – нормальный, т.е. он определяется двумя параметрами: математическим ожиданием и дисперсией.

Рассматривают выдвинутую гипотезу и противоречащую ей. Если выдвинутая гипотеза будет отвергнута, то принимается противоречащая ей. Выдвинутую гипотезу H_0 называют нулевой, противоречащую ей H_1 – конкурирующей (альтернативной). Если будет отвергнута правильная гипотеза, то это *ошибка первого рода.* Вероятность совершить такую ошибку обозначают через α – уровень значимости. Обычно его принимают равным 0,05 или 0,01. *Ошибка второго рода* состоит в том, что будет принята неправильная гипотеза. Вероятность такой ошибки – β.

В рассматриваемом примере требуется при заданном уровне значимости α проверить нулевую гипотезу о равенстве генеральных дисперсий:

$$H_0 : D(X) = D(Y).$$

Поскольку исправленные дисперсии являются несмещенными оценками генеральных дисперсий, то можно записать:

$$H_0 : M(S^2_X) = M(S^2_Y),$$

Если нулевая гипотеза справедлива, то различие между исправленными дисперсиями незначимо и объясняется случайными причинами.

Для проверки нулевой гипотезы используют *статистический критерий* - специально подобранную случайную величину, точное или приближенное распределение которой известно. Значение критерия, вычисленное по выборкам, называют наблюдаемым значением $K_{набл.}$ Вероятность попадания критерия в критическую область при условии, что справедлива конкурирующая гипотеза, называют *мощностью критерия*. Она равна 1- β.

В качестве критерия проверки нулевой гипотезы в рассматриваемом примере используют случайную величину $F = S_б^2/S_м^2$, где $S_б^2$ – большая из дисперсий. При условии справедливости нулевой гипотезы она имеет распределение Фишера (пункт 4.3.3) со степенями свободы $к_б = n_б - 1$ и $к_м = n_м - 1$, где индексы указывают на объемы выборок с большей и меньшей дисперсией.

После выбора критерия множество всех его возможных значений разбивают на 2 непересекающихся подмножества. Одно из них содержит совокупность значений критерия, при которых нулевую гипотезу отвергают. Это *критическая область*. Другое – содержит совокупность значений критерия, при которых гипотезу принимают. Это *область принятия гипотезы.*

Точки, отделяющие критическую область от области принятия гипотезы, называют критическими точками $к_{кр}$ (границами). Критическую область, определяемую неравенством $К > к_{кр}$, где $к_{кр}$ – положительное число, называют правосторонней. Левосторонней называют критическую область, определяемую неравенством $К < к_{кр}$, где $к_{кр}$ – отрицательное число. Двусторонняя критическая область определяется неравенствами $К < к_1$, $К > к_2$, где $к_2 > к_1$ Если критические точки симметричны относительно нуля, то двусторонняя критическая область определяется неравенством $|К| > к_{кр.}$ *Критическая область строится в зависимости от вида конкурирующей гипотезы* H_1.

Для отыскания критических областей достаточно найти критические точки. Сначала задаются уровнем значимости α. Затем ищут критические точки, исходя из требования, чтобы *при условии справедливости нулевой гипотезы* выполнялись следующие требования.

для правосторонней критической области – вероятность того, что критерий $К$ примет значение, большее $к_{кр,}$ была бы равна принятому уровню значимости:

$$P(К > к_{кр}) = α.;$$

для левосторонней критической области: $P(К < к_{кр}) = α.$;

для двусторонней критической области: $P(К < к_1) + P(К > к_2) = α.$.

Для каждого критерия имеются соответствующие таблицы, по которым находят критические точки, удовлетворяющие предъявляемым требованиям. В частности, если для правосторонней критической области окажется, что $К > к_{кр}$, то нулевую гипотезу отвергают; если $К < к_{кр}$, то нет оснований отвергнуть нулевую гипотезу. Это обусловлено тем, что вероятность события $К > к_{кр}$ мала. При справедливости нулевой гипотезы такое событие в силу принципа невозможности маловероятных событий не должно наступить в единичном испытании. Если оно наступило (наблюдаемое значение критерия оказалось больше критического), то это свидетельствует о том, что нулевая гипотеза ложная и должна быть отвергнута. Тем не менее, *если нулевая гипотеза принята, то это не значит, что она доказана.*

Если конкурирующей является гипотеза H_1: $D(X) > D(Y)$, то строят правостороннюю критическую область. Исходят из требования, чтобы вероятность попадания критерия в эту область в случае справедливости нулевой гипотезы была равна уровню значимости:

$$P[F > F_{КР}(α;\ k_{1,}\ k_2) = α.$$

Критическую точку $F_{КР}(α;\ k_{1,}\ k_2)$ находят по таблице критических точек распределения Фишера. Правосторонняя критическая область определяется неравенством $F > F_{КР}$. Если $F_{набл} < F_{кр}$,, где $F_{набл}$ – отношение большей исправленной дисперсии к меньшей, то нет оснований отвергнуть нулевую гипотезу.

4.4. Распределение *“хи*-квадрат”

Распределение *“хи*-квадрат” имеет случайная величина

$$\chi^2 = \sum_{i=1}^{n} X_i^2, \qquad (4.15)$$

где X_i – нормальные независимые случайные величины, у которых

$$M(X_i) = 0, \quad \sigma(X_i) = 1.$$

Это распределение определяется одним параметром – числом степеней свободы $k = n$. Если X_i связаны одним линейным соотношением (например, если $\sum X_i = n\bar{x}$), то $k = n -1$, и т.п. С увеличением k это распределение медленно приближается к нормальному.

Задание. Подтвердить, что случайная величина

$$\chi_k^2 = \sum_{i=1}^{k} (X_i - a)^2 / \sigma^2$$

имеет χ^2-распределение с $k = n$ степенями свободы, если $X_1, X_2, ..., X_k$ независимы и $X_i \sim N(a,\sigma)$.

Пример (из области геохимии).

Поиски коренных месторождений многих полезных ископаемых осуществляют, отбирая пробы в рыхлых отложениях. Объем проб может изменяться в широких пределах. Использование проб малого объема менее трудоемко, но при низком содержании полезного минерала может привести к пропуску ореола. Пробы большого объема требуют больших затрат на отбор и обработку. Требуется определить объем единичной пробы, гарантирующий обнаружение ореола с вероятностью 0,99 при уровне значимости $\alpha = 0{,}05$.

Для оценки оптимального объема пробы проводят экспериментальное опробование. Пусть отобрано по n проб, объемы которых- 2, 5 и 10 дм^3, содержание минерала в пробе - от 0 до 3 зерен. Статистическая модель выбирается с учетом того, что включение зерен минерала в пробу может рассматриваться как редкое событие, описываемое распределением Пуассона.

Вычисляются:

математическое ожидание числа зерен в пробах объема $V_{пр}$: $\lambda = qV_{пр}$, где $q = Q / V_{общ}$ – среднее содержание зерен минерала (в шт/дм^3) на участке; Q - общее число зерен в пробах, $V_{общ}$ - суммарный объем проб;

теоретическая вероятность попадания в пробу k-го числа зерен минерала

$$p(k) = e^{-\lambda} \lambda^k / k\,!,$$

где $k = 0, 1, ..., x, ..., m;$

ожидаемая частота отбора проб с k-м числом зерен минерала в выборке объема n:

$$n_k = np(k);$$

эмпирическая частота отбора проб с k-м числом зерен минерала $n_{k,э}$.

Данные заносят в таблицу следующего типа

Таблица 4.1

k	$p(k)$	n_k	$n_{k,э}$
0	0,8943	35,77	32
1	0,0999	3,998	5
2	0,0056	0,223	1

3	0,0002	0,008

Требуется при уровне значимости α = 0,05 проверить нулевую гипотезу: генеральная совокупность подчиняется распределению Пуассона. Рассмотрим применение критерия согласия Пирсона, в частности, для проверки гипотезы о нормальном распределении генеральной совокупности. Будем сравнивать эмпирические (фактически наблюдаемые) и теоретические (вычисленные в предположении нормального распределения) частоты.

Критерий проверки нулевой гипотезы – случайная величина

$$\chi^2 = \sum_{0}^{m} (n_i - n_{i,э})^2 / n_i \text{ ,}$$

закон распределения которой при $n \to \infty$ стремится к закону распределения χ^2 с k степенями свободы независимо от того, какому закону распределения подчинена генеральная совокупность.

Число степеней свободы $k = s - 1 - r$, где s – число групп выборки, r – число параметров предполагаемого распределения, которые оценены по данным выборки. Для нормального распределения $r = 2$ (математическое ожидание и среднее квадратическое отклонение), так что $k = s - 3$, для χ^2 распределения. $r = 1$ (объем выборки).

Строят правостороннюю критическую область. В предположении справедливости нулевой гипотезы требуют, чтобы

$$P[\chi^2 > \chi^2_{кр}(\alpha;\ k)] = \alpha.$$

Рассчитывают наблюдаемое значение критерия и сравнивают его с критическим, найденным по соответствующей таблице для заданного уровня значимости α и числа степеней свободы $k = m - 2$ (возвращается встроенной функцией Excel ХИ2ОБР). Если $\chi^2_{набл} > \chi^2_{пр}(1\text{-}\alpha;\ k)$, то нулевую гипотезу отвергают, иначе – нет оснований ее отвергнуть, т.е. гипотеза о соответствии эмпирического распределения закону Пуассона принимается. В частности, по данным табл. 4.1 $\chi^2_{набл}$ (3,36) > $\chi^2_{пр}(1\text{-}\alpha;\ k)$ (3,84), т.е. гипотеза принимается.

Принятие гипотезы позволяет использовать соотношение между вероятностью включений k-го числа зерен минерала в пробе и математическим ожиданием последних.

Вероятность фиксации ореола единичной пробой (0,99 при k > 0) соответствует вероятности его пропуска p(0) = 0,01. Так как $\lambda = -\ln\, p(0) + \ln(\lambda^0 / 0!) = -\ln\, p(0)$, т.е. λ = 4,6, то объем пробы $V_{пр} = \lambda / q$ = 4,6 / q.

4.5. Распределение Стьюдента (t-распределение)

Это распределение имеет случайная величина

$$T = Z/\sqrt{V/k}, \tag{4.16}$$

где Z-нормальная случайная величина, имеющая $M(Z) = 0$ и $\sigma(Z) = 1$, а V-независимая от Z случайная величина, распределенная по закону χ^2 с k степенями свободы. С увеличением k оно быстро приближается к нормальному распределению.

Распределение Стьдента, в частности, используется для оценки с помощью доверительных интервалов математического ожидания нормального распределения при неизвестном σ. По данным выборки строят случайную величину

$$T = \frac{\overline{X} - a}{S/\sqrt{n}},$$

где $\overline{X}$ – выборочная средняя, S – исправленное среднее квадратическое отклонение, n – объем выборки. Случайная величина T имеет распределение Стьюдента с $k = n - 1$ степенями свободы. Ее плотность распределения $S(t, n)$, где t – возможные значения T – четная функция от t. Ввиду этого

$$P\left(\left|\frac{\overline{X} - a}{S/\sqrt{n}}\right| < t_\gamma\right) = \int_0^{t_\gamma} S(t,n)dt = \gamma,$$

или

$$P(\overline{X} - t_\gamma S/\sqrt{n} < a < \overline{X} + t_\gamma S/\sqrt{n}) = \gamma.$$

Заменяя случайные величины $\overline{X}$ и S на неслучайные величины $\bar{x}$ и s, найденные по выборке, и находя в соответствующей таблице t_γ по заданным n и γ, определяют доверительный интервал

$$(\bar{x} - t_\gamma S/\sqrt{n},\ \bar{x} + t_\gamma S/\sqrt{n}),$$

покрывающий неизвестный параметр a с надежностью γ.

Пример (из области геохимии).

Найти оптимальное соотношение между химическим и полуколичественным спектральным анализом при оконтуривании геохимических аномалий (например, при разведке месторождений полезных ископаемых). В контур включать участки с концентрацией компонента A в пробах выше заданного предела $C_{пр}$.

Содержание компонентов в пробах обычно определяется с помощью трудоемких химических анализов. Чтобы не подвергать таким анализам пробы с заведомо низким или высоким содержанием компонента A, их предварительно подвергают полуколичественному спектральному анализу. Этот высокопроизводительный и дешевый метод отличается невысокой точностью. На химический анализ отправляют только те пробы, отклонения результатов анализа которых от заданного предела находятся в пределах погрешности измерения. Соответствующие нижнюю и верхнюю границы интервала концентраций компонента в таких пробах определяют по выборке данных спектрального анализа специально подготовленной пробы с содержанием компонента *A*, равным заданному пределу C_{np}. Сначала нужно определить, какому закону распределения подчиняется заданная выборка и какова систематическая ошибка измерений.

Распределение эмпирических данных обычно удовлетворительно описывается нормальным или логнормальным законами распределения. Для проверки соответствия распределения данных нормальному закону (методом моментов) вычисляются:

среднее содержание компонента

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

где n - объем выборки, x_i - i-е значение содержания компонента;

оценка дисперсии

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2;$$

асимметрия эмпирического распределения

$$A = m_3 / S^3,$$

где центральный эмпирический момент третьего порядка

$$m_3 = \frac{1}{n}\sum_{i}^{n}(x_i - \bar{x})^3;$$

эксцесс эмпирического распределения

$$E_k = m_4 / S^4 - 3,$$

где центральный эмпирический момент четвертого порядка

$$m_4 = \frac{1}{n}\sum(x_i - \bar{x})^4 .$$

Значения *A* и *E* сравнивают с теоретическими, для нормального распределения равными нулю. Если отношения $|A| / \sigma_A$ (где $\sigma_A = 2{,}45 / \sqrt{n}$ - среднеквадратическое отклонение А) и $|E| / \sigma_E$ ($\sigma_E = 4{,}9 / \sqrt{n}$) меньше трех, то гипотеза о соответствии эмпирического распределения теоретическому принимается.

В этом случае оценивают систематическую ошибку анализа $\delta = \bar{x} - C_{np}$, проверяя гипотезу о равенстве среднего содержания компонента А в исходной пробе истинному содержанию. Гипотеза принимается, если вычисленное по данным выборки (наблюдаемое) значение критерия

Стьюдента

$$t = (\bar{C} - C_{пр})\sqrt{n} / S$$

меньше табличного (возвращаемого встроенной функцией Excel СТЬДРАСПОБР) для заданного уровня значимости α (0,05) и числа степеней свободы $k = n$-1.

Поскольку выборка имеет нормальное распределение, для которого

$$F(x)= F_0((x - a) / \sigma),$$

то для определения границ искомого интервала надо сначала найти значения $z_н = (x_н - a) / \sigma$ и $z_в = (x_в - a) / \sigma$, используя таблицу стандартизованного нормального распределения F_0 (или встроенную функцию Excel НОРМСТОБР). Нижняя граница интервала определяется для $\alpha/2$, верхняя – для $1 - \alpha / 2$. Переходя от нормированных значений случайной величины к наблюдаемым, получим:

$$x_н = a + z_н \sigma – \delta, (z_н < 0),$$

$$x_в = a + z_в \sigma – \delta, (z_в > 0).$$

Систематическую ошибку δ учитывают, если она статистически значима. Следовательно, если по результатам полуколичественного спектрального анализа окажется, что концентрация компонента *A* находится в интервале $(x_н,x_в)$, то такую пробу надо направить на химический анализ.

4.6. Распределение Фишера

Это *F*-распределение, которое имеет случайная величина

$$F = (U / k_1) / (V / k_2), \tag{4.17}$$

где *U* и *V* – независимые случайные величины, распределенные по закону χ^2 со степенями свободы k_1 и k_2, соответственно.

Пример (на стыке географии и геохимии). Оценить влияние ландшафтных условий на результаты геохимической съемки.

Чтобы избежать оконтуривания ложных аномалий, учитывают влияние ландшафтных условий на геохимическую съемку. Если такое влияние установлено, то обычно обрабатывают данные по каждому ландшафтному таксону раздельно.

Пусть опробованы *p* участков (типов ландшафтов), заведомо не имеющие аномальных концентраций соответствующих элементов. Разные типы ландшафтов рассматриваются как фактор, влияющий на однородность исходной выборки. Данные по каждому типу ландшафтов объединены в группы. Оценка влияния фактора на однородность выборки основывается на сопоставлении остаточной (внутригрупповой) и факторной (межгрупповой) дисперсии признака.

Вычисляются:

среднее содержание элемента в каждой группе $\bar{x}_i$;

общее среднее всей совокупности

$$\hat{x} = \sum_{i=1}^{p}\sum_{j=1}^{n_i} x_{ij} \Big/ \sum_{i=1}^{p} n_i \ ;$$

общая дисперсия

$$S_0^2 = \frac{1}{n-p}\sum_{i=1}^{p}\sum_{j=1}^{n_i}(x_{ij}-\hat{x})^2 \ ,$$

где i = 1, 2, ..., p - количество групп, соответствующих типам ландшафтов; j=1,2,..., n_i - количество данных в выборке по каждой группе; факторная дисперсия

$$S_\phi^2 = \frac{1}{p-1}\sum_{i=1}^{p} n_i(\bar{x}_i - \hat{x})^2 \ ;$$

внутригрупповая дисперсия $S^2_{‰} = S_0^2 - S^2_\phi$.

Гипотеза о равенстве S_0^2 и $S_в^2$ проверяется по F-критерию Фишера - отношению межгрупповой дисперсии к остаточной. Если вычисленное значение F меньше табличного (возвращается встроенной функцией Excel FРАСПОБР) при выбранном уровне значимости (0,05) и при числе степеней свободы $k_1 = n - p$ и $k_2 = p - 1$, то гипотеза о равенстве дисперсий принимается и влияние ландшафта на результаты геохимической съемки считается незначительным. Если гипотеза о равенстве дисперсий отвергается, то делается вывод о необходимости раздельной оценки фоновых и минимально аномальных значений содержания элемента по каждому ландшафту.

4.7. Показательное распределение

Показательное распределение широко применяется, в частности в теории надежности.

4.7.1. Плотность показательного распределения

$$f(x) = \lambda \exp(-\lambda x) \text{ при } x \geq 0 \text{ и } 0 \text{ при } x < 0, \qquad (4.20)$$

где постоянная $\lambda > 0$.

Задание. Подтвердить, что

$$F(x) = 1 - \exp(-\lambda x) \text{ при } x \geq 0 \text{ и } 0 \text{ при } x < 0, \qquad (4.21)$$

$$P(a<X<b) = exp(-\lambda a) - exp(-\lambda b), \qquad (4.22)$$

$$M(X) = 1/\lambda, \quad D(X) = 1/\lambda^2. \qquad (4.23)$$

На рис. 4.2 представлены результаты расчета функции и плотности вероятностей показательного распределения.

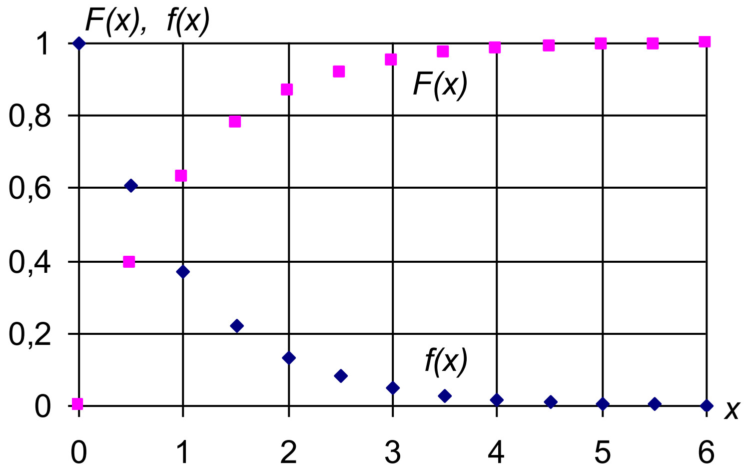


Рис. 4.2.

4.7.2. Функция надежности $R(t)$ определяет вероятность безотказной работы элемента (объекта) за время длительностью t.

Пусть $t_0 = 0$ – время начала работы, t – время, по истечении которого происходит отказ, T - непрерывная случайная величина, длительность безотказной работы элемента. Вероятность отказа за время длительностью t определяет функция распределения $F(t) = P(T< t)$, а вероятность противоположного события $R(t) = P(T> t) = 1- F(t)$

Обычно функция распределения длительности безотказной работы элемента

$$F(t)=1 - exp(-\lambda t), \qquad (4.24)$$

где λ - интенсивность отказов.

Функцию надежности, определяемую равенством

$$R(t) =1 - F(t) = exp(-\lambda t),$$

называют показательным законом надежности.

Пример. Найти вероятность того, что элемент проработает безотказно 100 часов, если при $t \geq 0$ $f(t) = 0,02 \ exp(-0,02t)$.

Решение. $R(100) = exp(-0,02*100) = 0,135.$

4.7.3. Задание. Пусть A – безотказная работа элемента на интервале $(0,t_0)$, B – на интервале (t_0, t_0+t). Тогда AB – безотказная работа на интервале (0, t_0+t). Подтвердить, что $P_A(B) = exp(-\lambda t)$, т.е. вероятность безотказной работы элемента на интервале времени длительностью t не зависит от времени предшествующей работы до начала рассматриваемого интервала, а зависит только от длительности времени (при заданной интенсивности отказов λ). Для этого найти $P(A)$, $P(B)$, $P(AB)$, $P_A(B)$.

5. Элементы однофакторного дисперсионного анализа

Дисперсионный анализ применяют, чтобы установить, оказывает ли существенное влияние на изучаемую величину X некоторый качественный фактор F, имеющий p уровней F_1, F_2, ... F_P. Он основан на сравнении "факторной дисперсии", порождаемой воздействием фактора, и "остаточной дисперсии", обусловленной случайными причинами. Фактор оказывает существенное влияние, если различие между дисперсиями значимо. Также значимо различие между групповыми средними (средними наблюдаемых значений на каждом уровне).

Пусть число наблюдений (испытаний) на каждом уровне одинаково и равно q. Наблюдалось $n = pq$ значений x_{ij} признака X, где i – номер испытания (i =1, 2, …, q), j – номер уровня фактора (j =1, 2, …, p).

Факторная дисперсия в этом случае равна отношению факторной суммы квадратов отклонений групповых средних от общей средней $\bar{x}$ к числу степеней свободы $p – 1$ факторной дисперсии:

$$s^2_{факт} = S_{факт} / (p-1)$$

где

$$s_{факт} = q\sum_{j=1}^{p}(\bar{x}_j - \bar{x})^2.$$

Остаточная дисперсия в этом случае равна отношению остаточной суммы квадратов отклонений наблюдаемых значений группы от своей групповой средней $\bar{x}_j$ к числу степеней свободы $p(q – 1)$ остаточной дисперсии:

$$s^2_{ост} = S_{ост} / p(q-1)$$

где

$$S_{ост} = \sum_{j=1}^{p}\sum_{i=1}^{q} (x_{ij} - \bar{x}_j)^2.$$

На практике остаточную сумму находят по равенству

$$S_{ост} = S_{общ} - S_{факт} ,$$

где

$$S_{общ} = \sum_{j=1}^{p}\sum_{i=1}^{q} (x_{ij} - \bar{x})^2.$$

Чем больше влияние фактора F на X, тем больше рассеяны групповые средние вокруг общей средней. Поскольку на X воздействуют и случайные причины, то наблюдаемые значения каждой группы рассеяны относительно своей групповой средней.

Пусть математические ожидания и дисперсии нормально распределенных генеральных совокупностей $X_1, X_2, ..., X_p$ неизвестны. У них одинаковые дисперсии, а математические ожидания могут быть различны. Требуется по выборочным средним при заданном уровне значимости проверить нулевую гипотезу $H_0 : M(X_1) = M(X_2) = ... = M(X_p)$ о равенстве всех математических ожиданий.

Решение этой задачи сводится к сравнению факторной и остаточной дисперсий по критерию Фишера. Если нулевая гипотеза справедлива, то эти дисперсии являются несмещенными оценками генеральной дисперсии. Следовательно, они различаются незначимо, и нулевую гипотезу о равенстве факторной и остаточной дисперсий принимают. По мере роста расхождения между групповыми средними увеличивается факторная дисперсия, и вместе с нею – отношение $F_{набл} = s^2_{факт} / s^2_{ост}$. Гипотеза отвергается, когда окажется, что наблюдаемое значение критерия больше критического: $F_{набл} > F_{кр}$.

Гипотеза принимается, если факторная дисперсия меньше остаточной. Предположение о равенстве дисперсий:

$$H_0 : D(X_1) = D(X_2) = ... = D(X_l)$$

можно проверить предварительно, например, по критерию Кочрена – отношению максимальной исправленной дисперсии к сумме всех исправленных дисперсий:

$$G = S^2_{max} / (S^2_1 + S^2_2 + \ldots + S^2_l).$$

Распределение этой случайной величины зависит только от числа степеней свободы $k = n - 1$ и количества выборок l. Строят правостороннюю критическую область, исходя из требования, чтобы в предположении справедливости нулевой гипотезы о равенстве дисперсий вероятность попадания критерия в эту область была равна принятому уровню значимости:

$$P[G > G_{кр}(\alpha; k, l)] = \alpha.$$

Если $G_{набл} < G_{кр}$, то нулевая гипотеза принимается.

6. ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ СВЯЗИ МЕЖДУ СЛУЧАЙНЫМИ ВЕЛИЧИНАМИ

Характеристики взаимосвязи двух случайных величин изучает парная корреляция на основе совместного распределения вероятностей: $F(x, y) = P(X < x, Y < y)$. Если условный закон распределения одной случайной величины изменяется в зависимости от значений, принимаемых другой случайной величиной, то их взаимосвязь называют стохастической (статистической, вероятностной). Если при этом изменяется среднее значение другой величины, то имеет место корреляционная зависимость.

Среднее арифметическое наблюдавшихся значений Y, соответствующих $X = x$, называют условным средним $\bar{y}_x$. Аналогично, среднее арифметическое наблюдавшихся значений X, соответствующих $Y = y$, называют условным средним $\bar{x}_y$. Условное среднее является оценкой соответствующего условного математического ожидания (регрессии Y на X, или X на Y).

6.1. Коэффициенты корреляции

6.1.1. Одной из характеристик последней корреляционный момент (ковариация $\mathrm{cov}(X_i,X_j)$ случайных величин X_i и X_j - второй смешанный центральный момент случайных величин X_i и X_j)

$$\mu_{ij} = \sigma_{ij} = \mathrm{cov}(X_i,X_j) = M[(X_i - M(X_i))(X_j - M(X_j))]. \qquad (6.1)$$

Корреляционный момент дискретных случайных величин X и Y вычисляют по формуле

$$\mu_{xy} = \sum_{i=1}^{n}\sum_{j=1}^{m} [x_i - M(X)]\,[y_j - M(Y)].p(x_i\,, y_j). \qquad (6.2)$$

Корреляционный момент непрерывных случайных величин X и Y

$$\mu_{xy} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [x - M(X)]\,[y - M(Y)].f(x\,, y)dx\,dy. \qquad (6.3)$$

Размерность корреляционного момента равна произведению размерностей случайных величин, т.е. его величина зависит от того, в каких единицах были измерены эти величины. Этого недостатка не имеет коэффициент корреляции.

6.1.2. Задание. Используя формулу (5.1) подтвердить, что

$$\sigma_{ij} = M(X_i X_j) - M(X_i)\,M(X_j)]. \qquad (6.4)$$

Сравнить эту формулу с формулой (3.19) для вычисления дисперсий.

6.1.3. Задание. Используя формулу (6.1) подтвердить, что для независимых случайных величин $\sigma_{ij} = 0$.

6.1.4. Если корреляционный момент случайных величин не равен нулю, то их называют коррелированными. Коррелированные случайные величины зависимы. Допуская обратное, придется заключить, что $\mu_{xy} \neq 0$, что противоречит условию. Зависимые величины могут быть некоррелированными. Вообще, из некоррелированности еще не вытекает независимость, но для нормально распределенных случайных величин это имеет место.

6.1.5. Задание. Подтвердить, что

$$|\sigma_{XY}| \le \sigma_X \sigma_Y . \qquad (6.5)$$

Методические указания. Найти дисперсии случайных величин $\sigma_X X \pm \sigma_Y Y$ и учитывая, что любая дисперсия неотрицательна, подтвердить соответствующие неравенства (в частности $D(\sigma_X X - \sigma_Y Y) = 2\sigma_X^2\sigma_Y^2 - 2\sigma_X\sigma_Y\sigma_{XY}$, откуда следует $\sigma_{XY} \le \sigma_X\sigma_Y$) .

6.1.6. Коэффициент корреляции r_{xy} случайных величин X и Y – это отношение корреляционного момента к произведению средних квадратических отклонений этих величин

$$r_{xy} = \mu_{xy} / (\sigma_X \sigma_Y) \qquad (6.6)$$

Задание. Учитывая формулу (5.8), подтвердить, что

$$| r_{xy} | \le 1. \qquad (6.7)$$

6.1.7. Задание. Получить статистическую оценку коэффициента корреляции генеральной совокупности - выборочный коэффициент корреляции

$$r_в = (\sum n_{xy} xy - n\bar{x}\bar{y}) / n\tilde{\sigma}_x \tilde{\sigma}_y ,$$

где x, y – варианты признаков X и Y;

n_{xy} – число наблюдений пар чисел (x, y);

n - объем выборки;

$\bar{x}, \bar{y}$ - выборочные средние;

$\tilde{\sigma}_ч, \tilde{\sigma}_y$ - выборочные средние квадратические отклонения.

6.1.8. Выборочный коэффициент корреляции $r_в$ служит для измерения линейной связи между X и Y. Последние независимы, если $r = 0$. Они связаны линейной функциональной зависимостью, если $r = 1$. Если гипотеза о равенстве нулю генерального коэффициента корреляции будет отвергнута, то X и Y коррелированны, иначе - нет.

6.1.9. Проверка гипотезы о значимости выборочного коэффициента корреляции

Требуется при заданном уровне значимости проверить нулевую гипотезу $H_0 : r_в = 0$ при конкурирующей гипотезе $H_1 : r_в \ne 0$. Если нулевая гипотеза отвергается, то выборочный коэффициент корреляции значимо отличается от нуля, а X и Y коррелированны

Критерий проверки нулевой гипотезы

$$T = r_в \sqrt{n-2} / \sqrt{1 - r_в}$$

при справедливости нулевой гипотезы имеет распределение Стьюдента с $k = n - 2$ степенями свободы. Если $|T_{набл}| > t_{кр}$, то нулевую гипотезу отвергают.

6.1.10. Плотность распределения двумерной нормальной случайной величины

$$f(x, y) = \exp\left\{-\frac{1}{2(1-r^2)}\left(z_1^2 - 2rz_1z_2 + z_2^2\right)\right\} / 2\pi\sigma_1\sigma_2\sqrt{1-r^2}, \qquad (6.8)$$

где

$$z_1 = (x-a_1) / \sigma_1, \; z_2 = (y-a_2) / \sigma_2 .$$

Задание. Полагая, что в формуле (6.8) $r = 0$, получить $f(x_1, x_2) = f(x_1)\, f(x_2)$. Что это означает?

6.2. Уравнения регрессии.

6.2.1. Линейным приближением случайной величины Y, взаимосвязанной со случайной величиной X, является среднеквадратическая регрессия Y на X:

$$\tilde{Y} = a_0 + b_0 X, \qquad (6.9)$$

где

$$a_0 = M(Y) - M(X)\, r_{XY}\, \sigma_Y/\sigma_X,$$

$$b_0 = r_{XY}\, \sigma_Y/\sigma_X.$$

Коэффициент b_0 называют коэффициентом регрессии Y на X

$$\beta = r_{XY}\, \sigma_Y/\sigma_X \qquad (6.10)$$

6.2.2. Коэффициенты функции (6.9) можно найти методом наименьших квадратов, исходя из условия, что математическое ожидание $M(Y-\tilde{Y})^2$ принимает наименьшее возможное значение. Для этого необходимо, чтобы математическое ожидание квадрата отклонения

$$M(Y-\tilde{Y})^2 = F(a_0, b_0) = M(Y - a_0 - b_0 X)^2 \qquad (6.11)$$

было минимальным. Это условие выполняется, если приравнять нулю следующие производные:

$$\partial F(a_0, b_0) / \partial a_0 = 0;$$

$$\partial F(a_0, b_0) / \partial b_0 = 0.$$

После преобразований получают систему уравнений:

$$M(Y) - a_0 - b_0 M(X) = 0;$$

$$r_{XY}\,\sigma_Y\,\sigma_X - b_0\,\sigma^2_x = 0,$$

решая которую находят коэффициенты a_0 и b_0 .

6.2.3. Задание. Решить систему последних уравнений с неизвестными a_0 и b_0 . Подтвердить, что

$$\tilde{y} - M(Y) = (x - M(X))\, r\sigma_Y / \sigma_X \ . \tag{6.12}$$

Эта формула задает функцию регрессии Y на X, если $M(Y/X = x) = ax + b$. Прямая, построенная по формуле (6.12.) называется прямой средней квадратичной регрессии Y на X.

Аналогично,

$$\tilde{x} - M(X) = (y - M(Y)) r\sigma_x / \sigma_y \ . \tag{6.13}$$

Задание. Используя формулы (6.12) и (6.13) показать, что обе прямые проходят через точку $(M(X), M(Y))$, которую называют центром совместного распределения случайных величин X и Y.

6.2.4. Подставляя значения a_0 и b_0 в формулу (6.11), находят минимальное значение математического ожидания квадрата отклонения среднеквадратичной регрессии Y на X от Y:

$$M(Y - (a_0 + b_0 X))^2 = \sigma_y^2(1 - r^2). \tag{6.14}$$

Чем ближе r к 1, тем меньше дисперсия отклонения Y от наилучшего линейного приближения (остаточная дисперсия). При $/ r / = 1$ остаточная дисперсия равна нулю, т.е. X и Y связаны линейной функциональной зависимостью (обе прямые регрессии совпадают).

6.2.4. Если обе функции регрессии линейны, то X и Y связаны линейной корреляционной зависимостью. Графики линейных функций регрессии – прямые линии. Они совпадают с прямыми среднеквадратической регрессии.

Если двумерная случайная величина (X,Y) распределена нормально, то X и Y связаны линейной корреляционной зависимостью.

6.2.5. Задание. Найти функцию регрессии Y на X, если (X,Y) – нормальная случайная величина.

Ответ: это нормальная случайная величина с параметрами:

$$a = M(Y/x) = M(Y) + r\frac{\sigma_y}{\sigma_x}(x - M(X)); \tag{6.15}$$

$$\sigma^2 = D(Y/x) = \sigma_x^2(1 - r^2), \tag{6.16}$$

т.е. функция регрессии Y на X является прямой линией, а условная дисперсия не зависит от x.

6.2.6. Задание. Получить выборочное уравнение прямой линии регрессии *Y* на *X*

$$\tilde{y}_в - \bar{y} = (x - \bar{x}) r_в \tilde{\sigma}_Y / \tilde{\sigma}_X \ ,$$

где выборочный коэффициент регрессии *Y* на *X*

$$\rho_{yx} = r_в \tilde{\sigma}_y / \tilde{\sigma}_x \, .$$

Аналогично, выборочное уравнение прямой линии регрессии *X* на *Y*:

$$\tilde{x}_в - \bar{x} = (y - \bar{y}) r_в \tilde{\sigma}_x / \tilde{\sigma}_y \ ,$$

где выборочный коэффициент регрессии *X* на *Y*

$$\rho_{xy} = r_в \tilde{\sigma}_x / \tilde{\sigma}_y \, .$$

6.3. Нелинейная связь

6.3.1. Пусть данные наблюдений над количественными признаками *X* и *Y* сведены в корреляционную таблицу. В ее первой строке указаны наблюдаемые значения признака *X*, в первом столбце – признака *Y*. На пересечении строк и столбцов – частоты n_{xy} наблюдаемых пар значений признаков. В последнем столбце – суммы частот строк n_y , в последней строке – суммы частот столбцов n_x. В правом нижнем углу таблицы – сумма всех частот (число всех наблюдений *n*).

Это, в частности, означает, что наблюдаемые значения Y разбиты на группы, каждая из которых содержит те значения Y, которые соответствуют определенному значению X. Условные средние можно назвать групповыми средними. Общую дисперсию признака можно представить в виде суммы внутригрупповой и межгрупповой дисперсий:

$$D_{общ} = D_{вг} + D_{мг} \, .$$

Если Y связан с X функциональной зависимостью,то в разных группах находятся разные значения Y, но в каждой из них – одни и те же. Поэтому групповая дисперсия каждой группы равна нулю, т.е. $D_{вг} = 0$, $D_{общ} = D_{мг}$ и

$$D_{мг} / D_{общ} = 1.$$

Если Y связан с X корреляционной зависимостью,то

$$D_{мг} / D_{общ} < 1.$$

Чем связь между Y и X ближе к функциональной, тем больше отношение $D_{мг} / D_{общ}$ приближается к единице. Именно поэтому оно, а точнее – корень квадратный из него, служит мерой тесноты корреляционной связи.

Анализ тесноты нелинейной корреляционной связи осуществляется с помощью выборочных корреляционных отношений *Y* к *X*:

$$\eta_{yx} = \sigma_{мж} / \sigma_{общ} \quad \text{или} \quad \eta_{yx} = \sigma_{\bar{y}_x} / \tilde{\sigma}_y$$

и выборочных корреляционных отношений X к Y

$$\eta_{xy} = \sigma_{\bar{x}_y} / \tilde{\sigma}_x,$$

где

$$\sigma_{\bar{y}_x} = \sqrt{(\sum n_x(\bar{y}_x - \bar{y})^2) / n};$$

$$\tilde{\sigma}_y = \sqrt{(\sum n_y(y - \bar{y})^2) / n};$$

n - объем выборки; n_x - частота значения x признака X; n_y - частота значения y признака Y; $\bar{y}$ - общая средняя признака Y; $\bar{y}_x$- условная средняя признака Y.

Корреляционное отношение служит мерой тесноты связи любой формы, но не указывает на то, как близко к кривой определенного типа расположены точки, найденные по данным наблюдений.

Установить форму и тесноту корреляционной связи - основная задача корреляционного анализа.

6.3.2. Свойства выборочного корреляционного отношения:

$$0 \leq \eta \leq 1;$$

если $\eta = 0$, то признак Y не связан корреляционной зависимостью с признаком X;

если $\eta = 1$, то признак Y связан функциональной зависимостью с признаком X;

выборочное корреляционное отношение не меньше абсолютной величины выборочного коэффициента корреляции $\eta \geq |r_в|$;

если выборочное корреляционное отношение равно абсолютной величине выборочного коэффициента корреляции, то имеет место точная линейная корреляционная зависимость.

6.3.3. Простейшие случаи криволинейной корреляции

Если график регрессии изображается кривой линией, то корреляцию называют криволинейной, например $\bar{y}_x = ax^2+bx+c$. Для определения вида функции регрессии строят точки $(x, \bar{y}_x)$ и по их расположению делают заключение о примерном виде функции регрессии; при окончательном решении учитывают особенности, вытекающие из сущности решаемой задачи.