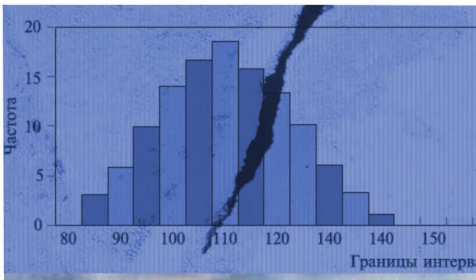
Е.Н. ГУСЕВА

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА



МАГНИТОГОРСК, 2009

Министерство образования и науки Российской Федерации
ГОУ ВПО «Магнитогорский государственный университет»

Е.Н. Гусева

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Учебное пособие

5-е издание, стереотипное

Москва
Издательство «ФЛИНТА»
2011

ББК В17/172
УДК 372.016:519.2
Г96

Р е ц е н з е н т ы:
доктор физико-математеческих наук, профессор
Магнитогорского государственного университета
С.И. Кадченко;
кандидат технических наук, доцент
Магнитогорского государственного технического университета
А.В. Леднов

Гусева Е.Н.
Г96 Теория вероятностей и математическая статистика : [электронный ресурс] учеб. пособие / Е. Н. Гусева. – 5-е изд., стереотип. – М.: ФЛИНТА, 2011. – 220 с.

ISBN 978-5-9765-1192-7

Пособие содержит теоретические основы курса «Теория вероятностей и математическая статистика», а также лабораторный практикум. Издание адресовано студентам высших учебных заведений, изучающим теорию вероятностей и математическую статистику.

ISBN 978-5-9765-1192-7

Оглавление

Основы теории вероятностей и математической статистики

Цель: *познакомиться с основными понятиями теории вероятностей, изучить аксиоматический подход к определению понятия «вероятность».*

Из истории статистики

Статистика имеет многовековую историю, уходя своими корнями в глубокую древность. Исторически развитие статистики было связано с возникновением государств, потребностями в их эффективном управлении. Первая публикация по статистике – это «Книга Чисел» в Библии в Ветхом Завете, в которой рассказано о переписи военнообязанных, проведенной Моисеем и Аароном.

Хозяйственные и военные нужды городов и государств древнего мира требовали знаний о населении, его составе, имуществе. Первые статистические сведения собирались для налогообложения, учета земель, призыва на военную службу. В античном мире подсчитывалось число родившихся детей, велись земельные кадастры, появились первые описания государств. Благодаря Аристотелю (384–322 гг. до н.э.) можно узнать о 157 городах и государствах того времени. Сбор и обработка данных о массовых общественных явлениях со временем приобрели регулярный характер.

Некоторые разделы статистики были разработаны на основе изучении теории азартных игр в XVI–XVII вв. Исследованиями в этой области занимались Д. Кардано, Х. Гюйгенс, Б. Паскаль, П. Ферма и др. Следующий этап развития науки связан с именем Я. Бернулли (1654–1705). Теорема Бернулли, названная «Зако-

ном больших чисел», была первым теоретическим обоснованием накопленных ранее фактов (при достаточно большом количестве испытаний вероятность события почти равна частоте этого события).

В конце XVII в. при страховании кораблей, начали подсчитывать, сколько шансов на то, что корабль вернется в порт невредимым, не будет потоплен бурей, что груз не испортится, не будет захвачен пиратами и т.д. Такой расчет позволял определять, какую страховую сумму следует выплачивать и какой страховой взнос брать, чтобы это было выгодно для компании.

В 1746 г. профессор философии и права Г. Ахенваль впервые в Марбургском университете начал преподавать новую дисциплину, названную им статистикой. С середины XIX в. благодаря усилиям бельгийца – математика, астронома и статистика А. Кетле (1796–1874) были выработаны правила переписей населения и регулярность их проведения в разных странах. По его инициативе проводились международные статистические конгрессы, а в 1885 году был основан Международный статистический институт. Международной статистикой занимаются такие организации – ООН, ФАО, ЮНЕСКО, МОТ, ЕС, Мировой банк и др. Эти организации занимаются сбором, представлением, сравнением и интерпретацией социально-экономических данных.

Свой вклад в теорию вероятностей внесли А. Муавр, П. Лаплас, К. Ф. Гаусс, С. Пуассон и др. Другой плодотворный период связан с П.Л. Чебышевым, А.А. Марковым, А.М. Ляпуновым. В это время теория вероятностей становится стройной математической наукой. Большое влияние на дальнейшее развитие науки связано с русскими математиками С. Н. Берштейном, В.И. Романовским и А.Н. Колмогоровым.

Элементы теории вероятностей

Событие (явление**) –** возможный исход испытания, опыта, наблюдения.

Все наблюдаемые нами явления можно условно разделить на три вида: *достоверные, невозможные и случайные.*

Достоверным называется событие, которое обязательно произойдет, если будут выполнены определенные условия. Например, если в сосуде находится вода, давление атмосферы нормальное, а температура воздуха 30°С, то событие «вода в сосуде находится в жидком состоянии».

Невозможным называется событие, которое заведомо не произойдет, если будет выполнена совокупность условий S. Например, при нагревании олова и меди вы не сможете получить золото.

Случайным называют событие, которое при осуществлении условий S может произойти, а может и не произойти. Например, при бросании монеты выпадение герба – является случайным событием, потому что оно может произойти, а может и не произойти.

Каждое случайное событие есть следствие действия очень многих сил и случайных причин, которые учесть просто невозможно. В нашем случае это сила броска, вес и размер монеты, ее симметричность, состояние здоровья человека, бросившего монету и т.д. Поэтому теория вероятностей не ставит себе задачу предсказать, произойдет или нет единичное случайное событие – она просто не в силах это сделать. Однако, когда речь идет о случайных событиях, которые многократно наблюдаются при осуществлении одних и тех же условий S, то есть происходят **массовые однородные случайные события**, то оказывается

что они подчиняются определенным закономерностям. Эти закономерности называются вероятностными.

Например, выпадение снега в Москве 10 октября является случайным событием. Ежедневный восход Солнца можно считать достоверным событием, а выпадение снега на экваторе – невозможным событием.

Предметом теории вероятностей является изучение вероятностных закономерностей массовых однородных случайных событий.

Знание закономерностей, которым подчиняются массовые события позволяет предвидеть, как эти события будут протекать. Например, нельзя заранее определить результат одного бросания монеты, но можно предсказать, причем с небольшой погрешностью, число выпадений «орла» или «решки», если монета будет брошена большое число раз в одних и тех же условиях.

Прежде чем мы введем основные понятия, теоремы, следствия теории вероятностей, попробуем рассмотреть общие принципы построения математических дисциплин. Теория вероятностей – математическая дисциплина, родственная таким дисциплинам, как, например, геометрия или теоретическая механика.

В каждой изучаемой дисциплине, как правило, существуют три аспекта:

а) формально-логическое содержание,

б) интуитивные представления,

в) приложения.

Характер дисциплины в целом и перспективы ее применения нельзя по-настоящему оценить, не рассматривая эти три аспекта в их взаимосвязи.

Формально-логическое содержание. Характерной особенностью математики является то, что она занимается исключительно соотношениями между неопределяемыми вещами. Невозможно “определить” шахматы иначе, как сформулировав систему правил игры. Аналогично этому геометрия не беспокоится о том, чем “на самом деле” являются точки и прямые. Они остаются неопределяемыми понятиями, и аксиомы геометрии лишь устанавливают связи между ними. Это правила игры, и в них нет ничего таинственного. Формально-логическое содержание статистики представляет собой совокупность понятий, общих представлений и закономерностей окружающего нас мира. В основе дисциплины лежат свои аксиомы и теоремы, которые являются фундаментом статистики.

Интуитивные представления. Каждый приобретает интуитивное представление о смысле самых разных понятий. Эта интуиция является достаточной предпосылкой для первых формальных правил теории вероятностей.

Приложения. Приложения теории вероятностей и математической статистики весьма обширны. Знания, полученные в результате статистического анализа явлений окружающего нас мира, применяются в экономике, политике, промышленности и других областях деятельности людей. Используются эти данные для изучения реальных процессов, а также эффективного управления ими и прогнозирования.

Основные определения вероятности

В отличие от математических дисциплин, изучающих “точные” закономерности, предметом теории вероятностей являются специфические закономерности, наблюдаемые при анализе случайных явлений. Эти закономерности проявляются в

массовых явлениях, и позволяют предсказывать с той или иной вероятностью исход испытаний. Тогда как, в единичном случае можно только предположить исход события.

Мы можем наблюдать широкий круг явлений, когда при **многократном** осуществлении комплекса условий $\sum$ доля той части случаев, когда событие А происходит, лишь изредка уклоняется сколько-нибудь значительно от некоторой средней цифры, которая таким образом может служить характерным показателем *массовой операции* (многократного повторения комплекса $\sum$) по отношению к событию А. Закономерности этого рода называются **вероятностными** или **стохастическими** закономерностями.

Итак, имеется схема для различных событий, наступающих при неизменном комплексе условий: **достоверное – случайное – невозможное**. Ясно, что большая часть событий в мире находится между достоверностью и невозможностью (интуитивное понимание!).

По мере развития теории вероятностей, а также областей её приложения, развивались и представления об основном понятии этой теории – вероятности.

В настоящее время существует четыре подхода к определению вероятности:

1. Определение математической вероятности как количественной меры “степени уверенности” познающего объекта.

2. Определения, сводящие понятие вероятности к понятию “равновозможности” как к более примитивному понятию (так называемое “классическое” определение вероятности).

3. Определения, основанные на “частоте” появления события в большом количестве испытаний (“статистическое” определение).

4. Аксиоматический подход, на основе теории множеств, формализующий теорию вероятностей.

Вероятностью события P(A) называют отношение числа благоприятных исходов испытания m к общему числу всех равновозможных несовместных элементарных исходов n:

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

Это определение вероятности базируется на классическом подходе и часто применяются для решения конкретных задач, поэтому мы часто будем обращаться к нему далее. Остановимся подробнее на аксиоматическом подходе к определению вероятности события.

Аксиоматическое определение вероятности

Прежде, чем рассмотреть вероятность с указанной позиции, вспомним, что аксиома – это исходное утверждение какой-либо научной теории, которое берется в качестве недоказуемого, и из которого выводятся все остальные предложения теории по принятым в ней правилам вывода.

Построение аксиом теории вероятностей А.Н. Колмогоровым означало переход от полуэмпирического, интуитивного понимания вероятности к строгому формализованному. Для введения аксиом нам необходимо принять следующие соглашения.

Зафиксируем комплекс условий $\sum$ и рассмотрим некоторую систему S событий A, B, C, ..., каждое из которых должно при каждом осуществлении комплекса $\sum$ ***произойти*** или ***не произойти***. Далее введём соглашения, которые, как увидит внимательный читатель, являются соглашениями теории множеств и математической логики.

1) Событие, состоящее в наступлении обоих событий A и B, будем называть ***произведением*** событий A и B и обозначать **AB** (или **A∩B**).
2) Событие, состоящее в наступлении хотя бы одного из событий A и B, будем называть ***суммой*** событий A и B и обозначать **A+B** (или **A∪B**).
3) Событие, состоящее в том, что событие A происходит, а событие B не происходит, будем называть ***разностью*** событий A и B и обозначать **A – B**.
4) Если при каждом осуществлении комплекса условий ∑, при котором происходит событие A, происходит и событие B, то мы будем говорить, ***что A влечет за собой B***, и обозначать это символом **A⊂ B** или **B ⊃ A**.
5) Если A влечет за собой B и в то же время B влечет за собой A, то есть если при каждой реализации комплекса условий ∑ события A и B оба наступают или оба не наступают, то мы будем говорить, что события A и B ***равносильны***, и обозначим это **A=B**. Равносильные события могут заменять друг друга или, по-другому, они **тождественны**.
6) Два события A и $\overline{A}$ называются ***противоположными***, если для них одновременно выполняются два соотношения:
7) $A+\overline{A} = U$ (достоверное событие),
   $A*\overline{A} = V$ (невозможное событие)

Пусть U – достоверное событие. Все достоверные события равносильны между собой.

V – невозможное событие. Все невозможные события тоже равносильны между собой.

8) Два события A и B называются ***несовместимыми***, если их совместное появление невозможно, то есть если **A*B = V**.

Если $\mathbf{A=B_1+B_2+...+B_N}$ и события $\mathbf{B_i}$ попарно несовместимы, то есть $\mathbf{B_iB_j = V}$ при $i \neq j$, то говорят, то событие А подразделяется на частные случаи B_1, B_2, ..., B_n. Например, при бросании игральной кости событие С, состоящее в выпадении *четного* числа очков, подразделяется на частные случаи E_2, E_4 и E_6, состоящие соответственно в выпадении 2, 4 и 6 очков.

События B_1, B_2, ..., B_N образуют ***полную группу событий***, если хотя бы одно из них непременно должно произойти (при каждом осуществлении комплекса $\sum$), то есть если $\mathbf{B_1+B_2+...+B_N = U}$.

Пример. В порту имеется два причала для приема судов. Можно рассмотреть три события: B_1 – отсутствие судов у причалов, B_2 – присутствие одного судна у одного из причалов, B_3– присутствие двух судов у двух причалов. Эти три события образуют полную группу.

В каждой задаче теории вероятностей приходится иметь дело с каким-либо определенным комплексом условий $\sum$ и с какой-либо определенной системой S событий, наступающих или не наступающих после каждой реализации комплекса условий $\sum$. Относительно этой системы целесообразно сделать следующие допущения:

а) если системе S принадлежат события А и В, то ей принадлежат также события АВ, А+В, А-В (замкнутость относительно операций);

б) система S содержит достоверное и невозможное события ("единица" и "ноль" в замкнутой системе).

Система событий, удовлетворяющая этим допущениям (1-9), называется ***полем событий***.

Всегда можно выделить такие события, которые не могут быть разложены на более простые: выпадение определенной грани при бросании игральной кости, попадание в определенную точку квадрата при рассмотрении диаграммы Венна (рис.1).

Назовем такие неразложимые события – э**лементарными событиями**.

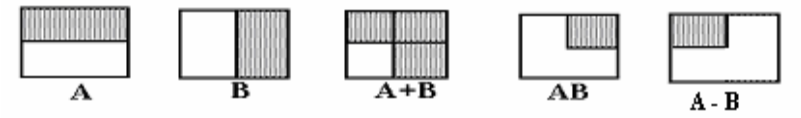


Рис. 1. Диаграммы Венна, описывающие события A и B, их сумму, произведение и разность

Для построения математической теории вероятностей требуется дополнительная формализация.

Введем понятие – ***пространство элементарных событий***, которое состоит из множества всех возможных элементарных событий. Элементами пространства могут быть точки евклидова пространства, функции одной или нескольких переменных и т. д. Множество точек пространства элементарных событий образуют случайные события. Имеется в виду любая доступная комбинация из элементарных событий, полученная в результате легальных операций. Событие, состоящее из всех точек пространства элементарных событий, называется ***достоверным событием***.

Для пространства элементарных событий, определенных выше указанным способом, имеют место следующие законы, пришедшие из алгебры (табл. 1).

Таблица 1

Законы пространства элементарных событий

коммутативный	A + B = B + A	AB = BA
ассоциативный	A+(B+C)= =(A+B)+C	A(BC) = (AB)C
дистрибутивный	A(B + C) = =AB + AC	A+(BC)=(A+B)(A+C)
тождества	A + A = A	A*A = A

Столь долгая процедура потребовалась, чтобы перейти на язык теории множеств и формальной алгебры. Следующим шагом будет выделение условий для ввода аксиом теории вероятностей.

Пусть задано некоторое множество Ω. Элементы этого множества называются **элементарными событиями**. Предположим, что фиксирована некоторая система подмножеств множества Ω; эти подмножества названы просто **событиями**. События обозначаются A, B, C и так далее. При этом потребуем, что:

I. Само множество Ω есть событие;

II. Если A – событие, то $\overline{A}$ - тоже событие; здесь символ $\overline{A}$ обозначает дополнение к подмножеству A в Ω;

III. Если A_1, A_2, ... события, то и $A_1+A_2+...$, а также $A_1A_2...$ – снова события. Под $A_1+A_2+...$ понимается **объединение** всех подмножеств A_1, A_2, ..., а под $A_1A_2...$ – их **пересечение**.

Таким образом, если строго следовать теоретико-множественному подходу, мы задаем **алгебру событий σ** на множестве Ω. **σ-алгебра** событий является системой подмножеств пространства элементарных исходов Ω, замкнутая относительно конечного числа теоретико-множественных операций.

Число подмножеств A_i может быть конечным и бесконечным.

Множество Ω называют **пространством элементарных событий**.

Два события А и В, не имеющие (как два подмножества) общих элементов, называются **несовместными**.

События А и $\overline{A}$ называются **противоположными**.

Событие Ω называется **достоверным**, событие $\overline{\Omega}$ (то есть пустое множество) – **невозможным**.

Аксиомы Колмогорова, задающие понятие вероятности:

Аксиома 1. Каждому событию А поставлено в соответствие неотрицательное число p(A), называемое вероятностью события А.

Аксиома 2. Если события A_1, A_2, ... попарно несовместны, то вероятность наступления хотя бы одного из них равна сумме вероятностей этих событий.

$$p(A_1+A_2+...+A_n)= p(A_1) + p(A_2) + ... p(A_n).$$

Для случая, когда пространство Ω конечно (аксиома 2), может быть заменено более слабым требованием:

p(A+B) = p(A) + p(B) , если А и В несовместны.

Аксиома 3. P(Ω) = 1. Вероятность полной группы событий равна 1.

Примеры полной группы событий

Приобретены два билета денежно-вещевой лотереи. Обязательно произойдет одно и только одно из следующих событий: «выигрыш выпал на первый билет и не выпал на второй», «выигрыш не выпал на первый билет и выпал на второй», «выигрыш выпал на оба билета», «на оба билета выигрыш не выпал». Эти события образуют полную группу попарно несовместных событий.

Стрелок произвел выстрел по цели. Обязательно произойдет одно из следующих двух событий: попадание, промах. Эти два несовместных события образуют полную группу. Все теоремы теории вероятностей выводятся из аксиом 1-3.

Задача 1. Из колоды игральных карт, содержащей 36 листов, наугад выбирается одна карта. Найти вероятность того, что: а) карта окажется красной масти; б) карта окажется картинкой; в) карта окажется дамой; г) эта карта туз буби.

Решение.

а) n = 36, m = 18; P(A) = 18/36=1/2.
б) n = 36, m = 16; P(A) = 16/36=4/9.
в) n = 36 , m = 4; P(A) = 4/36=1/9.
г) n = 36, m = 1; P(A) = 1/36.

Задача 2. Пусть вероятность того, что студент получит на экзамене по статистике «пятерку» равна 0,17, «четверку» – 0,38, «тройку» – 0,32, а «двойку» – 0,13. Найти вероятность того, что очередной студент получит оценку, не меньше тройки.

Решение. Искомое событие ***D*** произойдет, если будет получена оценка 5 (событие ***A***), оценка 4 (событие ***B***), или оценка 3 (событие ***C***), то есть событие ***D*** есть сумма событий ***A, B, C***. События ***A***, ***B*** и ***C*** несовместимы. Поэтому, применяя теорему сложения вероятностей, получим*:*

$$\mathbf{P(D) = P(A+B+C) = P(A) + P(B)+ P(C)} = 0{,}17 + 0{,}38 + 0{,}32 = 0{,}87.$$

Задача 3. Испытатель проводит опыты с пирамидкой, подбрасывая ее и определяя какая грань выпадет при очередном испытании. В результате опытов были определены вероятности выпадения каждой из четырех граней: 1/3, 1/6, 1/3, 1/6. Определите вероятность полной группы событий.

Решение. Вероятность полной группы событий определяется как сумма вероятностей всех элементарных исходов данной группы:

$P(\Omega) = \mathbf{P(A) + P(B)+ P(C)+ P(D)=}\ 1/3+1/6+1/3+1/6=1.$

Следствия из аксиом 1-3:

Следствие 1. Вероятность достоверного события равна единице. Действительно, если событие достоверно, то каждый исход испытания благоприятствует событию, то есть m=n, а значит и его вероятность **P(A)=m/n=1.**

Следствие 2. Вероятность невозможного события равна нулю. Раз ни один из исходов испытания не благоприятствует событию, то m=0, а тогда **P(A)=m/n=0.**

Следствие 3. Вероятность случайного события есть положительное число, заключенное между 0 и единицей. Поскольку случайному событию благоприятствует лишь часть из общего числа элементарных исходов испытания. То есть 0<m<n, а значит 0< m/n<1.

$$0 \le P(A) \le 1$$

Таким образом, введя аксиоматическое понятие вероятности, мы получили возможность использовать для доказательства теорем и следствий аппарат формальной логики.

Контрольные вопросы

1. Назовите основные этапы в истории развития статистики как науки.
2. Как в статистике определяется термин «событие»?
3. Какие события называются случайными, достоверными, невозможными?

4. Приведите примеры достоверных и невозможных событий.
5. Какие события называются совместными?
6. Приведите примеры совместных и несовместных событий.
7. Сформулируйте основные теоремы Колмогорова.
8. Что называется суммой событий?
9. Что называется произведением событий?

Классическая и статистическая модели вероятности

Цель: *Сформировать у студентов понятие «вероятности» с точки зрения классического и статистического подходов, познакомить с методами решения задач.*

Основные теоремы классической теории вероятностей

Рассмотрим вероятность с тех же позиций, с каких она исследовалась несколько веков математиками.

Классический подход тоже имеет дело с **событием**, как основным понятием теории вероятностей. Выделяется *достоверное*, *невозможное* и *случайное* события. Однако, если при построении аксиоматической теории единственным ограничением для событий, составляющих поле, было требование достоверности всех событий (т.е. по крайней мере, одно их них, должно произойти при выполнении заданного комплекса условий Σ), то классический подход основан на дополнительном требовании – равновозможности всех событий в выбранном поле событий.

Равновозможность означает равноправность (симметрию) отдельных исходов испытания относительно некоторого комплекса условий.

Например, имея колоду игральных карт из 36 листов, можно вытащить из нее любую карту. Вероятность вытащить каждую из 36 карт одинакова и равна 1/36. Это событие является равновозможным и несовместимым с другими. Требование равновозможности является жестким ограничением и сужает круг задач, которые можно решить при помощи классического подхода.

Вероятность события **A** равна отношению числа случаев **m**, благоприятствующих этому событию, к общему числу единственно возможных, равновозможных и несовместимых исходов испытания **n**:

$$\mathbf{P(A) = m/n}.$$

Пример 1. Брошены две игральные кости (кубики с нумерованными гранями). Найти вероятность, что сумма очков на выпавших гранях – четная, причем на грани хотя бы одной из костей появится шестерка.

Решение. На выпавшей грани “первой” кости может появиться одно очко, два очка, ..., шесть очков. Аналогичные шесть элементарных[1] исходов возможны при бросании “второй” кости. Каждый из исходов бросания “первой” кости может сочетаться с каждым из исходов бросания “второй”. Таким образом, общее число возможных элементарных исходов испытания равно 6*6 = 36. Эти исходы единственно возможны и, в силу симметрии костей, равновозможны.

---

[1]**Элементарное событие** – событие, которое не может быть разложено на составляющие события.

Благоприятными относительно интересующего нас события (хотя бы на одной грани появится шестерка и сумма выпавших очков – четная) являются следующие пять исходов:

6, 2; 6+2 = 8 2, 6; 2+6 = 8

6, 4; 6+4 = 10 4, 6; 4+6 = 10

6, 6; 6+6 = 12

Первым записано число очков, выпавших на “первой” кости, вторым – число очков, выпавших на “второй” кости, а затем указана сумма очков.

Искомая вероятность равна отношению числа исходов, благоприятствующих событию, к числу всех исходов:

$$P(A) = 5/36.$$

Если теперь обратиться к аксиоматическому определению вероятности и сравнить его с классическим, то мы увидим, что классическое определение является частным случаем аксиом. Действительно, если аксиому 1 дополнить требованием равенства введенных чисел p_i друг другу, а их сумму ограничить единицей (аксиома 3), то мы получим классическое определение.

Теоремы классической теории вероятностей

Для эффективного использования вычислений вероятности, основанной на классическом подходе было доказано несколько теорем и следствий из них. Как видно, эти теоремы перекликаются с теоремами аксиоматического подхода.

Теорема 1. Вероятность любого события не может быть отрицательной и больше единицы.

Теорема 2. Вероятность достоверного события равна единице.

Теорема 3. Вероятность невозможного события равна нулю.

Теорема 4. Теорема сложения вероятностей для несовместных событий.

Вероятность появления одного из нескольких *несовместных* событий равна сумме вероятностей этих событий.

$$\mathbf{P(A_1+A_2+...+A_n) = P(A_1) + P(A_2) +... +P(A_n).}$$

Теорема 5. Теорема сложения вероятностей для совместных событий. Вероятность появления одного из нескольких *совместных* событий равна сумме вероятностей этих событий минус вероятность произведения этих событий:

$$P(A_1 + A_2 + A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1 \cdot A_2 \cdot A_3).$$

Теорема 6. Теорема произведения вероятностей для независимых событий. Вероятность произведения независимых событий равна произведению вероятностей каждого события

$$\mathbf{P(A_1 \cdot A_2 \cdot ... \cdot A_n) = P(A_1) \cdot P(A_2) \cdot ... \cdot P(A_n).}$$

Вероятность произведения зависимых событий вычисляется по формуле условной вероятности.

Задача 1. В урне 30 шаров: 10 красных, 5 синих и 15 белых. Из урны наудачу вытащили один шар. Найти вероятность появления цветного шара.

Решение. Появление цветного шара обозначает появление или красного, или синего шара. Вероятность появления красного шара событие А: P(A)=10/30=1/3.

Вероятность выпадения синего шара (событие В):

P(B)=5/30=1/6.

События А и В несовместны, поэтому применима теорема сложения вероятностей 4.

P(A+B) = P(A) + P(B) = 1/3 + 1/6 = 1/2.

Задача 2. Для отправки груза со склада может быть выделена только одна из двух машин различного вида. Известны вероятности выделения каждой машины: $P(A_1) = 0{,}2$; $P(A_2) = 0{,}4$. Определить вероятность того, что к складу будет подана хотя бы одна из этих машин.

Решение. Поскольку одновременное выделение двух машин для отправки груза – событие невозможное, а это значит, что события «выделение машины первого вида» и «выделение машины второго вида» являются несовместными. Тогда, вероятность того, что к складу будет подана хотя бы одна из этих машин будет: $P(A_1 + A_2) = 0{,}2 + 0{,}4 = 0{,}6$.

Задача 3. Бросают два кубика с нумерованными гранями. Определить вероятность того, что в очередном испытании хотя бы одна из выпавших граней будет четной.

Решение. На каждом кубике – шесть граней. Три из них являются четными, поэтому вероятность выпасть четной грани на первом кубике P(A)=1/2 и на втором кубике P(B)=1/2. Появление четного количества очков на первой и второй кости – события совместные, поэтому вероятность появления четной грани хотя бы на одной кости определяется по теореме сложения совместных событий:

$$P(A+B) = P(A) + P(B) - P(A \cdot B)$$

Вероятность одновременного появления четных граней на обоих костях:

$$P(A \cdot B) = P(A) \cdot P(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

Итак, вероятность появления четной грани хотя бы на одной кости равна

$$P(A+B) = \frac{1}{2} + \frac{1}{2} - \frac{1}{4} = \frac{3}{4}.$$

Задача 4. По мишени стреляют три стрелка. Вероятности попадания соответственно равны 0,7; 0,8 и 0,9. Найти вероятность того, что попадут все три.

Решение. Пусть событие А – «попадание в мишень первого стрелка», событие В – «попадание в мишень второго стрелка», а событие С – «попадание в мишень третьего стрелка». Поскольку эти события независимые, то применяя теорему произведения вероятностей, получим:

$$P(A \cdot B \cdot C) = P(A) \cdot P(B) \cdot P(C) = 0,7 \cdot 0,8 \cdot 0,9 = 0,504.$$

Задача 5. Три стрелка производят по одному выстрелу. Вероятности попадания в цель каждого стрелка равны 0,9; 0,8; 0,85 соответственно. Найти вероятность того, что в цель попадут только два стрелка.

Решение. Искомое событие произойдет, если случатся такие события:

D - попадут первый и второй стрелки, третий промажет;

E - попадут первый и третий стрелки, второй промажет;

F - попадут второй и третий стрелки, первый промажет.

Попадания каждого стрелка в цель не зависят друг от друга, поэтому с помощью теоремы умножения вероятностей для независимых событий можно рассчитать все возможные исходы испытания А, В и С.

Вероятность события D:

$$P(D) = P(A \cdot B \cdot \overline{C}) = 0,9 \cdot 0,8 \cdot 0,15 = 0,108;$$

вероятность события E:

$$P(E) = P(\overline{A} \cdot B \cdot C) = 0,1 \cdot 0,8 \cdot 0,85 = 0,068;$$

вероятность события **F**:

$$P(F) = P(A \cdot \overline{B} \cdot C) = 0,9 \cdot 0,2 \cdot 0,85 = 0,153.$$

События D, E и F несовместимы. Тогда, применяя теорему сложения вероятностей, получим, что вероятность того, что в цель попадут два стрелка

P (D+E+F)= P(D)+ P(E)+ P(F)= 0,108+0, 068+0, 153=0,329.

Статистическое определение вероятности

Иной подход к определению вероятности предлагается, если воспользоваться понятием статистических испытаний, проведенных при условии соблюдения комплекса условий Σ и с фиксацией наступления или не наступления интересующего нас события. Поле событий, таким образом, будет ограниченно только количеством испытаний, а вероятность благоприятного события определяется post factum, т.е. по фактическому результату, в предположении, что все исходы – равновероятны. Дополнительным требованием будет требование независимости каждого последующего испытания от предыдущего. Таким образом, статистическая вероятность фактически является относительной частотой интересующего нас события в серии испытаний.

Относительная частота события A или **статистическая вероятность** определяется равенством $W(A) = \frac{m}{n}$, где **m** – число испытаний, в которых событие **A** наступило; **n** – общее число произведенных испытаний.

Считается, что при достаточно большом количестве испытаний статистическая вероятность (частота) приближается асимптотически к классической вероятности.

Задача 6. Игральная кость брошена десять раз. Шесть очков выпало 3 раза. Какова вероятность и частота выпадения грани с шестью очками?

Решение. Вероятность выпадения шести очков определяется как отношение P(A)=1/6 (из шести возможных исходов при подбрасывании кости выпадению шестерки благоприятствует один), а частота выпадения шести очков равна W(A)=3/10 (событие наступило три раза в десяти испытаниях).

Геометрическая вероятность

Пусть Ω – некоторая область, имеющая меру μ(Ω) (длину, площадь, объем и т. д.), такую, что $0 < \mu(\Omega) < \infty$.

Точка равномерным образом попадает в Ω (реализуется принцип геометрической вероятности), если вероятность P(A) попадания ее в каждую область A, являющейся подобластью Ω, пропорциональна мере этой области μ(A):

$$P(A) = \frac{\mu(A)}{\mu(\Omega)}.$$

Проще говоря, геометрическая вероятность определяется отношением площадей: общей – фигуры, и области, в которую должна попасть точка.

Задача 7. В круг вписан правильный шестиугольник. Найти вероятность того, что точка, наудачу брошенная в круг, не попадет в правильный шестиугольник, вписанный в него.

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

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

Вероятность искомого события:

$$P(A) = \frac{S - s}{S} = (\pi R^2 - \frac{3\sqrt{3} \cdot R^2}{2}) \cdot \frac{1}{\pi R^2} = \frac{\pi - 3\sqrt{3}}{2\pi} = 0{,}174.$$

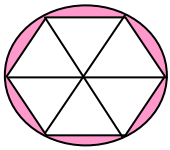


Рис. 2. Правильный шестиугольник, вписанный в круг с радиусом R

Задача 8**.** Два человека договорились встретиться у кинотеатра между 12 и 14 часами и договорились, что тот, кто придет первым, ждет другого в течение 30 минут, после чего уходит. Найти вероятность их встречи, если приход каждого в течение указанного времени может произойти в любой момент, а приходы людей не зависят друг от друга?

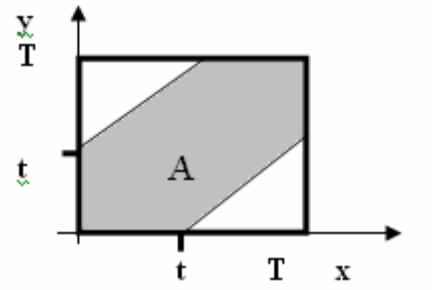


Рис. 3. Графическое представление примера № 4

Решение. Обозначим моменты прихода людей через значения **x** и **y** на соответствующих осях координат. Пусть **T** – интервал времени возможной встречи, равный в примере 120 минутам, а **t** – время ожидания, равное 30 минутам. Ясно, что для того, чтобы встреча произошла, нужно, чтобы разность между моментами прихода людей была меньше или равна 30 мин. Изобразим **x** и **y** как точки внутри квадрата со сторонами **T**. Тогда исходы, благоприятные для встречи, будут соответствовать заштрихованной области A. Согласно принципу геометрической

вероятности, искомая вероятность равна отношению площади заштрихованной фигуры к площади всего квадрата:

$$P(A) = \frac{T^2 - (T - t^2)}{T^2}.$$

Элементы комбинаторики в теории вероятностей

Многие задачи, основанные на классическом подходе, и как мы уже знаем, понятии равновозможности, сводятся к вычислению количества всех возможных событий, наступающих при выполнении комплекса условий Σ, а также количества интересующих нас благоприятных событий. Такие вычисления достаточно часто сводятся к расчету количества различных комбинаций элементов из какого-либо множества. В этом случае мы имеем дело с разделом математики – **комбинаторикой**. Введем основные определения и соотношения комбинаторики.

Пусть имеется ***k*** групп элементов, причем ***i***-я группа состоит из n_i элементов. Выберем по одному элементу из каждой группы. Тогда общее число N способов, которыми можно произвести такой выбор, определяется соотношением:

$\mathbf{N = n_1 \cdot n_2 \cdots n_k}$ – основная формула комбинаторики.

Пусть имеется 4 аудитории, в которых стоят компьютеры. В первой аудитории $\mathbf{n_1}$=12 компьютеров, во второй $\mathbf{n_2}$=10 компьютеров, в третьей $\mathbf{n_3}$=11 компьютеров, а в четвертой $\mathbf{n_4}$=13 компьютеров. Для расчета количества комбинаций этих компьютеров, взятых из каждого класса по одному, нужно рассчитать

$$N = 12 \cdot 10 \cdot 11 \cdot 13 = 17160.$$

Перестановкой из **n** элементов называется любой **упорядоченный** набор этих элементов. Например, перестановки чисел 1, 2, 3: (1,2,3), (1,3,2), (2,1,3), (2,3,1), (3,1,2), (3,2,1).

$$\mathbf{P_n = n(n-1) \cdot \cdot \cdot 2 \cdot 1 = n!}$$

Размещением из **n** элементов по **m** называется любой **упорядоченный** набор из **m** различных элементов, выбранных из общей совокупности в **n** элементов.

Например, размещения из четырех чисел 1, 2, 3, 4 по два:

(1,2), (1,3), (1,4), (2,1), (2,3), (2,4), (3,1), (3,2), (3,4), (4,1), (4,2), (4,3).

$$A_n^m = (n)_m = n(n-1) \cdots (n-m+1) = \frac{n!}{(n-m)!}.$$

Сочетанием из **n** элементов по **m** называется любой **неупорядоченный** набор из **m** различных элементов, выбранных из общей совокупности в **n** элементов.

Сочетаниями из четырех чисел 1, 2, 3, 4 по два:

(1,2), (1,3), (1,4), (2,3), (2,4), (3,4).

$$C_n^m = \binom{n}{m} = \frac{(n)_m}{m!} = \frac{n!}{(n-m)!\,m!}.$$

Задача 9. (*перестановка*). На пяти карточках написаны буквы “п”, “л”, “а” ,“м”, “а”. Карточки перемешиваются и выкладываются в ряд. Найти вероятность того, что образовавшееся слово будет “лампа” (событие А).

Решение. В соответствии с комбинаторными принципами для определения общего числа элементарных исходов нужно подсчитать число упорядоченных наборов из четырех букв. Мы имеем дело с числом перестановок, поэтому число элементар-

ных исходов ***n = 5! = 120***. Слово “лампа” образует две перестановки, то есть число благоприятных для события **А** элементарных исходов ***m = 2***. Поэтому ***P(A) = m/n = 2/120=1/60***.

Задача 10**.** Набирая номер телефона, абонент забыл последние две цифры и, помня лишь, что эти цифры различны, набирал их наудачу. Найти вероятность того, что набраны нужные цифры.

Решение. Обозначим через через В событие – набраны две нужные цифры. Всего можно набрать столько различных цифр, сколько может быть составлено размещений из десяти цифр по две, то есть

A^2_{10}= n!/(n-m)!=10!/8!=90. Тогда P(B)=1/90.

Задача 11. В коробке сидят 12 котят. Среди них три белых котенка, три черных, три серых и три рыжих. Из коробки наудачу вытаскивают трех котят. Определить вероятность вытащить из коробки трех котят так, чтобы один из них был непременно рыжий, а два других имели различный окрас и не были бы рыжими.

Решение. Общее число элементарных исходов испытания равно числу сочетаний из двенадцати элементов по три, то есть С из 12 по 3.

$$C_{12}^3 = \frac{12!}{(12-3)!\cdot 3!} = \frac{12!}{9!\cdot 3!} = \frac{10\cdot 11\cdot 12}{6} = 220.$$

Число исходов, благоприятствующих появлению двух котят различного окраса (и не рыжих), равно числу сочетаний из одиннадцати элементов по два, то есть С из 11 по 2.

$$C_{11}^2 = \frac{11!}{(11-2)!\cdot 2!} = \frac{11!}{9!\cdot 2} = 55.$$

Искомая вероятность равна отношению числа исходов, благоприятствующих интересующему нас событию, к общему числу возможных элементарных исходов:

$$P(A) = \frac{C_{11}^2}{C_{12}^3} = \frac{55}{220} = \frac{1}{4}.$$

Задача 12. В партии из 10 деталей 7 стандартных. Найти вероятность того, что среди взятых наудачу шести деталей четыре (4) оказались стандартными.

Решение. Общее число возможных исходов испытания равно числу способов, которыми можно извлечь 6 деталей из десяти, то есть числу сочетаний из 10 элементов по 6 – количество сочетаний C_{10}^6.

Число благоприятствующих исходов (среди 6 взятых деталей 4 стандартных). Четыре стандартных деталей можно взять из 7 стандартных C_7^4 способами, при этом остальные 6-4=2 должны быть нестандартными. Взять же 2 нестандартные детали из 10-7=3 нестандартных можно C_3^2 способами. Следовательно, число благоприятствующих исходов равно $C_7^4 * C_3^2$. Искомая вероятность равна отношению числа исходов, благоприятствующих событию, к числу всех элементарных исходов: $P(A)= C_7^4 * C_3^2 / C_{10}^6 =1/2$.

Задача 13. Известно, что в поступившей партии из 30 швейных машинок 10 имеют внутренний дефект. Определить вероятность того, что из пяти наудачу взятых машинок три окажутся бездефектными.

Решение. Введем следующие обозначения: N – общее число машинок, n – число бездефектных машинок, m – число отобранных в партию машинок, k – число бездефектных машинок в отобранной партии.

Общее число комбинаций по m машинок (общее число возможных исходов) будет равно числу сочетаний из N элементов по m, т. е. $C_N^m = C_{30}^5$. Но в каждой отобранной комбинации должно содержаться по три бездефектные машинки. Число таких комбинаций равно числу сочетаний из n элементов по k, т. е. $C_n^k = C_{20}^3$.

Оставшиеся дефектные машинки (элементы) тоже образуют множество комбинаций, число которых равно числу сочетаний из

$N - n$ элементов по $m - k$, т. е. $C_{N-n}^{m-k} = C_{10}^2$.

Это значит, что общее число благоприятствующих исходов определяется произведением $C_n^k \cdot C_{N-n}^{m-k}$.

$$C_{20}^3 \cdot C_{30-20}^{5-3} = C_{20}^3 \cdot C_{10}^2$$

Подставив в эту формулу численные значения данного примера, получим

$$P(A) = \frac{C_{20}^3 \cdot C_{10}^2}{C_{30}^5} = \frac{20!\cdot 10!}{17!\cdot 3!\cdot 8!\cdot 2!} \cdot \frac{15!\cdot 5!}{30!} \approx 0{,}36.$$

Задачи для самостоятельного решения

Классическая вероятность:

1. В магазине ежемесячно составляется отчет о продажах, который должен быть представлен руководству в последний день месяца, независимо от того является этот день выходным или нет. Какова вероятность того, что отчет будет представлен: а) 28– го числа? б)30 – го числа? в)31- го числа? Год является невисокосным.

2. Брошены две игральные кости. Найти вероятность того, что сумма очков на выпавших гранях равна пяти, а произведение – четырем.
3. Участники жеребьевки тянут из ящика жетоны с номерами от 1 до 100. Найти вероятность того, что номер первого наудачу извлеченного жетона не содержит цифры 5.
4. Куб, все грани которого окрашены, распилен на тысячу кубиков одинакового размера, которые затем тщательно перемешаны. Найти вероятность того, что наудачу извлеченный кубик будет иметь окрашенных граней: а) одну; б) две; в) три.
5. В замке на общей оси пять дисков. Каждый диск разделен на шесть секторов, на которых написаны различные буквы. Замок открывается только в том случае, если каждый диск занимает одно определенное положение относительно корпуса замка. Найти вероятность того, что при произвольной установке дисков замок можно будет открыть.
6. Из тщательно перемешанного полного набора 28 костей домино наудачу извлечена кость. Найти вероятность того, что вторую наудачу извлеченную кость можно приставить к первой, если первая кость: а) оказалась дублем; б) не есть дубль.

Классическая вероятность с использованием формул комбинаторики:

1. Бросают две одинаковые монеты. Какова вероятность комбинации "орла" и "решки" и комбинации два "орла"?
2. Из 28 костей домино наугад выбирают одну. Что вероятнее: что сумма цифр на ней будет что сумма цифр на ней будет равна 6 или 8 (равна 3 или 4)?
3. Какова вероятность того, что дата рождения человека приходится на 29 февраля?

4. Из букв слова «апельсин» последовательно выбирают 4 буквы. Найдите вероятность того, что выбранные буквы в порядке их выбора образуют слово: «лиса»? «плен»?
5. Какова вероятность того, что при случайном расположении в ряд кубиков, на которых написаны буквы “а”, “г”, “и”, “л”, “м”, “о”, “р”, “т” получится слово “алгоритм”?
6. В коллекции одного из известных ученых МГПИ 200 монет, из которых 25 монет XVIII века. Какова вероятность того, что наудачу выбранная монета датирована XVIII веком?
7. Библиотечка состоит из десяти различных книг, причем пять книг стоят по 4 рубля каждая, три книги по одному рублю и две книги по 3 рубля. Найти вероятность того, что взятые наудачу две книги стоят 5 рублей. (*Отв.* $p = C^1_5 * C^1_3 / C^2_{10} = 1/3$).
8. В ящике имеется 50 одинаковых деталей, из них 5 окрашенных. Наудачу вынимают одну деталь. Найти вероятность того, что извлеченная деталь окажется окрашенной.
9. Брошена игральная кость. Найти вероятность того, что выпадет четное число очков.
10. Участники жеребьевки тянут из ящика жетоны с номерами от 1 до 100. Найти вероятность того, что номер первого наудачу извлеченного жетона не содержит цифры 5.
11. В мешочке имеется 5 одинаковых кубиков. На всех гранях каждого кубика написана одна из следующих букв: о, п, р, с, т. Найти вероятность того, что на вынутых по одному и расположенных "в одну линию" кубиков можно будет прочесть слово "спорт".
12. На каждой из шести одинаковых карточек напечатана одна из следующих букв: а, т, м, р, с, о. Карточки тщательно перемешаны. Найти вероятность того, что на четырех, вынутых по

одной и расположенных "в одну линию" карточках, можно будет прочесть слово "трос".

13. Куб, все грани которого окрашены, распилен на тысячу кубиков одинакового размера, которые затем тщательно перемешаны. Найти вероятность того, что наудачу извлеченный кубик будет иметь окрашенных граней: а) одну; б) две; в) три.
14. Из тщательно перемешанного полного набора 28 костей домино наудачу извлечена кость. Найти вероятность того, что вторую наудачу извлеченную кость можно приставить к первой, если первая кость: а) оказалась дублем; б) не есть дубль.
15. В замке на общей оси пять дисков. Каждый диск разделен на шесть секторов, на которых написаны различные буквы. Замок открывается только в том случае, если каждый диск занимает одно определенное положение относительно корпуса замка. Найти вероятность того, что при произвольной установке дисков замок можно будет открыть.
16. Восемь различных книг расставляются наудачу на одной полке. Найти вероятность того, что две определенные книги окажутся поставленными рядом.
17. Библиотечка состоит из десяти различных книг, причем пять книг стоят по 4 рубля каждая, три книги – по одному рублю и две книги – по 3 рубля. Найти вероятность того, что взятые наудачу две книги стоят 5 рублей. В партии из 100 деталей отдел технического контроля обнаружил 5 нестандартных деталей. Чему равна относительная частота появления нестандартных деталей?
18. При стрельбе из винтовки относительная частота попадания в цель оказалась равной 0,85. Найти число попаданий, если всего было произведено 120 выстрелов.

Геометрическая вероятность

1. На отрезок OA длины L числовой оси Ox наудачу поставлена точка B (x). Найти вероятность того, что меньший из отрезков OB и BA имеет длину, меньшую чем L/3.
2. Внутрь круга радиуса R наудачу брошена точка. Найти вероятность того, что точка окажется внутри вписанного в круг квадрата. Предполагается, что вероятность попадания точки в квадрат пропорциональна площади квадрата и не зависит от его расположения относительно круга.
3. Какова вероятность того, что наудачу поставленная в данном круге точка окажется внутри вписанного в него квадрата?
4. Два лица условились встретиться в определенном месте между 10 и 11 часами и договорились, что пришедший первым ждет другого в течение 15 минут, после чего уходит. Найти вероятность их встречи, если приход каждого в течение указанного часа может произойти в любое время и моменты прихода независимы?
5. После бури на участке между 40-м и 70-м километрами телефонной линии произошел обрыв провода. Какова вероятность того, что обрыв провода произошел между: 50-м и 55-м километрами? 60-м и 66-м километрами?
6. В круг случайным образом бросают две точки. Найдите вероятность того, что обе точки окажутся внутри вписанного в этот круг правильного шестиугольника? треугольника?

Условная вероятность. Полная вероятность. Формула Байеса

Цель: *Сформировать у студентов понятия зависимых и независимых событий, условной и полной вероятности, научить решать задачи, в которых определяются полная вероятность и вероятность гипотезы.*

При совместном рассмотрении двух случайных событий A и B часто возникает вопрос: насколько связаны эти события друг с другом, в какой мере наступление одного из них влияет на возможность наступления другого?

Событие B называют *независимым от события* A, если появление события A не изменяет вероятности события B.

Произведением двух событий A и B называют событие AB, состоящее в совместном появлении (совмещении) этих событий. Например, если A – деталь годная, B – деталь окрашенная, то AB – деталь годна и окрашена.

Теорема произведения вероятностей. Вероятность совместного появления нескольких независимых событий равна произведению вероятностей этих событий:

$$\mathbf{P\ (A \cdot B \cdot C) = P(A) \cdot P(B) \cdot P(C).}$$

Пример 1. Три человека договорились о встрече в определенном месте, в определенный час. Известны вероятности прихода на встречу каждого человека: $P(A_1) = 0{,}2$; $P(A_2) = 0{,}4$; $P(A_3) = 0{,}5$. Определить вероятность того, что все три человека придут на встречу.

Решение:

Вероятность того, что все эти люди придут на встречу, будет определяться по формуле произведения вероятностей независимых событий:

$$\mathbf{P\ (A \cdot B \cdot C) = 0{,}2 \cdot 0{,}4 \cdot 0{,}5 = 0{,}04.}$$

Случайное событие – это событие, которое при осуществлении совокупности условий S может произойти или не произойти. Если при вычислении вероятности события никаких других ограничений, кроме условий S, не налагается, то такую вероятность называют ***безусловной*;** если же налагаются и другие дополнительные условия, то вероятность события называют ***условной.*** Например, часто вычисляют вероятность события B при ***дополнительном условии,*** что произошло событие A. Заметим, что и безусловная вероятность, строго говоря, является условной, поскольку предполагается осуществление условий S.

Условной вероятностью P_A (B) называют вероятность события B, вычисленную в предположении, что событие A уже наступило.

Условная вероятность события B при условии, что событие A уже наступило, по определению, равна

$$P_A(B) = \frac{P(A \cdot B)}{P(A)}, при$$

$$P(A) \neq 0.$$

Рассмотрим два события: A и B; пусть вероятности P (A) и P_A (B) известны. Как найти вероятность совмещения этих событий, т. е. вероятность того, что появится и событие A и событие B? Ответ на этот вопрос дает теорема умножения зависимых друг от друга событий.

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

$$\mathbf{P\,(AB) = P\,(A)\,P_A\,(B)} \quad (1)$$

Доказательство

Применив формулу (1) к событию BA, получим

$$\mathbf{P\,(BA) = P\,(B)\,P_B\,(A),}$$

или, поскольку событие BA не отличается от события AB,

$$\mathbf{P\,(AB) = P\,(B)\,P_B\,(A)} \quad (2)$$

Сравнивая формулы (1) и (2), заключаем о справедливости равенства

$$\mathbf{P\,(A)\,P_A\,(B) = P\,(B)\,P_B\,(A)} \quad (3)$$

Пример 2. У продавца имеется 3 красных воздушных шарика и 7 синих шариков. Продавец взял наугад один шарик из мешка, а затем второй. Найти вероятность того, что первый из взятых шариков – красный, а второй – синий.

Решение. Вероятность того, что первый шарик окажется красным

(событие A), P (A) = 3/10.

Вероятность того, что второй шарик окажется синим (событие B), вычисленная в предположении, что первый шарик – красный, т. е. условная вероятность P_A (B) = 7 / 9.

По теореме умножения, искомая вероятность

P (AB) = P (A) P_A (B) = (3 / 10) * (7 / 9) = 7 / 30.

Заметим, что сохранив обозначения, легко найдем:

P (B) = 7/10,

$P_B(A) = 3/9$,

$P(B)P_B(A) = 7/30$,

что наглядно иллюстрирует справедливость равенства (3).

Пример 3. В урне 5 белых, 4 зеленых и 3 синих кубика. Каждое испытание состоит в том, что наудачу извлекают один кубик, не возвращая его обратно. Найти вероятность того, что при первом испытании появится белый кубик (событие А), при втором – зеленый (событие В) и при третьем – синий (событие С).

Решение. Вероятность появления белого кубика в первом испытании

$P(A) = 5/12$.

Вероятность появления зеленого кубика во втором испытании, вычисленная в предположении, что в первом испытании появился белый кубик, т. е. условная вероятность

$$P_A(B) = 4 / 11.$$

Вероятность появления синего кубика в третьем испытании, вычисленная в предположении, что в первом испытании появился белый кубик, а во втором – зеленый, т. е. условная вероятность

$P_{AB}(C) = 3 / 10$.

Искомая вероятность:

$P(ABC) = P(A)\, P_A(B)\, P_{AB}(C) = (5/12) * (4/11) * (3/10) = 1/22$.

Формула полной вероятности

Предположим, что событие А может наступить только вместе с одним из нескольких попарно несовместных событий

$$\mathbf{H_1, H_2, ..., H_n.}$$

Условимся называть эти события по отношению к А гипотезами.

Общая гипотеза – это научно обоснованное предположение о законах и закономерностях природных и общественных явлений, а также закономерностях психической деятельности человека. Приведем примеры некоторых гипотез.

- Гипотеза Демокрита: «Вещество состоит из атомов».
- Гипотеза биохимической эволюции**:** «Жизнь это результат длительной эволюции углеродных соединений» (Авторы биохимик А. И. Опарин в 1924 г. и Дж. Холдейном в 1929 г.).
- Гипотеза большого взрыва: «Наша вселенная произошла около 13 млрд лет назад в результате взрыва чрезвычайно плотного и горячего вещества – космологической сингулярности».

Для определения полной вероятности события используются вероятности гипотез и условные вероятности событий.

Формула полной вероятности:

$$\mathbf{p(A) = p(A/H_1)p(H_1) + p(A/H_2)p(H_2) + ... + p(A/H_n)p(H_n).}$$

Вероятность события А равна сумме произведений условных вероятностей этого события по каждой из гипотез на вероятность самих гипотез.

Пример 4. Имеются три урны. В первой находятся 5 белых и 3 черных шара, во второй – 4 белых и 4 черных, в третьей – 8 белых. Наугад выбирается одна из урн и из нее вытаскивается один шар. Какова вероятность того, что он окажется черным (событие А)?

Решение.

Шар может быть вытащен из первой урны, либо из второй, либо из третьей; обозначим эти события H_1, H_2, H_3. Так как имеются одинаковые шансы выбрать любую из урн, то

$$p(H_1) = p(H_2) = p(H_3) = 1/3.$$

Далее находим вероятности события А при каждом из условий H_1, H_2, H_3:

$$p(A/H_1) = 3/8,\ p(A/H_2) = 4/8,\ p(A/H_3) = 0/8.$$

Отсюда: $p(A) = p(A/H_1)\,p(H_1) + p(A/H_2)\,p(H_2) + p(A/H_3)\,p(H_3)$

$$= (1/3)\cdot(3/8) + (1/3)\cdot(4/8) + (1/3)\cdot(0/8) = 7/24.$$

Во многих задачах на полную вероятность, рассматриваемый опыт можно представить происходящим в два этапа; гипотезы H_1, H_2, ..., H_n исчерпывают все возможные предположения относительно исхода первого этапа, событие же А есть один их возможных исходов второго этапа. В рассмотренном примере первый этап заключался в выборе урны, второй – в извлечении из нее шара.

Формула Т. Байеса

Формула священника и математика **Томаса Байеса** была опубликована в 1764 г. Ученые и религиозные деятели того времени искали ответы на сложнейшие вопросы мироздания. Среди них: возникновение и устройство космоса, эволюция, появление человека – на многие трудные вопросы должен был быть найден математический ответ. Результатом таких исследований и стала формула Т. Байеса, которая используется в теории вероятностей до сих пор.

Формула Байеса относится к той же ситуации, что и формула полной вероятности, но определяет вероятность гипотезы, которая привела к наступлению события. Событие А может наступить только вместе с одним из попарно несовместных собы-

тий H_1, H_2, ... H_n. Пусть произведен опыт, в результате которого произошло событие А. Сам по себе этот факт еще не позволяет сказать, какое из событий H_1, H_2, ..., H_n имело место в проделанном опыте. Поставим задачу: найти вероятности $p(H_i/A)$ каждой из гипотез в предположении что событие А наступило. Эта задача решается при помощи формулы Байеса.

В примере из предыдущего раздела вероятность гипотезы H_3 – шар извлечен из третьей урны – до того, как произведен опыт, равнялась 1/3. Однако, если опыт произведен и наступило событие А – вытащенный шар оказался черным, то это снижает шансы гипотезы H_3 до нуля. Послеопытная, "апостериорная" вероятность гипотезы H_3 будет в данном случае ниже, чем доопытная, "априорная".

$$p(H_i / A) = \frac{p(H_i)\,p(A / H_i)}{\sum_{j=1}^{n} p(H_j)\,p(A / H_j)}$$ – формула Байеса.

Вывод формулы весьма прост – из формулы полной вероятности.

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

$$p(H_iA) = p(H_i / A)\,p(A).$$

Приравнивая правые части, получим:

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

откуда следует:

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

или, если воспользоваться формулой полной вероятности,

$$p(H_i / A) = \frac{p(A / H_i)}{p(A / H_1)p(H_1) + p(A / H_2)p(H_2) + ... + p(A / H_n)p(H_n)}.$$

Пример 5. В студенческом стройотряде 2 бригады первокурсников и одна – второкурсников. В каждой бригаде первокурсников 5 юношей и 3 девушки, а в бригаде второкурсников 4 юношей и 4 девушки. По жеребьевке из отряда выбрали одну из бригад и из нее одного человека для поездки в город. а) Какова вероятность того, что выбран юноша? б) Выбранный человек оказался юношей. Какова вероятность, что он первокурсник?

Решение

Обозначим через *A* событие – выбор группы студентов для поездки в город. Можно выдвинуть две гипотезы:

H1 – выбрана группа первокурсников, а поскольку первокурсников 2 группы, то P(H1) = 2/3;

H2 – выбрана группа второкурсников, причем P(H2) = 1/3.

Условная вероятность того, что выбранный человек является юношей первокурсником: P(A/H1) = 5/8.

Условная вероятность того, что выбранный человек является юношей второкурсником: P(A/H2) = 1/2.

Вероятность того, что наудачу выбранный человек – юноша определяется по формуле полной вероятности:

P(A) = P(A/H1)P(H1) + P(A/H2)P(H2) = 5/8*2/3 + 1/2* 1/3 =7/12.

Искомая вероятность того, что выбранный человек – юноша с первого курса определяется по формуле Байеса:

$$P(H_1 / A) = \frac{P(H_1)P(A / H_1)}{P(A)} = \frac{\frac{2}{3} \cdot \frac{5}{8}}{\frac{7}{12}} = \frac{5}{7}.$$

Задачи для самостоятельного решения

1. В ящике лежат 12 белых, 8 черных и 10 красных шаров. Какова вероятность того, что наугад выбранный шар: будет красным, если известно, что он не черный? будет черным, если известно, что он не белый?

2. На заводе 50% деталей типа А1 производит рабочий Орлов, 30% – рабочий Чайкин и 20% – рабочий Воронин. Вероятность брака у этих рабочих составляет 5%, 3%, и 2% соответственно. Из партии деталей наугад выбирается одна. Найдите вероятность того, что эта деталь: 1) качественная; 2) бракованная; 3) бракованная и изготовлена Орловым? 4) качественная и изготовлена Чайкиным?

3. В цехе 10 станков марки А, 6 – марки В и 4 – марки С. Вероятность выпуска качественной продукции для каждого станка составляет 0,9; 0,8 и 0,7 соответственно. Какой процент качественной бракованной продукции выпускает цех в целом?

4. Число грузовых автомашин, проезжающих по шоссе, на котором стоит бензоколонка, относится к числу легковых машин, проезжающих по тому же шоссе, как 3 : 2. Вероятность того, что будет заправляться грузовая машина, равна 0,1; для легковой машины эта вероятность равна 0,2. К бензоколонке подъехала для заправки машина. Найти вероятность того, что это грузовая машина.

5. В пирамиде пять винтовок, три из которых снабжены оптическим прицелом. Вероятность того, что стрелок поразит мишень при выстреле из винтовки с оптическим прицелом, равна 0,95; для винтовки без оптического прицела эта вероятность равна 0,7. Найти вероятность того, что мишень будет поражена, если стрелок произведет один выстрел из наудачу взятой винтовки.

6. Из 20 студентов, пришедших на экзамен, 8 подготовлены отлично, 6 – хорошо, 4 – посредственно и 2 – плохо. В экзаменационных билетах имеется 40 вопросов. Студент, подготовленный отлично, знает все вопросы, хорошо – 35, посредственно – 25 и плохо – 10 вопросов. Некоторый студент ответил на все 3 вопроса билета. Найдите вероятность того, что он подготовлен: а) хорошо; б) плохо.

7. Из 100 девушек, посетивших салон красоты, 45 – блондинки, 20 – брюнетки, 30 – шатенки и 5 – рыжие. Салон оказывает посетителям 24 различных услуги. Блондинки в среднем пользуются 8 услугами, брюнетки – 5, шатенки – 4, рыжие – 7 услугами. В салон зашла клиентка и заказала несколько услуг. Найдите вероятность того, что она: а) блондинка; б) брюнетка, в) шатенка, г) рыжая.

Контрольные вопросы

1. Какое событие называется зависимым?
2. Приведите примеры независимых событий.
3. Объясните разницу между классическим и статистическим подходами в определении вероятности. Приведите пример задачи, в которой классическая и статистическая вероятность имеют разные значения.
4. Как определяется полная вероятность события?
5. Что такое условная вероятность события?
6. Как определяется вероятность гипотезы?
7. Что называется геометрической вероятностью события?
8. Как определить геометрическую вероятность события?

Распределения дискретных случайных величин

Цель: *получить представление о случайных величинах, освоить законы распределения дискретных случайных величин.*

Все процессы, происходящие в природе, делятся на непрерывные и дискретные. Примерами непрерывных процессов являются различные природные объекты и их свойства: температура, давление и влажность воздуха, объекты технологических производственных процессов: давление и температура теплоносителя в ядерном реакторе.

В определенный момент времени непрерывная случайная величина может быть выражена в численной форме, но впоследствии это значение будет непрерывно изменяться.

Дискретными являются сигналы тревоги, языковые сообщения в виде звука и письма, жесты и т.п. Например, такие величины, как количество человек в студенческой группе, число солнечных дней в году, высота горы, уровень интеллекта являются дискретными величинами, потому что имеют конкретный количественный признак, который некоторое время не изменяется.

Для обработки непрерывной и дискретной информации существуют и разные вычислительные машины: аналоговые и цифровые. Аналоговые машины следят за постоянно изменяющимся аналоговым сигналом. Например, в отделениях интенсивной терапии такие ЭВМ могут измерять давление, снимать кардиограмму и др.

Рассмотрим вероятностное пространство (Ω, σ, P), то *есть пространство элементарных исходов* Ω, σ-*алгебру событий* (определенную нами на пространстве путем введения замкнутых операций), *вероятность P* (как меру нашего множества). Множества вида $\{\omega : \xi(\omega) = x\} \subset \Omega$ являются событиями.

Случайной называют величину, которая в результате испытания примет одно и только одно возможное значение, заранее не известное и зависящее от случайных причин, которые не могут быть заранее учтены.

Например, число родившихся мальчиков среди 100 новорожденных есть случайная величина, которая имеет возможные значения: 0,1,2,3…100.

Расстояние, которое пролетит снаряд при выстреле орудия – есть случайная величина, которая зависит от прицела, силы и направления ветра, температуры воздуха. Возможные значения этой величины принадлежат промежутку (a,b).

Далее будем обозначать случайные величины прописными буквами X, Y, Z, а их возможные значения x,y,z. Например случайная величина X имеет три возможных значения x_1,x_2,x_3.

Случайной величиной *ξ называется произвольная функция, ставящая в соответствие каждому элементарному исходу (событию) ω число $\xi = \xi(\omega)$.*

Так как Ω - ограничено своим набором возможных событий, то случайная величина ξ принимает не более чем счетное число значений: $x_1,...x_k,...$

Распределением дискретной случайной величины ξ назовем таблицу 2.

Таблица 2

Распределением дискретной случайной величины

ξ	x_1	x_2	...	x_k	...
P	p_1	p_2	...	p_k	...

где $p_k = P\{\omega : \xi(\omega) = x_x\} = P\{\xi = x_k\}$.

Таким образом, с точки зрения функционального анализа, случайная величина представляет собой обычную числовую функцию, заданную на пространстве элементарных исходов (событий) Ω. Специфика теории вероятностей проявляется в том, что на пространстве Ω задана также вероятность P.

Пример 1. Два игрока играют в “орлянку” на следующих условиях: если при подбрасывании монеты выпадает “орел”, то первый игрок платит второму $1, если “решка”, то второй игрок платит первому $2. Опишем случайную величину ξ, равную выигрышу первого игрока в этой игре (при одном подбрасывании монеты).

Решение. Пространство элементарных исходов (событий) Ω состоит из двух исходов: ω_1 – выпадение “орла” и ω_2 – “решки”. σ-Алгебра событий насчитывает 4 события: ∅, $\{\omega_1\}$, $\{\omega_2\}$, Ω. Это следует из аксиом Колмогорова. Предполагая, что монета симметричная, найдем вероятности всех событий из множества алгебры событий: $P(\varnothing) = 0$, $P(\omega_1) = 1/2$, $P(\omega_2) = 1/2$, $P(\Omega) = 1$. Вероятностное пространство – определено.

Вероятностное пространство, как было определено выше, включает в себя пространство элементарных событий, σ - алгебру, вероятность Р – как меру, ограничение.

Случайная величина ξ принимает значения: -1, если выпал "герб" ($\xi(\omega_1) = -1$), и 2, если выпала "цифра" ($\xi(\omega_2) = 2$).

Таблица 3

Значения случайной величины

Элементарные исходы	ω_1	ω_2	∅	Ω
$\xi(\omega)$	-1	2		
P(ω)	½	½	0	1

Функция распределения случайной величины

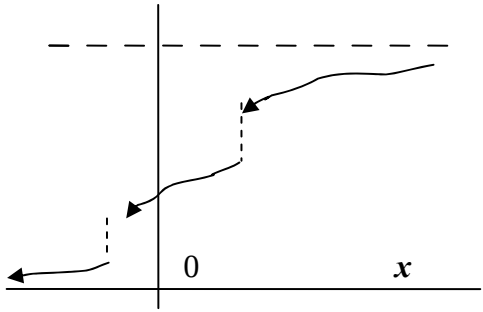


Рис. 4. Функция распределения случайной величины

Функцией распределения *(вероятностей) случайной величины ξ называется функция F(x), значение которой в точке x равно вероятности события { ξ < x }, т.е. события, состоящего из тех и только тех элементарных исходов ω, для которых* $\xi(\omega)<x$:

$$F(x) = P\{\xi < x\}.$$

Или говорят, что значение функции распределения в точке x равно вероятности того, что случайная величина ξ примет значение, меньшее x.

Свойства функции распределения

1. Функция F(x) является ограниченной, то есть ее значения лежат в интервале от 0 до 1.

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

2. Функция F(x) является неубывающей. Если $\boldsymbol{x_2 > x_1}$, то $\mathbf{F(x_2) \ge F(x_1)}$, так как вероятность любого события неотрицательна.

3. Поскольку событие $\{\xi < -\infty\}$ является невозможным, а событие $\{\xi < \infty\}$ – достоверным, то имеем

$$F(-\infty) = \lim_{x \to -\infty} F(x) \quad \text{и} \quad F(+\infty) = \lim_{x \to +\infty} F(x) .$$

4. Вероятность попадания случайной величины ξ на отрезок (x_1, x_2) определяется формулой: $\mathbf{P\{x_1 \le \xi \le x_2\}=F(x_2) - F(x_1).}$

Событие $\{\xi < x_2\}$ при $x_1 < x_2$ представляет собой объединение двух непересекающихся событий: $\{\xi < x_1\}$ – случайная величина ξ приняла значение, меньшее x_1, и $\{x_1 \le \xi \le x_2\}$ – случайная величина приняла значение, лежащее в интервале (x_1, x_2). Поэтому из аксиомы сложения получаем:

$$\mathbf{P\{x_1 \le \xi \le x_2\}=F(x_2) - F(x_1).}$$

Зная функцию распределения F(x), можно однозначно определить вероятность попадания случайной величины ξ не только на интервал [x1, x2[, но и в любое множество на прямой.

Итак, с любой случайной величиной связана ее функция распределения. После общего определения функции распреде-

ления случайной величины, перейдем к частным случаям – дискретной и непрерывной функциям распределения.

Дискретные случайные величины

Случайные величины могут быть дискретными т.е. принимать только конечное или счетное множество определенных значений (например, число очков при бросании игральной кости; число телефонных звонков, поступающих конкретному абоненту в течение суток). У таких величин F(x) имеет разрывы в точках, соответствующих принимаемым значениям. Такие величины удобнее характеризовать указанием возможных значений и их вероятностей.

Таблица 4

Ряд распределения дискретной случайной величины числа выпавших очков при бросании кости

Значения x_i:	0	1	2	3	4	5	6
Вероятности $p(x_i)$	0	1/6	1/6	1/6	1/6	1/6	1/6

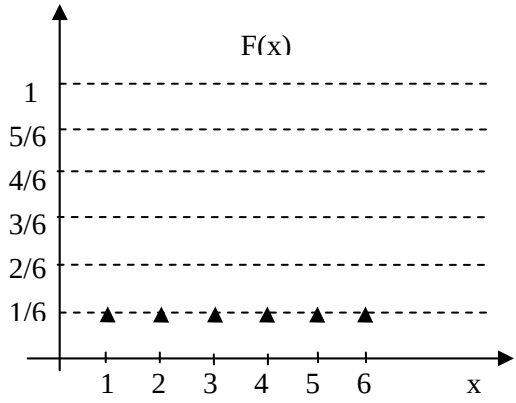


Рис. 5. Функция распределения вероятностей выпадения очков при бросании кости

Рис. 6. Кумулятивная вероятность распределения числа очков при бросании кости

Хотя случайная величина принимает только дискретные значения, ее функция распределения определена для любых x.

Например: $F(-1) = 0$, $F(0) = 0$, $F(0.999) = 0$, $F(1.001) = 1/6$, $F(3.5) = 3/6$, $F(7) = 1$.

Дискретную случайную величину удобно характеризовать *рядом распределения*.

Таблица 5

Ряд распределения

ξ	X_1	X_2	$\cdots$	X_i	$\cdots$	X_n
P	p_1	p_2	$\cdots$	p_i	$\cdots$	p_n

ξ – все возможные значения случайной величины.

P – вероятности $p_i = P\{\xi = X_i\}$, того, что случайная величина примет эти значения.

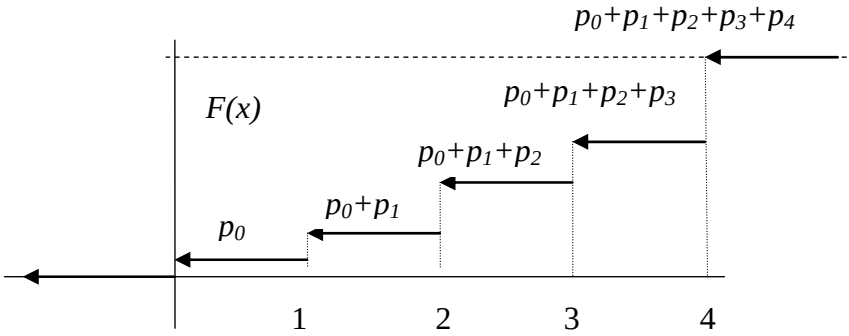


Рис. 7. Распределение вероятностей дискретной случайной величины

Пример 2. Дискретное распределение.

В неком обществе организована лотерея. Разыгрываются две вещи стоимостью по $10 и одна стоимостью $30. Составить закон распределения суммы чистого выигрыша для субъекта, который приобрел один билет за $1; всего продано 50 билетов.

Решение. Искомая случайная величина X может принимать три значения: -1, (если субъект не выиграет, а фактически проиграет $1, уплаченный за билет); $9, $29. Первому результату благоприятны 47 случаев из 50, второму – 2 из 50, третьему – 1 из 50. Следовательно, вероятности, соответствующие этим случаям равны: *P(X=-1) = 47/50 = 094; P(X=9) = 2/50 = 0,04; P(X=29) = 1/50 = 0,02.*

Закон распределения X имеет вид:

Сумма выигрыша	-1	9	29
Вероятность	0,94	0,04	0,02

Виды дискретных функций распределения

Биноминальное распределение

Биноминальное распределение является распределением числа успехов μ в n испытаниях Бернулли с вероятностью успеха p и неудачи $q = 1 - p$.

Схема Бернулли

Рассмотрим последовательность независимых одинаковых испытаний: появление или не появление некоторого наблюдаемого события в каждом испытании не будет зависеть от исходов предыдущих испытаний. Это и есть *схема Бернулли*.

Опыт состоит в n-кратном повторении одинаковых испытаний, в каждом из которых может с вероятностью p наступить некоторое событие (будем говорить в этом случае, что произошел "успех") или с вероятностью $q = 1 - p$ не наступить (произошла "неудача"). Результат каждого опыта можно записать в виде последовательности УНН...У, "У" – успех, "Н" – неудача. Пространство элементарных исходов Ω состоит из 2^n исходов, каждый из которых отождествляется с определенной последовательностью УНУ... (σ-алгебра событий включает 2^{2^n} событий). В силу независимости испытаний сопоставим каждому элементарному исходу ω = УННУ...У вероятность Р(ω) = Р(УННУ...У) = $pqqp...p$, p – повторяется столько раз, сколько раз произошел успех, а q – сколько раз была неудача. Типичный представитель схемы Бернулли – n-кратное подбрасывание несимметричной монеты.

Вычислим вероятность $P_n(m)$ получить в n испытаниях ровно m успехов. Событие A_m – в n испытаниях произошло ровно m успехов – состоит из тех элементарных исходов, в которых буква "У" появляется ровно m раз. Число таких исходов совпадает с числом *сочетаний* (не важен порядок выпадения "У"). С другой

стороны, каждый элементарный исход, в котором буква “У” встречается ровно *m* раз имеет вероятность $p^m q^{n-m}$. Окончательно получаем:

$$P_n(m) = C_n^m p^m q^{n-m} \quad (m = 0, 1, 2, ..., n).$$

Данное выражение носит также название биноминального закона, поскольку $P_n(m)$ можно получить как коэффициент при z^m бинома $(pz+q)^n$:

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

Дискретная случайная величина μ распределена по биномиальному закону (рис. 8), если она принимает значения 0, 1, 2, ..., *n* в соответствии с рядом распределения, представленным в табл. 6, где $0 < p, q < 1$ и $p + q = 1$.

Таблица 6

Распределение случайной величины μ

μ	0	1	. . .	k	. . .	n
P	q^n	npq^{n-1}	. . .	$C_n^k p^k q^{n-k}$	. . .	p^n

Последний член разложения p^n определяет вероятность наступления рассматриваемого события n раз в n независимых испытаниях; предпоследний член определяет вероятность наступления события рассматриваемого события n раз в n независимых испытаниях; предпоследний член определяет вероятность наступления события n-1 раз, а первый член определяет вероятность того, что событие не появится ни разу.

Основные характеристики распределения:

M(X)=np; D(X)=npq; $\sigma(X) = \sqrt{n \cdot p \cdot q}$.

Биноминальное распределение

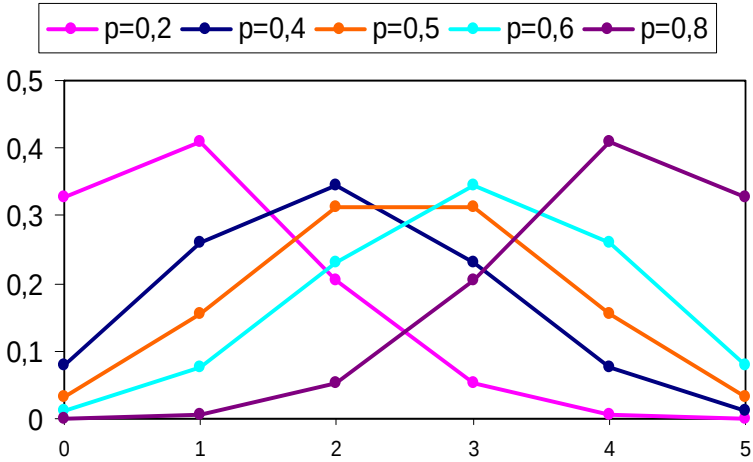


Рис. 8. Биноминальное распределение для n=5

Задача 1. Монета брошена 2 раза. Определить закон распределения случайной величины X – числа выпадений герба.

***Решение*:** Вероятность появления герба при каждом бросании монеты равна ½, следовательно, вероятность непоявления герба равна q=1-p=1-1/2=1/2.

При бросании монеты герб может появиться или 2 раза или 1 раз или совсем не появиться. Найдем вероятности этих событий по формуле Бернулли.

$$P_2(2) = C_2^2 p^2 = (\frac{1}{2})^2.$$

$$P_2(1) = C_2^1 pq = 2(\frac{1}{2})(\frac{1}{2}) = 0{,}5; \qquad P_2(0) = C_2^0 q^2 = (\frac{1}{2})^2 = 0{,}25.$$

X	2	1	0
P	*0,25*	*0,5*	*0,25*

Задача 2. На зачете студент получил n = 4 задачи. Вероятность решить правильно каждую задачу p = 0,8. Определим ряд распределения и построим функцию распределения случайной величины μ – числа правильно решенных задач.

Решение. В данном случае мы имеем дело с биноминальным законом: $P_n(k) = C_n^k p^k q^{n-k}$.

Таблица 7

Ряд биноминального распределения для задачи 2

μ	0	1	2	3	4
P	0.0016	0.0256	0.1536	0.4096	0.4096

Пуассоновское распределение

Распределение Пуассона моделирует случайную величину, представляющую собой число событий, произошедших за фиксированное время, при условии, что данные события происходят с фиксированной средней интенсивностью и независимо друг от друга (рис. 9).

Дискретная случайная величина ξ распределена по закону Пуассона и принимает целые неотрицательные значения с вероятностями, представленными рядом распределения (см. табл. 8). Параметр пуассоновского распределения λ>0 определяет интенсивность поступления событий и определяется формулой**:** λ=n*p, где n – общее число испытаний, а P – вероятность благоприятного исхода испытания.

Таблица 8

Пуассоновское распределение

ξ	0	1	2	…	k	…
P	$e^{-\lambda}$	$\lambda e^{-\lambda}$	$\frac{\lambda^2}{2!}e^{-\lambda}$	…	$\frac{\lambda^k}{k!}e^{-\lambda}$	…

Распределение Пуассона носит также название закона редких событий, поскольку оно всегда появляется там, где производится большое число испытаний, в каждом из которых с малой вероятностью происходит “редкое” событие. По закону Пуассона распределены, например, число вызовов, поступивших на телефонную станцию; число метеоритов, упавших в определенном районе; число распавшихся нестабильных частиц и т. д. При условии $p \to 0, n \to \infty, n \cdot p \to \lambda = const$ закон распределения Пуассона является предельным случаем биномиального закона.

Основные характеристики распределения:

$$M(X) = \lambda; D(X) = \lambda; \sigma(X) = \sqrt{\lambda}.$$

Распределение Пуассона

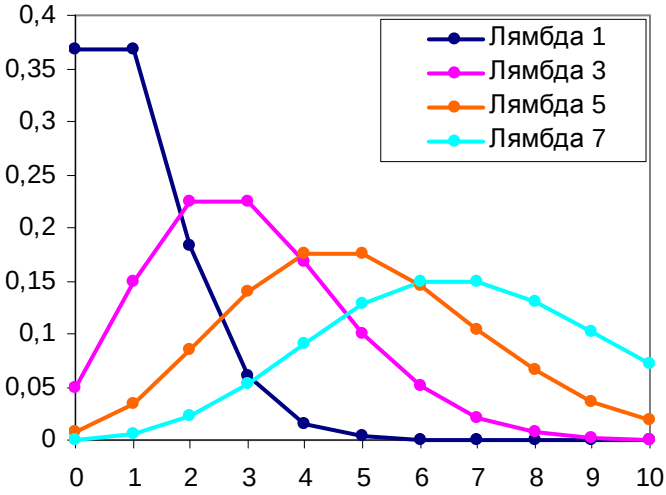


Рис. 9. Распределение Пуассона с различными λ

Формула Пуассона

Формула Пуассона применяется тогда, когда наряду с большим значением числа испытаний n "мала" вероятность успеха p. Она относится к приближенным формулам для вычисления $P_n(m)$ при больших n. Формула Пуассона наиболее простая из них.

Строго математически теорема Пуассона опирается на понятие *схемы серий*, здесь приведена "инженерная" интерпретация теоремы.

Теорема Пуассона: Пусть число испытаний n в схеме Бернулли "велико", а вероятность успеха p в одном испытании "мала", причем "мало" также произведение $\lambda=np$. Тогда $P_n(m)$ определяется по приближенной формуле (формула Пуассона)

$$P_n(m) \approx \frac{\lambda^m}{m!} e^{-\lambda} \quad (m=0, 1, 2, \ldots, n).$$

Доказательство. Формула Бернулли

$$P_n(m) = C_n^m \cdot p^m \cdot q^{n-m}$$

$$P_n(m) = \left(\frac{n}{m}\right) p^m (1-p)^{n-m} = \frac{n(n-1)\cdots(n-m+1)}{m!} p^m (1-p)^{n-m},$$

или с учетом обозначения $\lambda = np$,

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

При больших n $(1-\lambda/n)^n \approx e^{-\lambda}$. Кроме того, если n – велико, то $(n-1)/n \approx 1$, ..., $(n-m+1)/n \approx 1$ и $(1-\lambda/n)^{-m} \approx 1$. Поэтому приходим к доказываемой формуле.

Задача 3. Завод отправил на базу 5000 доброкачественных изделий. Вероятность того, что в пути изделие повредится равно 0,0002. Найти вероятность того, что на базу прибудут 3 негодных изделия.

Решение. По условию $n = 5000$, p=0,0002, k=3.

Найдем $\lambda = np = 5000 * 0.0002 = 1$

По формуле Пуассона искомая вероятность равна

$$P_{5000}(3) = \frac{\lambda^k \cdot e^{-\lambda}}{k!};$$

$$P_{5000}(3) = \frac{1^3}{e \cdot 3!} = 0{,}061.$$

Потоком событий называют последовательность событий, которые наступают в случайные моменты времени. Например, поступление вызовов на станцию скорой помощи, прибытие самолетов в аэропорт, клиентов в пункт сервиса, покупателей в магазин и т.д.

Простейшим (Пуассоновским) – называется поток событий, обладающий следующими свойствами:

- вероятность появления двух и более событий ничтожно мала по сравнению с вероятностью появления только одного события (*ординарность*);
- *отсутствие последствий*: вероятность появления m событий на любом промежутке времени не зависит от того, появлялись ли события раньше;
- *стационарность:* в одинаковые промежутки времени вероятность происхождения события одинакова.

Интенсивностью потока λ называют среднее число событий, которые появляются в единицу времени.

Задача 4. Среднее число вызовов, поступающих на АТС в одну минуту, равно – 2. Найти вероятности того, что за 5 минут поступит: а) 2 вызова; б); в) не менее 2 вызовов.

Решение: по условию $\lambda=2$, $t=5$, $m=4$.

По формуле Пуассона: $P_n(m) \approx \frac{\lambda t^m}{m!} e^{-\lambda t}$.

А) Вероятность, что за 5 минут поступят 2 вызова:

$$P_5(2) \approx \frac{10^2}{2!} e^{-10} = 0,00225 \quad .$$

Это событие практически невозможно.

Б) События «не поступило не одного вызова» и «поступил 1 вызов» – несовместны, поэтому по теореме сложения вероятностей: вероятность того, что за 5 минут поступят менее 2 вызовов, равна:

$$P_5(m<2) = \frac{10^2}{2!} e^{-10} = 0,00225.$$

Геометрическое распределение

Рассмотрим схему Бернулли. Пусть ξ – число испытаний, которое необходимо провести, прежде чем появится первый успех. Тогда ξ – дискретная случайная величина, принимающая значения 0, 1, 2, …, n, …. Определим вероятность события $\{\xi = n\}$. Очевидно, что $\xi = 0$, если в первом же испытании произойдет успех. Поэтому $P\{\xi = 0\} = p$. Далее, $\xi = 1$ в том случае, когда в первом испытании произошла неудача, а во втором – успех. Вероятность такого события – qp, то есть $P\{\xi = 1\} = qp$. Аналогично, $\xi = 2$, если в первых двух испытаниях произошли неудачи, а в третьем – успех: есть $P\{\xi = 2\} = qqp$. Продолжая эту процедуру, получим ряд распределения:

Таблица 9

Геометрическое распределение

ξ	0	1	2	…	k	…
P	p	qp	q^2p	…	$q^{k-1}p$	…

Случайная величина с таким рядом распределения называется распределенной по геометрическому закону (рис. 10).

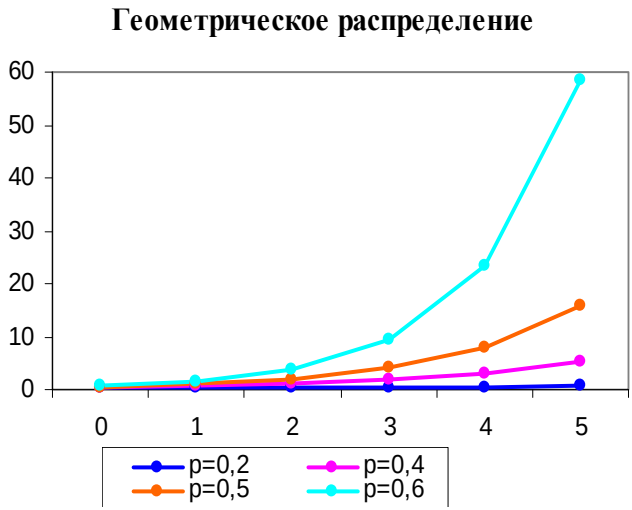


Рис. 10. Геометрическое распределение

Задача 5. Из орудия производится стрельба по цели до первого попадания. Вероятность попадания в цель p=0,6. Найти вероятность того, что попадание произойдет при третьем выстреле.

Решение. По условию, p=0,6, q=0,4, k=3. Искомая вероятность определяется по формуле:

$$p = q^{k-1} p = 0{,}4^2 * 0{,}6 = 0{,}096 .$$

Задача 6. Вероятность поражения цели равна 0,6. Производится стрельба по мишени до первого попадания (число патронов не ограничено). Требуется составить ряд распределения числа сделанных выстрелов, найти математическое ожидание и дисперсию этой случайной величины. Определить вероятность того, что для поражения цели потребуется не более трёх патронов.

Решение. Случайная величина X – число сделанных выстрелов – имеет геометрическое распределение с параметром p=0,6. Ряд распределения X имеет вид:

x_i	1	2	3	...	k	...
p_i	0,6	0,24	0,096	...	$0,6 \cdot 0,4^k$	...

Вероятность того, что для поражения цели потребуется не более трёх патронов равна

$P(X \le 3)=P(X=1)+P(X=2)+P(X=3)=0,6+0,24+0,096=0,936$.

Гипергеометрическое распределение

Определение. Дискретная случайная величина X имеет *гипергеометрическое распределение*, если она принимает значения 0, 1, 2, ..., $min\ \{n, M\}$ с вероятностями

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

где m=1, 2, ..., $min\ \{n, M\}$, $m \le N$, $n \le N$; n, N, M – натуральные числа.

N – общее количество объектов в генеральной совокупности;

M – количество объектов с определенным свойством в генеральной совокупности;

n – объем выборки;

m – количество деталей с определенным свойством.

Гипергеометрическое распределение имеет случайная величина $X=m$ – число объектов, обладающих данным свойством, среди n объектов, случайно извлечённых (без возврата) из совокупности N объектов, M из которых обладают этим свойством.

Гипергеометрическое распределение широко используется в практике статистического приёмочного контроля качества промышленной продукции, в задачах, связанных с организацией выборочных обследований, и некоторых других областях.

Задача 7. В национальной лотерее "6 из 45" денежные призы получают участники, угадавшие от трёх до шести чисел из случайно отобранных 6 из 45 (размер выигрыша увеличивается с увеличением числа угаданных чисел). Найти закон распределения, математическое ожидание и дисперсию случайной величины X – числа угаданных чисел среди случайно отобранных шести. Какова вероятность получения денежного приза?

Решение. Случайная величина X – число угаданных чисел среди случайно отобранных шести – имеет гипергеометрическое распределение с параметрами $n=6$, $N=45$, $M=6$. Ряд распределения X, рассчитанный по формуле:

$$P(X=m)=\frac{C_M^m \cdot C_{N-M}^{n-m}}{C_N^n}, где\_m=0,1,2...6.$$

Таблица 10

Гипергеометрическое распределение

x_i	0	1	2	3	4	5	6
p_i	0,40056	0,42413	0,15147	0,02244	0,00137	0,00003	0,0000001

Вероятность получения денежного приза

$$P(3\le X\le 6)=P(X=3)+P(X=4)+P(X=5)+P(X=6)$$

$$P(3\le X\le 6)=0,02244+0,00137+0,00003+0,0000001\approx 0,024.$$

Контрольные вопросы

1. Дать определение дискретной случайной величины.

2. Перечислить примеры распределений дискретной случайной величины.
3. Привести ряд распределения случайной величины, подчиняющейся биноминальному закону.
4. Привести ряд распределения случайной величины, подчиняющейся закону Пуассона.
5. Привести ряд распределения случайной величины, подчиняющейся геометрическому закону.
6. Привести ряд распределения случайной величины, подчиняющейся гипергеометрическому закону.

Распределения непрерывных случайных величин

Цель: *изучить свойства непрерывных случайных величин и явления, в которых они наблюдаются. Исследовать графики плотности распределения непрерывных случайных величин.*

Случайную величину назовем непрерывной, если ее функция распределения не имеет скачков и разрывов.

Непрерывной называется случайная величина ξ, *функцию распределения которой* $F(x)$ *можно представить в виде:*

$$F(x) = \int_{-\infty}^{x} p(x)dx.$$

Функция $p(x)$ *называется* **плотностью распределения** *(вероятностей) случайной величины* ξ (рис. 11).

Практически все, реально встречающиеся плотности распределения являются непрерывными функциями, и, следовательно, для них $p(x) = F'(x)$, то есть производную от функции распределения.

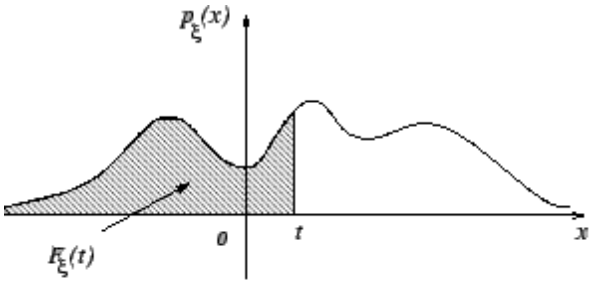


Рис. 11. Плотность распределения (вероятностей) случайной величины ξ

Функция $P\xi(x)$, обладающая вышеперечисленными свойствами, называется ***плотностью распределения случайной величины*** ξ**.**

Определим простейшие свойства плотности распределения *p(x) непрерывной случайной* величины:

1. Плотность распределения непрерывной случайной величины неотрицательная функция, поскольку она является производной от функции распределения, а функция распределения – неубывающая **p(x)>=0** для всех *x*.

2. Площадь, целиком заключенная под всей кривой плотности распределения, равна единице.

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

3. Вероятность попадания случайной величины на интервал [a, b] численно равна площади криволинейной трапеции (рис. 12): $P(a \le x < b) = \int_a^b p(x)dx = F(b) - F(a);$

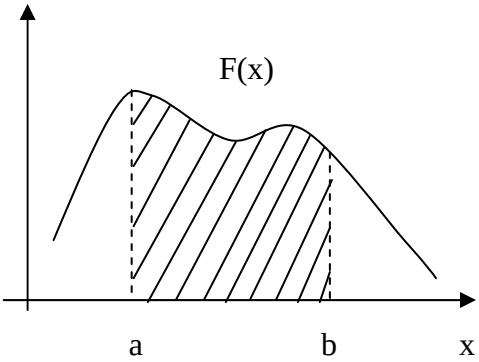


Рис. 12. Иллюстрация к свойству 3

$F(x) = \int_{-\infty}^{x} p(u)du$. Следовательно, $p(x) = \frac{dF(x)}{dx}$.

Равномерное распределение

Равномерно распределенная на отрезке [a,b] случайная величина имеет функцию распределения (рис. 13):

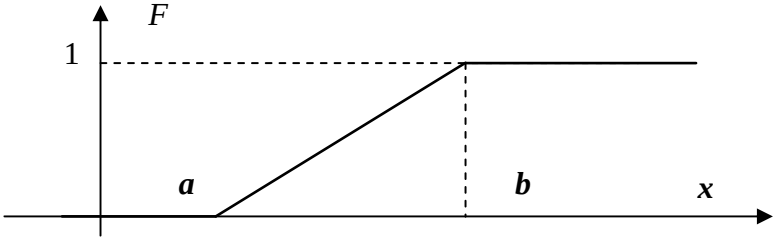


Рис. 13. Функция равномерного распределения

Функция равномерного распределения имеет вид:

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

Плотность равномерного распределения представлена ниже (рис.14).

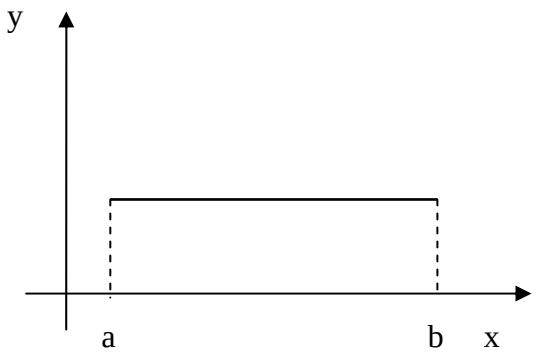


Рис. 14. Плотность равномерного распределения

$$p(x)=\begin{cases}\dfrac{1}{b-a}, \text{ если } a\le x\le b,\\ 0, \textbf{\textit{если }} x<a \textbf{\textit{ или }} x>b.\end{cases}$$

Основные характеристики распределения:

$$M(X)=\frac{a+b}{2};D(X)=\frac{(b-a)^2}{12}\lambda;\sigma(X)=\frac{b-a}{2\sqrt{3}}.$$

Вероятность попадания равномерно распределенной случайной величины на интервал (x_1, x_2), лежащий внутри отрезка (a,b), равна $F(x_2)-F(x_1)=(x_2-x_1)/(b-a)$, то есть пропорциональна длине этого интервала. Равномерное распределение реализует принцип геометрической вероятности при бросании точки на отрезок (a,b).

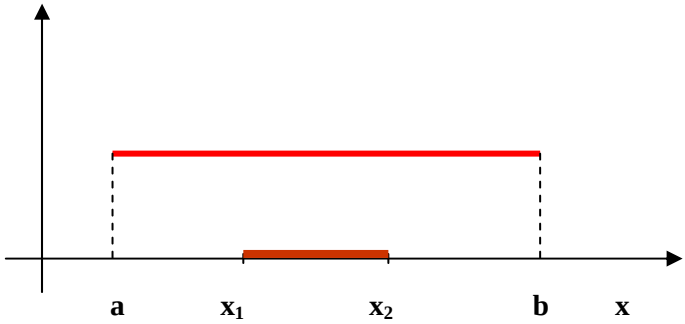


Рис. 15. Принцип геометрической вероятности при бросании точки на отрезок (a,b)

Экспоненциальное распределение

Случайная величина подчиняется экспоненциальному (показательному) закону, если она имеет функцию распределения:

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

Плотность экспоненциального распределения можно получить интегрированием функции распределения:

$$p(x)=\begin{cases} 0, \text{если} x<0, \\ \lambda e^{-\lambda x}, \text{если} x\geq 0. \end{cases}$$

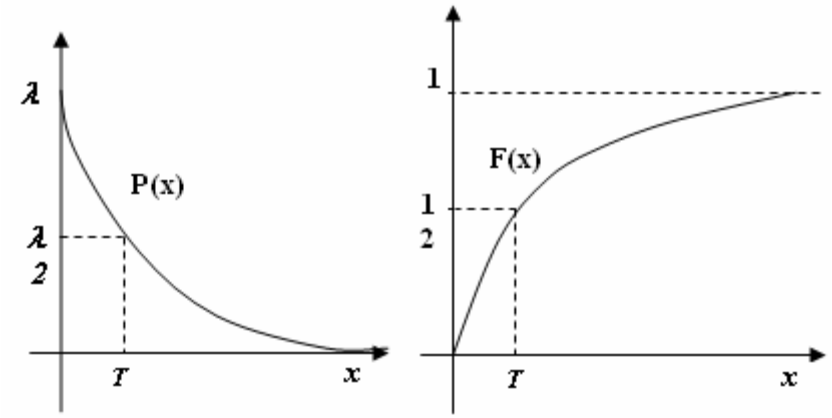


Рис. 16. Плотность и функция экспоненциального распределения

Экспоненциально распределенная случайная величина может принимать только положительные значения. Экспоненциальному распределению подчинено время распада атомов различных элементов. При этом число $T = 1/\lambda$ носит название *среднего времени распада*. Кроме того, употребляется также число $T_0 = ln2/\lambda$ – называемое периодом полураспада.

Экспоненциально распределенная случайная величина ξ обладает свойством – *отсутствием последействия*. Это можно трактовать как независимость поведения случайной величины в момент времени $x+\Delta x$ от того, что с ней произошло до этого.

Основные характеристики распределения:

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

Нормальное распределение

Случайная величина распределена по нормальному или гауссову закону, если она имеет плотность распределения

$$\varphi_{m,\sigma}(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-m)^2}{2\sigma^2}} \qquad (-\infty<m<\infty,\sigma>0).$$

Нормальное распределение зависит от двух параметров: где **m** – математическое ожидание или среднее значение нормального закона; **σ**- среднее квадратичное отклонение (рис. 17).

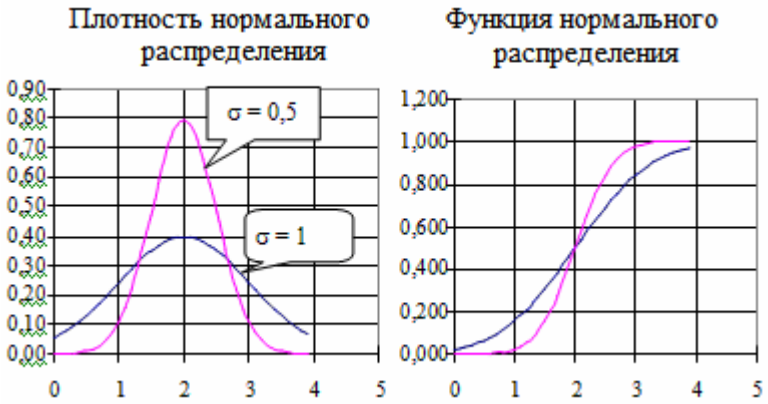


Рис. 17. Графики плотности и функции нормального распределения

Графики для плотности распределения с одинаковым средним арифметическим и различными среднеквадратическими отклонениями. Плотность нормального распределения зависит от значений среднего арифметического и среднеквадратичного отклонения (рис. 18).

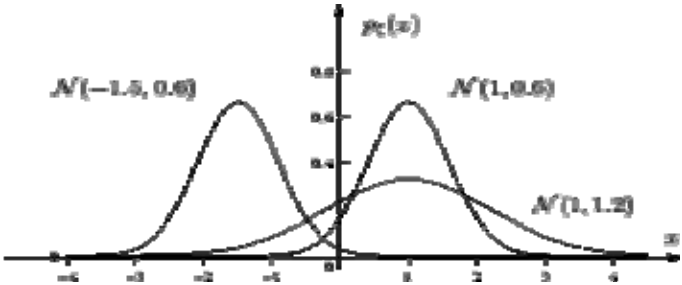


Рис. 18. Зависимость плотности нормального распределения от m и σ

Параметр **m** определяет положение центра нормальной плотности, а **σ** – разброс относительно центра. Если m=0, σ = 1, то такой нормальный закон называется *стандартным* и его функция распределения обозначается через Φ(x).

Основные характеристики распределения:

$$M(X) = m; D(X) = \sigma^2; \sigma(X) = \sigma.$$

Нормальное распределение возникает обычно в явлениях, подверженных действию большого числа “малых” случайных воздействий.

Распределение Вейбулла

Случайная величина распределена по закону Вейбулла, если она имеет плотность распределения

$$p(x) = \begin{cases} 0, \text{если}..\ x < 0, \\ \alpha \cdot \beta \cdot x^{\beta-1} e^{-\alpha \cdot x^{\beta}}, \ \text{если}..x \geq 0 \ (\alpha > 0, \beta > 0). \end{cases}$$

Семейство распределений Вейбулла является двухпараметрическим и описывает положительные случайные величины (рис.19). Считается, что распределению Вейбулла подчиняются времена безотказной работы многих технических устройств. Ес-

ли β=1, то распределение Вейбулла превращается в экспоненциальное распределение, а если β=2 – в так называемое распределение Релея.

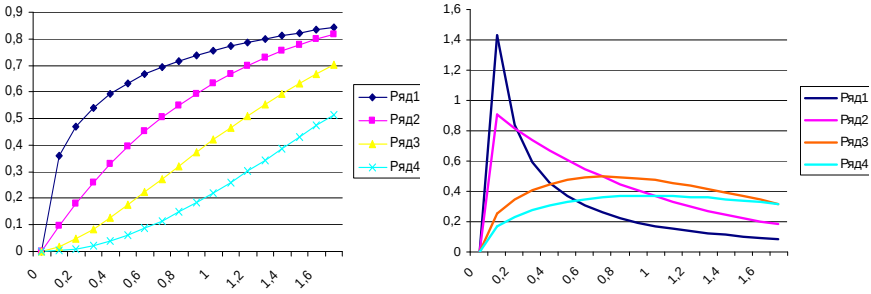


Рис. 19. Семейство функций и плотностей распределения Вейбулла

Гамма-распределение

Другим распределением, также достаточно хорошо описывающим времена безотказной работы различных технических устройств, является гамма-распределение с плотностью:

$$p(x)=\begin{cases}0, \text{ если } x\le 0,\\ \dfrac{\lambda^{\gamma}x^{\gamma-1}}{\Gamma(\gamma)}e^{-\lambda x}, \text{если } x\ge 0 \ \ (\lambda>0,\gamma>0).\end{cases}$$

где $\Gamma(\gamma)=\int_0^{\infty}x^{\gamma-1}e^{-x}dx$ – гамма-функция Эйлера. Свойства гамма-функции: $\Gamma(\gamma+1)=\gamma\Gamma(\gamma)$ и $\Gamma(n)=(n-1)!$ Для целых n. Функция и плотность Гамма-распределения представлены на рис. 20.

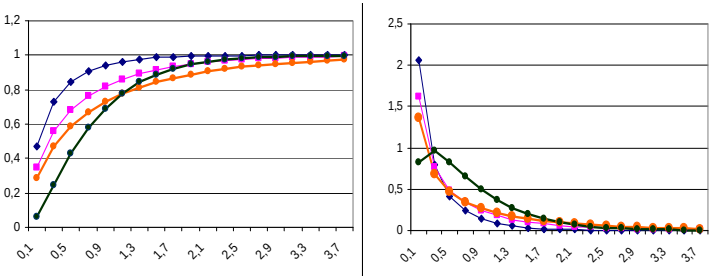


Рис. 20. Семейство функций и плотностей Гамма-распределения

Если $\gamma = k/2$ – полуцелое, а $\lambda = 1/2$, то гамма-распределение превращается в так называемое распределение χ^2(хи-квадрат). Параметр k – называется в этом случае числом степеней свободы распределения χ^2.

Функции от случайной величины

Для дальнейшего развития понятия – *случайная величина*, следует перейти к понятию – *функция от случайной величины*.

Пусть на вероятностном пространстве (Ω, σ, P) задана случайная величина $\xi = \xi(\omega)$. Возьмем обычную (измеримую) числовую функцию $g(x)$ числового аргумента x. Сопоставляя каждому элементарному исходу ω число $\eta(\omega)$ по формуле $\eta(\omega)=g(\xi(\omega))$, получим новую случайную величину η, которую назовем *функцией* $g(\xi)$ *от случайной величины* ξ.

Функция $\eta = g(\xi)$ от дискретной случайной величины также является дискретной случайной величиной, поскольку она не может принимать больше значений, чем случайная величина ξ. Ряд распределения случайной величины $\eta = g(\xi)$ можно представить таблицей:

η	$g(X_1)$	$g(X_2)$	…	$g(X_n)$
P	p_1	p_2	…	p_n

При этом, если в верхней строке таблицы появляются одинаковые значения $g(X_i)$, то соответствующие столбцы надо объединить в один, приписав им суммарную вероятность.

Функция $\eta = g(\xi)$ от непрерывной случайной величины ξ может быть как непрерывной, так и дискретной (дискретной она будет, например, если множество значений функции $g(x)$ не более, чем счётно). Найдем *функцию распределения* $F_\eta(x)$ по заданной плотности $p_\xi(x)$. По определению $F_\eta(x)$ представляет собой вероятность события $\{\eta < x\}$, состоящего из тех элементарных исходов ω, для которых $g(\xi(\omega))<x$. В свою очередь, вероятность события $\{g(\xi(\omega))<x\}$ можно определить, используя аксиому сложения вероятностей, "просуммировав" вероятности всех возможных значений y (промежуточная переменная) случайной величины ξ, для которых $g(y)<x$. Так как вероятность случайной величине ξ принять значение в промежутке от y до $y+dy$ приближенно равна $p_\xi(y)dy$, то, заменяя сумму на интеграл, получаем

$$F_\eta(x) = \int\limits_{g(y)<x} p_\xi(y)dy.$$

Рассматривая любую функцию от случайной величины (как, впрочем, и любую другую функцию), необходимо четко представлять *область определения* этой функции. В случае функции распределения область определения строится на тех значениях x (то есть случайных величинах, полученных на множестве событий ω_i), для которых $g(y) < x$.

Контрольные вопросы

1. Сформулировать равномерный закон распределения. Привести формулы его функции и плотности.
2. Описать формулу для вычисления математического ожидания равномерно распределенной случайной величины.
3. Сформулировать нормальный закон распределения. Записать дифференциальную и интегральную функции.
4. Описать свойства плотности нормально распределенной случайной величины. Пояснить геометрический смысл параметров нормального распределения.
5. Сформулировать экспоненциальный закон распределения. Привести его плотность и функцию распределения.

Числовые характеристики случайных величин

Цель: *изучение количественных характеристик моментов высших порядков.*

Основной целью статистического анализа является выяснение некоторых свойств изучаемой генеральной совокупности. Если генеральная совокупность конечна, то наилучшая процедура – рассмотрение каждого ее элемента. Однако в большинстве интересных задач используются либо бесконечные генеральные совокупности, либо конечные, но трудно обозримые. В этой ситуации необходимо отобрать из генеральной совокупности подмножество из n элементов, называемое *выборкой объема* n, исследовать его свойства, а затем обобщить эти результаты на всю генеральную совокупность. Это обобщение называется *статистическим выводом*.

Генеральная совокупность (популяция) W – полный набор объектов w, с которыми связана данная проблема. Эти объекты

могут быть людьми, животными, изделиями и так далее. С каждым объектом связана величина (или величины), называемая исследуемым признаком (x_i).

Различные значения признака, наблюдающиеся у членов генеральной совокупности (или выборки), называются вариантами, а числа, показывающие сколько раз встречается каждый вариант – их частотами.

В данном определении предполагается дискретное изменения признака. Однако, если мы измеряем непрерывную величину, то точность измерения и количество измерений в единицу времени тоже дадут некий дискретный набор.

Мы предполагаем, что измеряемый или исследуемый признак изменяется некоторым случайным образом. Произведя серию измерений, получим набор данных, которые, скорее всего, будут случайной выборкой из генеральной совокупности. Чтобы провести *первичную* обработку этой выборки, необходимо построить *экспериментальное распределение данных по частотам* или (если данные имеют явно непрерывный характер) по *интервалам частот*.

Пример 1. При регистрации размеров продаваемой магазином женской верхней одежды были получены данные о 100 покупках (табл. 11).

Таблица 11

Размеры одежды, купленной в магазине

42	48	50	46	50	48	48	46	50	50
50	50	48	48	44	48	50	46	50	52
46	50	46	46	50	50	42	48	48	46
52	48	54	48	46	50	48	54	46	50
50	50	50	44	50	48	46	48	46	52

54	50	46	48	52	48	46	46	46	44
48	46	54	48	46	50	44	48	52	50
46	46	48	46	50	48	50	48	54	46
48	48	46	46	46	52	54	46	46	46
48	44	44	48	52	54	48	48	52	50

Построим экспериментальное распределение данных по частотам. Для этого нужно определить количество признаков или интервалов частот, а затем подсчитать сколько вариантов в выборке соответствуют каждому интервалу, то есть частоту. Результаты этих расчетов для удобства заносят в таблицу, аналогичную табл. 12.

Таблица 12

Построение признаков и частот по выборке

Признаки	Частоты
42	2
44	6
46	27
48	27
50	23
52	8
54	7

По выборочным частотам строят гистограмму частот, которая характеризует опытное распределение (рис. 21).

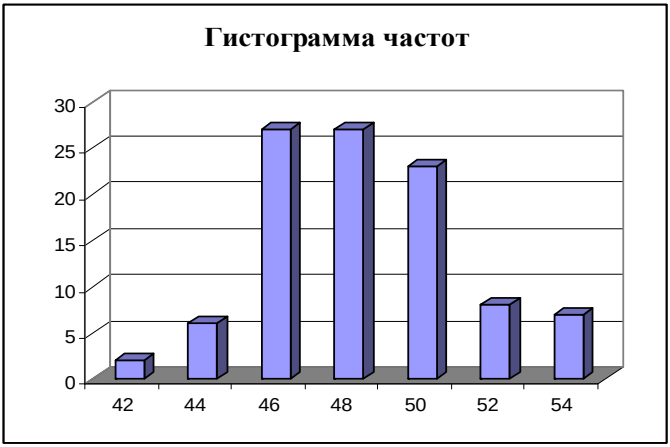


Рис. 21. Гистограмма частот

Другое представление получается, если значения на оси Y, соответствующие значениям на оси X, соединить ломаной кривой. Эта фигура называется *полигоном частот* или *многоугольником распределения*. Полигон частот дает информацию о законе распределения генеральной совокупности.

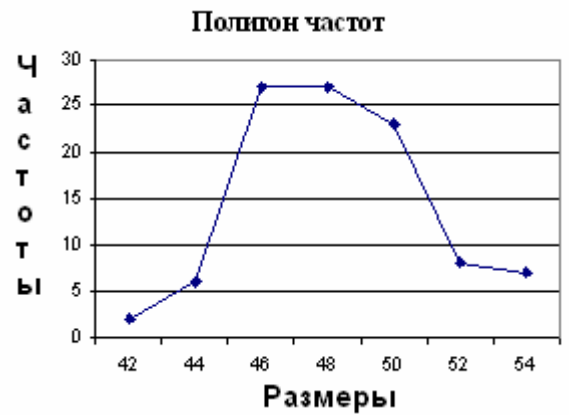


Рис. 22. Полигон частот выборки

Графическая иллюстрация статистических данных, геометрическая интерпретация отдельных вопросов статистики дает им наглядность, а в ряде случаев позволяет подвергнуть их анализу в наиболее простой и доступной форме. *В примере использованы графики, полученные в результате обработки данных в Excel.*

Под *формой* статистического распределения понимается форма его графика – полигона частот (рис. 22). Различают симметричные формы и несимметричные (асимметричные).

Распределение называется симметричным, если веса любых вариантов, равноотстоящих от среднего, равны между собой.

На практике такого совпадения для всех вариантов обычно нет и симметричными считаются распределения, в которых веса вариантов, равноотстоящих от среднего, отличаются незначительно. Полигон частот из пример похож на симметричное распределение, но совпадение приблизительное, поэтому данное распределение является умеренно асимметричным.

Асимметричные распределения можно разбить на три вида:

- умеренно асимметричные – распределения у которых частоты, находящиеся по одну сторону от наибольшей, больше (или меньше) частот, находящихся по другую сторону от наибольшей на таком же “расстоянии”;
- крайне асимметричные – распределения, у которых частоты или все время возрастают, или все время убывают.
- U-образные – частоты сначала убывают, а затем возрастают.

Числовые характеристики статистического распределения

В качестве характеристик измеримого признака вместо исходных значений величин или таблиц их частот используют чи-

словые характеристики, называемые также *статистическими мерами*.

- ***Среднее арифметическое*** $\bar{x}$ определяется по формуле

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i ,$$

где x_i – значения вариант.

- ***Медиана*** $\tilde{x}$ – срединное значение для ряда измерений n. Для ее вычисления необходимо все наблюдения расположить в порядке возрастания или убывания результатов. Если n – нечетное число, то медиана просто является числом, находящимся в середине упорядоченной последовательности. При четном n $\tilde{x}$ равна среднему арифметическому двух расположенных в середине значений упорядоченной последовательности.
- ***Мода*** – (наиболее вероятное значение) является наиболее часто встречающейся в выборке величиной.
- ***Размах вариации*** R – разность между максимальным и минимальным значениями признака в ряде измерений.

$$R = x_{max} - x_{min}.$$

- ***Среднее линейное отклонение*** d – среднее арифметическое абсолютных величин отклонений вариантов от их средней арифметической.

$$d = \frac{\sum_{i=1}^{n} |x_i - \bar{x}| \cdot n_i}{n} ,$$ n_i – частота признака x_i.

- ***Дисперсия D*** – среднее арифметическое квадратов отклонений вариантов от их средней:

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

- ***Среднее квадратичное отклонение*** σ **–** квадратный корень из дисперсии.

Каждая случайная величина характеризуется своей функцией распределения. С точки зрения наблюдателя, две случайные величины, имеющие одинаковые функции распределения, неразличимы, несмотря на то, что они могут быть заданы на различных вероятностных пространствах и описывать разные явления.

Функция распределения или плотность распределения вероятностей являются наиболее полными характеристиками случайных величин. Однако во многих задачах оказывается трудно или даже невозможно полностью описать функцию распределения. В то же время, для решения многих задач достаточно знать лишь некоторые параметры, характеризующие случайную величину с той или иной точки зрения. Наиболее распространенными числовыми параметрами, получившими название *числовых характеристик* или *моментов случайных величин*, являются *математическое ожидание* и *дисперсия* или среднеквадратичное отклонение.

Математическое ожидание случайной величины

Математическим ожиданием (средним значением) $M\xi$ дискретной случайной величины ξ называется сумма произведений значений X_i случайной величины на вероятности $p_i = P\{\xi=X_i\}$, с которыми эти значения принимаются:

$$M\xi = \sum_i X_i p_i .$$

Если рассматривать экспериментальные данные, аналогом математического ожидания является *среднее арифметическое значение* набора данных. Среднее арифметическое значение

приближается к математическому ожиданию при увеличении числа испытаний.

Пример 1. Найдем математическое ожидание случайной величины μ, распределенной по биномиальному закону (число успехов в n испытаниях Бернулли с вероятностью успеха p):

$$M\mu = \sum_{i=0}^{n} iP_n(i) = \sum_{i=0}^{n} i\binom{n}{i} p^i q^{n-i} = \sum_{i=0}^{n} i\frac{n!}{i!(n-i)!} p^i q^{n-i} =$$

$$= \sum_{i=1}^{n} np\frac{(n-1)!}{(i-1)!(n-i)!} p^{i-1} q^{n-1} = np\sum_{j=0}^{n-1}\binom{n-1}{j} p^j q^{n-1-j} =$$

$$M_\mu = np\sum_{j=0}^{n-1} P_{n-1}(j) = np.$$

Следовательно, $\mu = np$. При выводе использовалось свойство вероятности – сумма вероятностей всех событий равна 1.

Пример 2. Пусть ξ имеет распределение Пуассона. Тогда математическое ожидание этой величины равно:

$$M\xi = \sum_{i=0}^{\infty} i\frac{\lambda^i}{i!} e^{-\lambda} = \lambda\sum_{i=1}^{\infty}\frac{\lambda^{i-1}}{(i-1)!} e^{-\lambda} = \lambda\sum_{j=0}^{\infty}\frac{\lambda^j}{j!} e^{-\lambda} = \lambda e^{\lambda} e^{-\lambda} = \lambda.$$

При больших n $(1-\lambda/n)^n \approx e^{-\lambda}$. Таким образом, параметр λ пуассоновского распределения совпадает с математическим ожиданием.

Математическим ожиданием (средним значением) $M\xi$ непрерывной случайной величины ξ называется инте-

грал: $M\xi = \int_{-\infty}^{\infty} xp(x)dx$, где p(x) – плотность распределения. Случайная величина ξ принимает значение x с вероятностью p(x)dx.

Пример. Пусть непрерывная случайная величина X задана плотностью распределения φ(x)=sinx в интервале (0;π/2), вне этого интервала φ(x)=0. Найти математическое ожидание случайной величины ξ= φ(X)=X^2.

Воспользуемся формулой:

$$M[\varphi(X)] = \int_0^{\pi/2} x^2 \sin(x)dx = \pi - 2.$$

Свойства математического ожидания

1. Если случайная величина ξ принимает всего одно значение *C* с вероятностью единица. Математическое ожидание постоянной величины равно этой постоянной:

$$\mathbf{MC = C \cdot 1 = C}$$

2. Пусть η = aξ + b – случайная величина, выраженная линейной функцией, тогда математическое ожидание этой случайной величины равно:

$$\mathbf{M(a\xi+b)=aM\xi+b}$$

Рассматривая пример доказательства этого свойства для непрерывного случая:

$$M\eta = M(a\xi + b) = \int_{-\infty}^{\infty} (ax+b)p_{\xi}(x)dx = a\int_{-\infty}^{\infty} xp_{\xi}(x)dx + b\int_{-\infty}^{\infty} p_{\xi}(x)dx$$

.

Аналогично свойство 2 доказывается для дискретной случайной величины (постоянные величины выносятся за знаки суммирования).

3. Пусть η – случайная величина, которая является суммой двух других величин: $\eta = \xi_1 + \xi_2$. Тогда математическое ожидание суммы двух случайных величин равно сумме математических ожиданий каждой из этих величин:

$$\boxed{M(\xi_1 + \xi_2) = M\xi_1 + M\xi_2.}$$

Рассмотрим доказательство этого свойства на примере дискретной случайной величины:

$$\boldsymbol{M\eta = M(\xi_1 + \xi_2) = \sum_{i,j} (X_i + Y_j)\, p_{ij} = \sum_{i,j} X_i p_{ij} + \sum_{i,j} Y_i p_{ij} =}$$

$$\boldsymbol{= \sum_i X_i \sum_j p_{ij} + \sum_j Y_j \sum_i p_{ij} = \sum_i X_i p_{\xi_1 i} + \sum_j Y_j p_{\xi_2 j} = M\xi_1 + M\xi_2.}$$

4. Если ξ_1 и ξ_2 *независимы,* то математическое ожидание их произведений $\eta = \xi_1\xi_2$ равно произведению их математических ожиданий:

$$\boxed{\mathbf{M(\xi_1\xi_2) = M\xi_1 M\xi_2.}}$$

Дисперсия. Моменты высших порядков

Наряду со средним значением необходимо иметь число, характеризующее “разброс” случайной величины вокруг своего среднего. Такой характеристикой обычно служит *дисперсия*.

Следует помнить, что существует много характеристик разброса, в частности, центральные моменты любого четного порядка.

Существует много случайных величин, которые имеют одинаковые математические ожидания, но различные возможные значения. Например, дискретные случайные величины, заданные законами распределения:

X	-0,01	0,01	**Y**	-100	100
P	0,5	0,5	**P**	0,5	0,5

Найдем математическое ожидание этих величин:

$$M(X)= -0,01*0,5+0,01*0,5=0.$$

$$M(Y)= -100*0,5+100*0,5=0.$$

Здесь математические ожидания обеих величин одинаковы, причем X – близки к математическому ожиданию, а Y – далеки. Таким образом, зная математическое ожидание, нельзя судить о значениях случайной величины, ни о том, как рассеяны возможные значения случайной величины вокруг ее математического ожидания. Для оценки отклонения случайной величины от ее математического ожидания используют дисперсию.

Отклонением называют разность между случайной величиной и ее математическим ожиданием **X – M(X).**

На практике для оценки рассеяния случайной величины вокруг ее среднего значения пользуются величиной, названной дисперсией. Например, в артиллерии важно знать, насколько кучно лягут снаряды вокруг цели, которая должна быть поражена.

Дисперсия характеризует разброс значений случайной величины вокруг ее математического ожидания.

$$D\xi = M\left(\left[\xi - M(\xi)\right]^2\right) = M(\xi^2) - (M(\xi))^2.$$

Дисперсия Dξ *дискретной случайной величины* ξ *определяется формулой*

$$D\xi = \sum_{i=1}^{n} (x_i - M(\xi))^2 \cdot p_i = \sum_{i=1}^{n} x_i^2 \cdot p_i - \left(\sum_{i=1}^{n} x_i \cdot p_i\right)^2.$$

На первый взгляд может показаться, что достаточно вычислить все возможные значения отклонения и найти их среднее значение, но такой путь ничего не даст, поскольку отклонения бывают отрицательными и их среднее значение равно 0. Поэтому лучше взять абсолютные значения отклонений или их квадраты.

Пусть случайная величина задана законом распределения:

X	**X_1**	**X_2**	**X_3**	**xn**
p	**P_1**	**P_2**	**P_3**	**pn**

Тогда квадрат отклонения имеет следующий закон распределения:

$[X-M(X)]^2$	$[X_1-M(X)]^2$	$[X_2-M(X)]^2$	$[X_3-M(X)]^2$	$[Xn-M(X)]^2$
p	P_1	P_2	P_3	pn

По определению дисперсии,

$D(X)=M\,[X-M(X)]^2 = [x_1-M(X)]^2\ p_1 + [x_2-M(X)]^2\ p2 + [x_{2n}-M(X)]^2\ p_n$.

Из определения следует, что дисперсия случайной величины есть неслучайная (постоянная) величина.

Вторым (начальным) моментом m_2 *случайной величины* ξ *называется математическое ожидание квадрата* ξ $(g(x) = x^2)$:

$$\boldsymbol{m_2 = M\xi^2 = \sum_i x_i^2 p_i}.$$

Дисперсией непрерывной случайной величины называют математическое ожидание квадрата ее отклонения. Если возможные значения X принадлежат отрезку [a,b], то

$$D\xi = \int_{-\infty}^{\infty} (x - M\xi)^2 \, p(x) dx.$$

Среднее квадратичное отклонение непрерывной случайной величины определяется равенством:

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

Дисперсия $D\xi$ представляет собой второй момент случайной величины ξ, из которой вычтено ее математическое ожидание $M\xi$, то есть *центрированной* (имеющей нулевое математическое ожидание) случайной величины $\dot{\xi} = \xi - M\xi$.

Поэтому дисперсию иногда называют *вторым центральным моментом*.

Свойства дисперсии

Определим некоторые свойства дисперсии (без вывода).

1. Если случайная величина ξ с вероятностью 1 принимает одно и тоже постоянное значение C, то из свойства 1 математического ожидания (MC = C·1=C) получаем

$$D\xi = M(\xi - c)^2 = (c - c)^2 \cdot 1 = 0$$

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

$$D(a\xi) = a^2 D\xi.$$

3. Дисперсия суммы случайной величины и постоянной равна:

$$D(x+C)=Dx$$

4. Дисперсия суммы двух независимых случайных величин ξ и η равна сумме дисперсий каждой из этих величин:

$$D(\xi+\eta) = D\xi + D\eta.$$

Дисперсия $D\xi$ имеет размерность квадрата размерности случайной величины ξ. Для практических целей удобно иметь меру разброса, размерность которой совпадает с размерностью ξ. В качестве такой меры естественно использовать $\sigma = \sqrt{D\xi}$, которую называют *средним квадратичным отклонением* случайной величины ξ (или *стандартным* отклонением).

Моменты высших порядков

В теории вероятностей и математической статистике, помимо математического ожидания и дисперсии, используются и другие числовые характеристики случайных величин. В первую очередь это *начальные* и *центральные* моменты.

Начальным моментом k-го порядка случайной величины называется математическое ожидание *k*-й степени случайной величины x , то есть:

$$\alpha_k = Mx^k.$$

Заметим, что математическое ожидание случайной величины – начальный момент первого порядка, $\alpha_1 = Mx^1$.

Центральным моментом k-го порядка случайной величины x называется величина m_k, определяемая формулой $\mu_k = Mx^k$

$$\mu_2 = M(x - \bar{x})^2 = Dx.$$

Дисперсия является центральным моментом второго порядка.

Существуют формулы, позволяющие выразить центральные моменты случайной величины через ее начальные моменты, например:

$$\mu_2 = a_2 - a_1^2;$$
$$\mu_3 = a_3 - 3a_2a_1 + 2a_1^3,$$

где m $_3$ – центральный момент третьего порядка.

Если плотность распределения вероятностей непрерывной случайной величины симметрична относительно прямой $x = \boldsymbol{M}x$, то все ее центральные моменты нечетного порядка равны нулю.

Асимметрия

В теории вероятностей и в математической статистике в качестве меры асимметрии распределения является коэффициент асимметрии, который определяется формулой:

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

Эксцесс

Нормальное распределение наиболее часто используется в теории вероятностей и в математической статистике, поэтому график плотности вероятностей нормального распределения стал своего рода эталоном, с которым сравнивают другие распределения. Одним из параметров, определяющих отличие рас-

пределения случайной величины x от нормального распределения, является эксцесс.

Эксцесс случайной величины определяется равенством:

$$K_э = \frac{\mu_4}{\sigma^4} - 3$$

У нормального распределения, естественно, **Кэ** = 0.
Если **Кэ**>0, то это означает, что график плотности вероятностей p_x (x) сильнее "заострен", чем у нормального распределения, если же **Кэ**< 0, то "заостренность" графика p_x (x) меньше, чем у нормального распределения**.**

Контрольные вопросы

1. Что такое частота выборочных данных и как она определяется?
2. Что представляет собой полигон частот?
3. Какие формы выборочных распределений существуют?
4. Какой смысл имеет математическое ожидание для выборочного распределения?
5. Как рассчитывается дисперсия дискретной и непрерывной случайных величин?
6. Что такое мода и медиана и как они определяются?
7. Приведите примеры моментов высших порядков.
8. Чем отличаются начальные моменты от центральных?
9. Что характеризует асимметрия распределения?
10. Как рассчитать эксцесс и что он описывает?

Введение в математическую статистику

Цель: *Освоить понятие статистическая гипотеза. Познакомиться с методами статистической проверки гипотез.*

В задачу математической статистики входит изучение массовых явлений в природе, обществе и технике и их научное обоснование. Везде, где приходится иметь дело с обработкой экспериментальных результатов, необходимыми и незаменимыми вспомогательными средствами являются методы математической статистики.

Зарождение математической статистики было связано со сбором данных и графическим представлением полученных результатов измерений. Так возникли первые сводки рождаемости, бракосочетаний и смертности в демографической статистике.

В 20-е годы нашего столетия, главным образом в США и Англии, были разработаны математико-статистические методы научной обработки результатов измерений, основанные на закономерностях теории вероятностей (К. Пирсон, Р.А. Фишер, Дж. Нейман, А. Вальд).

Генеральная совокупность (популяция) **W** – полный набор объектов, с которыми связана данная проблема. Эти объекты могут быть людьми, животными, изделиями и так далее. С каждым объектом связана величина (или величины), называемая исследуемым признаком (x_i).

Основной целью статистического анализа является выяснение некоторых свойств рассматриваемой генеральной совокупности. Если генеральная совокупность конечна, то наилучшая процедура – рассмотрение каждого ее элемента. Однако в большинстве задач используются либо бесконечные генеральные совокупности, либо конечные, но трудно обозримые. В этой си-

туации необходимо отобрать из генеральной совокупности подмножество из n элементов, называемое выборкой объема n, исследовать его свойства, а затем обобщить эти результаты на всю генеральную совокупность. Это обобщение называется статистическим выводом.

Способы получения различных выборок и оценка их представительности будут рассмотрены в лабораторном практикуме. Различные значения признака, наблюдающиеся у членов генеральной совокупности (или выборки), называются вариантами, а числа, показывающие сколько раз встречается каждый вариант, частотами.

В данном определении предполагается дискретное изменения признака. Однако, если измерять непрерывную величину, то точность измерения и количество измерений в единицу времени тоже дадут некий дискретный набор.

Мы предполагаем, что измеряемый или исследуемый признак изменяется некоторым случайным образом. Произведя серию измерений, получим набор данных, которые, скорее всего, будут случайной выборкой из генеральной совокупности. Чтобы провести первичную обработку этой выборки, необходимо построить экспериментальное распределение данных по частотам или (если данные имеют явно непрерывный характер) по интервалам частот.

Выборочные функции

Для любой случайной величины X существует (теоретическая) функция распределения F(x), или по-другому “Генеральная совокупность имеет теоретическое распределение F(x)”.

Вероятностный закон генеральной совокупности на практике почти всегда неизвестен. Единственным источником информации о нем служит взятая из этой совокупности выборка объе-

ма ***n***, элементы которой $x_1, x_2, ..., x_n$ являются реализациями ***X***; по ней рассчитывается эмпирическое распределение и статистические параметры (еще говорят – статистики числовых характеристик): среднее арифметическое, дисперсия, моменты высших порядков и др.

Эмпирическое распределение выборки рассматривается в качестве оценки теоретической функции распределения F(x) генеральной совокупности.

Пусть дана выборка значений случайной величины $X = (x_1, x_2, ...x_n)$ из неизвестного совместного распределения F(x). Тогда любое утверждение, касающееся природы F(x), называется статистической гипотезой. Гипотезы различают по виду предположений, содержащихся в них:

Статистическая гипотеза, однозначно определяющая распределение F(x), то есть $H = \{F(x) = F_0\}$, где F_0 какой-то конкретный закон, называется простой.

Статистическая гипотеза, утверждающая принадлежность распределения F(x) к некоторому семейству распределений, то есть вида $H = \{F(x) \in F\}$, где F – семейство распределений, называется сложной.

Например, для экспоненциального распределения гипотеза H_0: $\lambda = 3$ – простая, H_0: $\lambda > 3$ – сложная, состоящая из бесконечного числа простых гипотез вида $\lambda = c$, где c – любое число, большее 3.

На практике обычно требуется проверить какую-то конкретную и как правило простую гипотезу . Такую гипотезу принято называть *нулевой*. При этом параллельно рассматривается противоречащая ей гипотеза , называемая *конкурирующей* или *альтернативной*.

Различают две группы математико-статистических методов:

- **статистическая проверка гипотез** (статистические тесты);
- **статистическая оценка параметров распределения**.

Статистическая проверка гипотез предполагает выдвижение определенных допущений (гипотез) относительно неизвестных параметров F(x). Правильность этих гипотез проверяется затем по числовым значениям, полученным из выборки, и, в зависимости от результата проверки, гипотезы принимаются или отвергаются.

Примеры непараметрических гипотез

$\mathbf{H_0}$: $F(x)=F_0(x)$; где $F_0(x)$ может быть функцией *нормального распределения* с определенными установленными параметрами μ_0 и σ_0^2, то есть $F_0(x) = \Phi(x; \mu_0, \sigma_0^2)$. Закон распределения выборочной совокупности является нормальным

$\mathbf{H_1}$: закон распределения выборочной совокупности не является нормальным.

$\mathbf{H_0}$: связь между ущербом в случае аварии и размером страхуемой машины отсутствует;

$\mathbf{H_1}$: связь между ущербом в случае аварии и размером машины существует.

Статистическая оценка параметров распределения предусматривает получение оценок (для отдельных значений или интервалов) неизвестных параметров вероятностного закона генеральной совокупности по параметрам выборки.

При статистической оценке параметров распределения и проверке гипотез используются числовые характеристики, рассчитанные по ***n*** наблюдениям выборки.

Пример параметрической гипотезы

Пусть дана независимая выборка из нормального распределения, где μ – неизвестный параметр. Тогда , где – фиксированная константа, является простой параметрической гипотезой, а конкурирующая с ней - сложная параметрическая гипотеза.

Выдвинутая гипотеза нуждается в проверке, которая осуществляется статистическими методами, поэтому гипотезу называют статистической. Для проверки гипотезы используют критерии, позволяющие принять или опровергнуть гипотезу. Статистической гипотезой называется любое предположение о виде неизвестного распределения или о параметрах известного распределения.

Статистическая проверка гипотез

Под **статистической гипотезой** понимают всякое высказывание о генеральной совокупности (случайной величине), проверяемое по выборке (по результатам наблюдений).

Располагая выборочными данными и руководствуясь конкретными условиями рассматриваемой задачи, формулируют гипотезу *Но,* которую называют основной или нулевой, и гипотезу H_1, конкурирующую с гипотезой H_0. Термин «конкурирующая» означает, что являются противоположными следующие два события:

- по выборке будет принято решение о справедливости для генеральной совокупности гипотезы H_0;
- по выборке будет принято решение о справедливости для генеральной совокупности гипотезы H_1.

Гипотезу H_1 называют также *альтернативной.*

Например, если нулевая гипотеза такова: математическое ожидание равно 5, то альтернативная гипотеза может быть сле-

дующей: математическое ожидание меньше 5, что записывается следующим образом:

Основная гипотеза: H_0: M(X)=5

Конкурирующая гипотеза: H_1: M(X)<5

Статистическая проверка гипотез применяется для того, чтобы использовать полученную по выборке информацию для суждения о законе распределения генеральной совокупности. При этом имеется определенное представление о неизвестном вероятностном законе *F(x)* и его параметрах, которое формулируется в виде статистической гипотезы, обозначаемой символом H или H_0 (нулевая, или основная, гипотеза).

Целесообразным оказался следующий способ записи: H_0: $F(x)=F_0(x)$; это означает допущение ("гипотезу") о том, что $F_0(x)$ есть функция распределения генеральной совокупности. Например, $F_0(x)$ может быть функцией нормального распределения с определенными установленными параметрами μ_0 и σ_0^2, то есть

$$F_0(x)=\Phi(x; \mu_0, \sigma_0^2).$$

С помощью статистических методов или критериев для проверки гипотезы устанавливается, соответствуют ли взятые из выборки данные выдвинутой гипотезе или нет, то есть нужно ли принять или отвергнуть гипотезу.

Если вид функции распределения F(x) задан отдельными параметрами и, если гипотеза строится именно по этим неизвестным параметрам, то говорят о параметрических гипотезах. Например, допущение о неизвестном параметре μ[2] нормального распределения является такой параметрической гипотезой. В

---

[2] μ – математическое ожидание случайной величины или "средняя арифметическая величина" по выборке.

противоположность этому статистические гипотезы общего порядка H_0: $F(x)=F_0(x)$ называются непараметрическими, а методы их проверки – непараметрическими тестами. Они, естественно, являются более общими, чем параметрические гипотезы и методы их проверки, так как не требуют дополнительных предположений о виде функции ***F(x)***. С другой стороны, они менее эффективны, чем соответствующие критерии параметрических гипотез.

Этапы проверки статистических гипотез

1. Формулировка основной гипотезы и конкурирующей гипотезы . Гипотезы должны быть чётко формализованы в математических терминах.

2. Задание вероятности , называемой уровнем значимости и отвечающей ошибкам первого рода, на котором в дальнейшем и будет сделан вывод о правдивости гипотезы.

3. Расчёт статистики критерия такой, что:

- её величина зависит от исходной выборки

$$X = (X_1, X_2, ...X_n) : \phi = \phi(X_1, X_2 ...X_n)$$

- по её значению можно делать выводы об истинности гипотезы ;
- сама статистика должна подчиняться какому-то известному закону распределения, т.к. сама является случайной в силу случайности .

4. Построение критической области. Из области значений выделяется подмножество таких значений, по которым можно судить о существенных расхождениях с предположением. Его размер выбирается таким образом, чтобы выполнялось равенство . Это множество и называется **критической областью**.

5. Вывод об истинности гипотезы. Наблюдаемые значения выборки подставляются в статистику и по попаданию (или не-

попаданию) в критическую область выносится решение об отвержении (или принятии) выдвинутой гипотезы $P(\phi \in C) = \alpha$.

Параметрические критерии

В группу параметрических критериев методов математической статистики входят методы для вычисления описательных статистик, построения графиков на нормальность распределения, проверка гипотез о принадлежности двух выборок одной совокупности. Эти методы основываются на предположении о том, что распределение выборок подчиняется нормальному (гауссовому) закону распределения. Среди параметрических критериев статистики нами будут рассмотрены критерий Стьюдента и Фишера.

Критерий Стьюдента (t-критерий)

Критерий позволяет найти вероятность того, что оба средних значения в выборке относятся к одной и той же совокупности. Данный критерий наиболее часто используется для проверки гипотезы: «Средние двух выборок относятся к одной и той же совокупности».

При использовании критерия можно выделить два случая. В первом случае его применяют для проверки гипотезы о равенстве генеральных средних двух независимых, несвязанных выборок (так называемый двухвыборочный t-критерий). В этом случае есть контрольная группа и экспериментальная (опытная) группа, количество испытуемых в группах может быть различно.

Во втором случае, когда одна и та же группа объектов порождает числовой материал для проверки гипотез о средних, ис-

пользуется так называемый парный t-критерий. Выборки при этом называют зависимыми, связанными.

Случай независимых выборок

Статистика критерия для случая несвязанных, независимых выборок равна:

$$t_{эмп} = \frac{\bar{x} - \bar{y}}{\delta_{x-y}}, \quad (1)$$

где $\bar{x}$, $\bar{y}$, – средние арифметические в экспериментальной и контрольной группах; σ_{x-y} – стандартная ошибка разности средних арифметических, которая находится из формулы:

$$\delta_{x-y} = \sqrt{\frac{\sum (x_i - \bar{x})^2 + \sum (y_i - \bar{y})^2}{n_1 + n_2 - 2} \cdot \left[\frac{1}{n_1} + \frac{1}{n_2}\right]}, \quad (2)$$

где n_1 и n_2 соответственно величины первой и второй выборки.

Если $n_1=n_2$, то стандартная ошибка разности средних арифметических будет считаться по формуле:

$$\delta_{x-y} = \sqrt{\frac{\sum (x_i - \bar{x})^2 + \sum (y_i - \bar{y})^2}{(n-1) \cdot n}}, \quad (3)$$

где n величина выборки.

Подсчет **числа степеней свободы** осуществляется по формуле:

$$k = n_1 + n_2 - 2. \quad (4)$$

При численном равенстве выборок $k = 2n - 2$.

Далее необходимо сравнить полученное значение $t_{эмп}$ с теоретическим значением t–распределения Стьюдента (см. приложение к учебникам статистики). Если $t_{эмп}<t_{крит}$, то гипотеза H_0 принимается, в противном случае нулевая гипотеза отвергается и принимается альтернативная гипотеза.

Рассмотрим пример использования t-критерия Стьюдента для несвязных и неравных по численности выборок.

Пример 1. В двух группах учащихся – экспериментальной и контрольной – получены следующие результаты по учебному предмету (см. табл. 13).

Таблица 13

Результаты эксперимента

Экспериментальная группа N1=11 человек	Контрольная группа N2=9 человек
12 14 13 16 11 9 13 15 15 18 14	13 9 11 10 7 6 8 10 11

Общее количество членов выборки: $n_1=11$, $n_2=9$.

Расчет средних арифметических: $X_{ср}=13{,}636$; $Y_{ср}=9{,}444$.

Стандартное отклонение: $\sigma_x=2{,}460$; $\sigma_y=2{,}186$.

По формуле (2) рассчитываем стандартную ошибку разности арифметических средних:

$$\delta_{xy}=\sqrt{\frac{60.545+38.222}{11+9-2}\cdot\left(\frac{1}{9}+\frac{1}{11}\right)}=1.053.$$

Считаем статистику критерия:

$$t=\frac{13.636-9.444}{1.053}=3.981.$$

Сравниваем полученное в эксперименте значение t с табличным значением с учетом степеней свободы, равных по формуле (4) числу испытуемых минус два.

Табличное значение $t_{крит}$ равняется 2,1 при допущении возможности риска сделать ошибочное суждение в пяти случаях из ста (уровень значимости α=5 % или 0,05).

Если полученное в эксперименте эмпирическое значение t превышает табличное, то есть основания принять альтернативную гипотезу (H_1) о том, что учащиеся экспериментальной группы показывают в среднем более высокий уровень знаний. В эксперименте t=3,981, табличное t=2,10, 3,981>2,10, откуда следует вывод о преимуществе экспериментального обучения.

Здесь могут возникнуть такие вопросы:

1. Что если полученное в опыте значение t окажется меньше табличного? Тогда надо принять нулевую гипотезу.

2. Доказано ли преимущество экспериментального метода? Не столько доказано, сколько показано, потому что с самого начала допускается риск ошибиться в пяти случаях из ста (p=0,05). Наш эксперимент мог быть одним из этих пяти случаев. Но 95% возможных случаев говорит в пользу альтернативной гипотезы, а это достаточно убедительный аргумент в статистическом доказательстве.

3. Что если в контрольной группе результаты окажутся выше, чем в экспериментальной? Поменяем, например, местами, сделав средней арифметической экспериментальной группы $\bar{x}$, а $\bar{y}$ – контрольной:

$$t = \frac{9.444 - 13.636}{1.053} = -3.981.$$

Отсюда следует вывод, что новый метод пока не проявил себя с хорошей стороны по разным, возможно, причинам. Поскольку абсолютное значение 3,9811>2,1, принимается вторая альтернативная гипотеза (H_2) о преимуществе традиционного метода.

Случай связанных выборок

В случае связанных выборок с равным числом измерений в каждой можно использовать более простую формулу t-критерия Стьюдента. Вычисление значения t осуществляется по формуле:

$$t_{эмп} = \frac{\overline{d}}{S_d}, \quad (5)$$

где $d_i = x_i - y_i$ – разности между соответствующими значениями переменной X и переменной У, а $\overline{d}$ - среднее этих разностей, а Sd вычисляется по следующей формуле;

$$S_d = \sqrt{\frac{\sum d_i^2 - \frac{(\sum d_i)^2}{n}}{n \cdot (n-1)}}. \quad (6)$$

Число степеней свободы *k* определяется по формуле k=n-1. Рассмотрим пример использования t-критерия Стьюдента для связных и, очевидно, равных по численности выборок.

Если $t_{эмп} < t_{крит}$, то нулевая гипотеза принимается, в противном случае принимается альтернативная.

Пример 2. Изучался уровень ориентации учащихся на художественно-эстетические ценности. С целью активизации формирования этой ориентации в экспериментальной группе проводились беседы, выставки детских рисунков, были организованы посещения музеев и картинных галерей, проведены встречи с

музыкантами, художниками и др. Закономерно встает вопрос: какова эффективность проведенной работы? С целью проверки эффективности этой работы до начала эксперимента и после давался тест. Из методических соображений в табл. 14 приводятся результаты небольшого числа испытуемых.

Таблица 14

Результаты эксперимента

Ученики	Баллы		Расчеты	
	начало экспер.	конец экспер.	d	d^2
Иванов	14	18	4	16
Новиков	20	19	-1	1
Сидоров	15	22	7	49
Пирогов	11	17	6	36
Агапов	16	24	8	64
Суворов	13	21	8	64
Рыжиков	16	25	9	81
Серов	19	26	7	49
Топоров	15	24	9	81
Быстров	9	15	6	36
Сумма	148	211	63	477
Среднее	14,8	21,1		

Вначале произведем расчет по формуле:

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{63}{10} = 6{,}3.$$

Затем применим формулу (6), получим:

$$Sd = \sqrt{\frac{\sum d_i^2 - \frac{(\sum d_i)^2}{n}}{n(n-1)}} = \sqrt{\frac{447 - (63 \cdot 63)/10}{10 \cdot 9}} = \sqrt{0{,}89} = 0{,}943.$$

И, наконец, следует применить формулу (5). Получим:

$$t_{эмп} = \frac{\bar{d}}{S_d} = \frac{6.3}{0.943} = 6.678.$$

Число степеней свободы: k=10-1=9 и по таблице находим $t_{крит}$ =2.262, экспериментальное t=6,678, откуда следует возможность принятия альтернативной гипотезы (H_1) о достоверных различиях средних арифметических, т.е. делается вывод об эффективности экспериментального воздействия.

В терминах статистических гипотез полученный результат будет звучать так: на 5%-м уровне гипотеза H_0 отклоняется и принимается гипотеза H_1 .

Критерий Фишера

F – критерий Фишера используют для сравнения дисперсий двух вариационных рядов. Он вычисляется по формуле:

$$F = \frac{S^2}{S'^2},$$

где $\mathbf{S^2}$ **–** большая выборочная дисперсия, $\mathbf{S'^2}$ – меньшая выборочная дисперсия. По двум выборкам объемами $\mathbf{n_1}$ и $\mathbf{n_2}$ строят выборочные функции:

$$\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n_1}(X_i - \overline{X})^2,$$

$$\overline{X}' = \frac{1}{n_2}\sum_{i=1}^{n_2} X_i', \quad S'^2 = \frac{1}{n-1}\sum_{i=1}^{n_1}(X_i' - \overline{X}')^2.$$

Если предположить, что соответствующие генеральные совокупности распределены нормально с параметрами μ_1, σ_1^2 и μ_2, σ_2^2 и что, кроме того, выполняется соотношение $\sigma_1^2 = \sigma_2^2$, то существует теорема: выборочная функция имеет непрерывную функцию распределения и ее плотность вероятности:

$$f_F(x) = E_{m_1,m_2} x^{(m_1/2-1)} (m_2 + m_1 x)^{-(m_1+m_2)/2}$$ для $x \geq 0$, где $m_1 = n_1 - 1$, $m_2 = n_2 - 1$. $E_{m1,m2}$ зависит лишь от m_1 и m_2.

Данное распределение названо в честь Р.А. Фишера F-распределением с (m_1, m_2) степенями свободы. Если вычисленное значение критерия F больше критического для определенного уровня значимости и соответствующих чисел степеней свободы для числителя и знаменателя, то дисперсии считаются различными.

Число степеней свободы числителя определяется по формуле: $k_1 = n_1 - 1$, где n_1 – число вариант для большей дисперсии.

Число степеней свободы знаменателя определяется по формуле: $k_2 = n_2 - 1$, где n_2 – число вариант для меньшей дисперсии.

Рассмотрим пример расчета критерия Фишера

Известны результаты женской эстафеты 5-го этапа Кубка мира 2010 г. по биатлону, прошедшей в Рупольдинге (в Германии).

H_0: D(X)=D(Y) – дисперсии времени прохождения дистанции спортсменками команд России и Швеции (двух генеральных совокупностей равны).

H_1: D(X)≠D(Y) – дисперсии времени прохождения дистанции спортсменками команд России и Швеции существенно различаются.

Таблица 15

Результаты эстафеты по биатлону

Россия	Время (X)	Швеция	Время(Y)
Романова Я.	19,35	Хогберг Э.	19,59
Булыгина А.	38,57	Олофссон А.	38,53
Медведцева О.	58,23	Нильссон А.	58,36
Зайцева О.	77,48	Йонссон Х.	77,31
Сумма	193,64	Сумма	193,80
Среднее	48,41	Среднее	48,45
Дисперсия	627,55	Дисперсия	620,78
$F_{экс}$	1,01	$F_{крит}$	9,27

По таблице критических точек распределения Фишера-Снедекора можно найти критическую точку для F-критерия при известных степенях свободы, равных: k= 4 – 1 = 3. Получаем $F_{крит}$=9,277. Поскольку, $F_{экс}<F_{кр}$: 1,01<9,277, следовательно, можно утверждать, что нулевая гипотеза H_0 принимается на 5%-м уровне значимости, а конкурирующая гипотеза H_1 в этом случае отвергается. Исследователь может сделать вывод, что по степени однородности показателя времени прохождения дистанции различие между двумя командами незначительные.

Непараметрические тесты

Чтобы определить, имеем ли мы дело с нормальным распределением, можно применять следующие методы.

1) В пределах осей можно нарисовать полигон частоты (эмпирическую функцию распределения) и кривую нормального распределения на основе данных исследования. Исследуя формы кривой нормального распределения и графика эмпирической функции распределения, можно выяснить те параметры, которыми последняя кривая отличается от первой.

2) Вычисляется среднее, медиана и мода и на основе этого определяется отклонение от нормального распределения. Если мода, медиана и среднее арифметическое друг от друга значительно не отличаются, мы имеем дело с нормальным распределением. Если медиана значительно отличается от среднего, то мы имеем дело с асимметричной выборкой.

3) Эксцесс кривой распределения должен быть равен 0. Кривые с положительным эксцессом значительно круче кривой нормального распределения. Кривые с отрицательным эксцессом являются более покатистыми по сравнению с кривой нормального распределения.

4) Правило трех сигм. После определения среднего значения распределения частоты и стандартного отклонения находят следующие четыре интервала распределения сравнивают их с действительными данными ряда:

а) $\bar{x} \pm 0.3\sigma$ – к интервалу должно относиться около 25% частоты совокупности,

б) $\bar{x} \pm 0.7\sigma$ – к интервалу должно относиться около 50% частоты совокупности,

в) $\bar{x} \pm 1.1\sigma$ – к интервалу должно относиться около 75% частоты совокупности,

г) $\bar{x} \pm 3\sigma$ – к интервалу должно относиться около 100% частоты совокупности.

Проверка гипотез о законе распределения по критерию χ2 (хи-квадрат)

Численным методом оценки того, принадлежит ли данная выборка генеральной совокупности с нормальным распределением, является критерий χ^2, разработанный К. Пирсоном. Согласно этому методу, наблюдаемое эмпирическое распределение выборки, выраженное абсолютными, относительными или относительными накопленными частотами сгруппированного ряда измерений, сравнивается с гипотетическим теоретическим распределением соответствующей генеральной совокупности. Для этого выдвигается гипотеза о неизвестной функции распределения F(x) генеральной совокупности, которая сопоставляется с подходящей выборочной функцией и, в зависимости от величины отклонения эмпирического распределения от теоретического, выдвинутая гипотеза принимается или отвергается. Так как статистическая гипотеза относится к неизвестной функции распределения F(x), а не к отдельным ее параметрам, мы говорим о непараметрическом методе проверки, или о критерии подобия. Критерий χ^2 и представляет собой один из таких критериев подобия.

Критерий χ^2 часто используют также для сравнения между собой двух выборок из некоторой генеральной совокупности.

Пусть в результате n наблюдений получен вариационный ряд с опытными частотами n_1, n_2, ..., n_m. Тогда их сумма равна . Анализируя опытные данные, выбираем некоторый закон теоретического распределения для рассматриваемого признака. По опытным данным найдем параметры этого закона (гипотеза). С помощью теоретического закона вычислим теоретические час-

тоты n_1^0, n_2^0, ..., n_m^0, соответствующие эмпирическим частотам. Сумма теоретических частот также должна быть равна объему выборки – n (соглашение).

В качестве меры расхождения теоретического и эмпирического рядов частот возьмем величину:

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - n_i^o)^2}{n_i^0}$$ – критерий согласия Пирсона.

Из выражения видно, что $\chi^2 = 0$ лишь при совпадении всех соответствующих эмпирических и теоретических частот: $n_i = n_i^0$. В противном случае $\chi 2$ отлично от нуля и тем больше, чем больше расхождение между указанными частотами.

Величина χ^2 является случайной и имеет распределение χ^2-распределение. Параметр k назван числом степеней свободы. Число k определяется:

$$k = m - s,$$

где m – число групп эмпирического распределения (или число интервалов);

s – число параметров теоретического закона, найденного с помощью этого распределения или число связей теоретического и эмпирического распределений. Например, если мы нашли – среднее арифметическое и σ^2 – дисперсию, используя данные опытного распределения и установили, сумма частот опытного распределения равна сумме частот теоретического распределения, то число связей s = 3.

Если же эмпирическое распределение не использовалось для нахождения параметров теоретического закона и теоретических частот, а эмпирические частоты не связаны никакими дополнительными соотношениями, то ***k*** равно числу групп эмпирического распределения. Количество частот в группе должно быть

больше 5. Если количество меньше, то соседние группы следует объединить.

Контрольные вопросы

1. Какие методы проверки гипотез существуют в статистике?
2. Опишите сущность параметрических методов проверки гипотез.
3. Объясните, чем отличаются непараметрические методы проверки гипотез от параметрических.
4. Как определить число связей и число степеней свободы?
5. Что такое доверительный интервал и как он определяется?

Выборочная совокупность. Вариационный ряд

Цель: *изучение типов выборок, методов анализа выборочной совокупности. Получение представлений о статистической оценке параметров и интервалов вариационного ряда.*

Основной целью статистического анализа является выяснение некоторых свойств рассматриваемой генеральной совокупности. Если генеральная совокупность конечна, то наилучшая процедура – рассмотреть каждый ее элемент.

Однако чаще всего на практике приходится ограничиваться выборочными значениями из генеральной совокупности. Основное требование к ***выборке*** – хорошо представлять (быть репрезентативной, представительной) генеральную совокупность.

Обычно считается, что чтобы иметь право судить о генеральной совокупности по выборке, выборка должна быть образована ***случайно***. Это можно достичь различными способами (наиболее распространенными):

- собственно-случайная выборка;
- механическая;
- типическая;
- серийная.

Собственно-случайная выборка

Существует два подхода к решению данной задачи:

Простая случайная выборка с возвращением – объект извлекается из генеральной совокупности случайным образом, и перед извлечением следующего, возвращается обратно. Например, после отбора деталей на анализ соответствия стандарту из большой партии, их снова возвращают назад и партию перемешивают.

Выборка без возвращения – извлеченный объект не возвращается в генеральную совокупность, а значит, может появиться в выборке только один раз. Например, отбор деталей производится с конвейера и после деструктивного анализа (разрушающего), возврат уже не возможен.

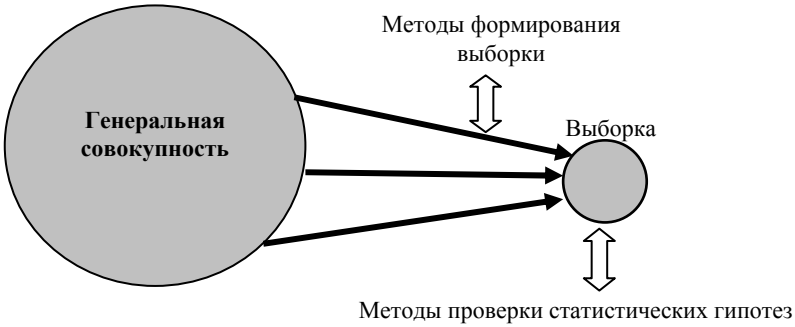


Рис. 23. Выборка элементов из генеральной совокупности

Если генеральная совокупность бесконечна, то процедуры выборки как с возвращением, так и без него, дают простую слу-

чайную выборку. Если генеральная совокупность конечна и велика по сравнению с размером выборки, то процедура извлечения без возвращения дает ***приблизительно*** простую случайную выборку. Если генеральная совокупность конечна и объем выборки составляет заметную долю от размера генеральной совокупности, то ***различие*** между этими двумя методами ***становится заметным***.

Механическая выборка

Механической называется выборка, в которую объекты из генеральной совокупности отбираются через *определенный интервал* (рис. 24).

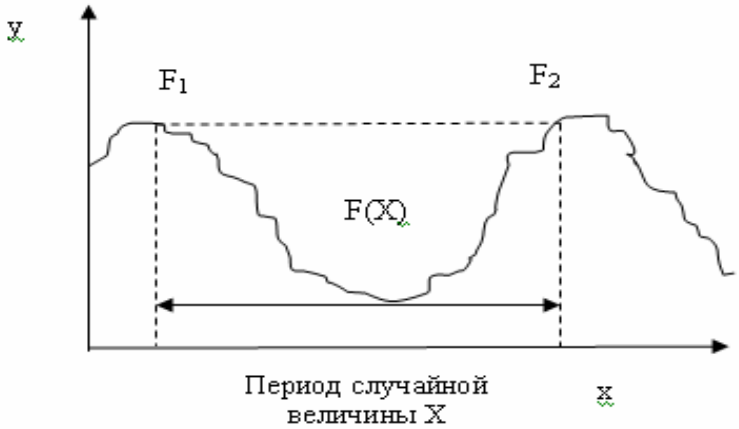


Рис. 24. Механическая выборка для периодической случайной величины

Например, если объем выборки должен составлять 5 % объема генеральной совокупности, то отбирается каждый двадцатый объект генеральной совокупности. Опасность, которая подстерегает исследователя при использовании этого метода – попасть в период циклически изменяющейся случайной величины.

Типическая выборка

Если генеральную совокупность предварительно разбить на непересекающиеся группы, а затем образовать собственно-случайные выборки элементов (с возвратом или без) из каждой группы и все отобранные объекты считать попавшими в выборку, то получим выборочную совокупность, называемую ***типической*** выборкой. Считается, что типическая выборка с большей достоверностью воспроизводит однородную генеральную совокупность.

Серийная выборка

Если генеральную совокупность предварительно разбить на непересекающиеся серии, а затем, рассматривая серии как элементы некой мегасовокупности, выбрать случайным образом несколько серий и все объекты отобранных серий поместить в выборку, то получим выборочную совокупность, которая называется серийной.

Всякая случайная выборка возникает из генеральной совокупности в результате случайного отбора, ее можно описать с помощью n-мерного случайного вектора $\Psi = (X_1; X_2; ...; X_n)$, составляющая которого X_i дает случайную величину X в i-м наблюдении ($i \in [1; n]$). Можно говорить о выборке объема n из распределенной согласно F(x) генеральной совокупности, когда случайные компоненты $X_i (i \in [1; n])$ независимы друг от друга и имеют ту же функцию распределения, что и X, а именно F(x).

Закон распределения случайного вектора $\Psi = (X_1; X_2; ...; X_n)$ в этом случае полностью определяется формулой: Эта формула получается из условий: F(x) – вероятность, x_i – независимые события. Здесь мы вторгаемся в область многомерных случайных величин и функций.

Отсюда следует, что каждая из рассчитанных по наблюдениям x_1, x_2, ..., x_n данной выборки числовая характеристика, например среднее арифметическое , есть реализация случайной величины, которая от выборки к выборке может принимать различные значения.

Такая случайная величина называется выборочной функцией и в общем случае обозначается как

$$T = T(X_1; X_2; ...; X_n).$$

Такая запись означает зависимость выборочной функции от случайных компонент $X_i(i \in [1; n])$ вектора Ψ.

Так как выборочная функция T является случайной величиной, то она имеет закон распределения, зависящий от закона распределения случайной величины X в генеральной совокупности. Для построения математико-статистических методов оценки параметров и проверки гипотез необходимо знание закона распределения, поэтому его расчет по распределению X для различных выборочных функций образует основную техническую проблему математической статистики.

Распределение среднего арифметического значения используется достаточно часто. Если из генеральной совокупности, распределение которой имеет математическое ожидание μ и дисперсию σ^2 (при этом закон распределения генеральной совокупности не обязательно должен быть нормальным) последовательно отбирать ряд выборок объема n, то каждая выборка даст реализацию величины . В итоге получается ряд средних арифметических для которых можно установить эмпирическое распределение и вычислить числовые характеристики. Тогда распределение частот с увеличением объема выборки ***n*** все более приближается по форме к нормальной кривой. Можно математически строго доказать, что имеет (для больших n) асимптоти-

чески нормальное распределение с математическим ожидание μ и дисперсией σ^2/n.

Некоторые важные распределения выборочных функций

Все приведенные ниже теоремы предполагают, что n компонент X_i(i ∈[1; n]) случайного вектора Ψ независимы и имеют нормальное распределение с математическим ожиданием μ и дисперсией σ^2, то есть имеем выборку объема n из нормально распределенной генеральной совокупности.

Теорема 1. Выборочная функция

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

распределена нормально согласно $N(\mu; \sigma^2/n)$.

Величина также приближенно распределена нормально согласно $N(\mu; \sigma^2/n)$, если соответствующая генеральная совокупность удовлетворяет произвольному распределению с математическим ожиданием μ и дисперсией σ^2. Приближение тем лучше, чем больше n.

Теорема 2. Выборочная функция (нормированная) удовлетворяет нормированному (стандартному) нормальному распределению с параметрами N(0; 1).

$$Z = \frac{X - \mu}{\sigma}\sqrt{n}.$$

Теорема 3. Образованная с помощью эмпирической дисперсии (деление на n-1 дает несмещенную оценку) выборочная функция удовлетворяет непрерывной функции распределения с плотностью для x>0 при m = n-1.

Значение C_m зависит только от m, но не от x, и его следует выбирать так, чтобы выполнялось условие нормирования для плотности распределения (). Определенная здесь плотность рас-

пределения называется распределением χ^2(хи-квадрат) с m = n-1 степенями свободы. При этом n и m являются параметрами.

Плотность распределения хи-квадрат асимметрична, но при $n \to \infty$ приближается к плотности нормального распределения с математическим ожиданием m=n-1 и дисперсией 2m = 2(n-1).

Теорема 4. Если из выборочных функций $\overline{x}$ и $\mathbf{S^2}$ образовать новую функцию $t = \frac{\overline{X} - \mu}{S}\sqrt{n}$, то для нее доказано, что эта выборочная функция непрерывна, а ее плотность равна: $f_t(x) = D_m(1 + \frac{x^2}{m})^{-(m+1)/2}$ для $-\infty < x < +\infty$, где m=n-1, а D_m — константа, зависимая только от m.

Эта плотность вероятности получена У. С. Гассетом и названа по его псевдониму – Student. Распределение Стьюдента или t-распределение с m=n-1 степенями свободы.

Кривая плотности вероятности тем более полога, чем меньше m, и при $m \to \infty$ переходит в плотность вероятности нормированного (стандартного) распределения.

Контрольные вопросы

1. Каковы задачи статистического анализа?
2. Что такое генеральная совокупность?
3. Приведите примеры генеральной совокупности, выборки и исследуемого признака.
4. Какие типы выборок существуют?
5. Для чего исследуются числовые характеристики выборок?

Статистические оценки параметров распределения

Цель: *получение представлений о статистической оценке параметров и интервалов выборочного распределения.*

Точечные оценки параметров распределения

Пусть требуется изучить количественный признак генеральной совокупности. Допустим, что из теоретических соображений удалось установить, какое именно распределение имеет признак. Возникает задача оценки параметров, которыми определяется это распределение.

Обычно в распоряжении исследователя имеются лишь данные выборки, полученные в результате n наблюдений (здесь и далее наблюдения предполагаются независимыми). Через эти данные и выражают оцениваемый параметр. Рассматривая значения количественного признака как независимые случайные величины, можно сказать, что найти статистическую оценку неизвестного параметра теоретического распределения – это значит найти функцию от наблюдаемых случайных величин, которая и дает приближенное значение оцениваемого параметра.

Итак, статистической оценкой неизвестного параметра теоретического распределения называют функцию от наблюдаемых случайных величин.

Для того чтобы статистические оценки давали «хорошие» приближения оцениваемых параметров, они должны удовлетворять определенным требованиям: оценка должна быть несмещенной, эффективной и состоятельной.

Несмещенной называют статистическую оценку $\mathbf{Q^*}$, математическое ожидание которой равно оцениваемому параметру Q при любом объеме выборки, т. е.

$$\mathbf{M(Q^*) = Q.}$$

Смещенной называют оценку, математическое ожидание которой не равно оцениваемому параметру.

Эффективной называют статистическую оценку, которая (при заданном объеме выборки n) имеет наименьшую возможную дисперсию.

При рассмотрении выборок большого объема (n велико!) к статистическим оценкам предъявляется требование состоятельности.

Состоятельной называют статистическую оценку, которая при $n\to\infty$ стремится по вероятности к оцениваемому параметру. Например, если дисперсия несмещенной оценки при $n\to\infty$ стремится к нулю, то такая оценка оказывается и состоятельной.

Рассмотрим точечные оценки параметров распределения, т.е. оценки, которые определяются одним числом $Q^* = f(x_1, x_2,\ldots,x_n)$, где $x_1, x_2,\ldots,x_n$- выборка.

Генеральная средняя

Пусть изучается генеральная совокупность относительно количественного признака X.

Генеральной средней называют среднее арифметическое значений признака генеральной совокупности.

Если все значения признака различны, то

$$\overline{x_г} = \frac{x_1 + x_2 + x_N}{N}.$$

Если значения признака распределены по частотам:

$N_1, N_2, \ldots, N_k$, где $N_1 + N_2 + \ldots + N_k = N$, то

$$\overline{x_г} = \frac{x_1 \cdot N_1 + x_2 \cdot N_2 + x_k \cdot N_k}{N}.$$

Выборочная средняя

Пусть для изучения генеральной совокупности относительно количественного признака X извлечена выборка объема n.

Выборочной средней называют среднее арифметическое значение признака выборочной совокупности.

Если все значения признака выборки различны, то

$$\overline{x_в} = \frac{x_1 + x_2 + x_n}{n},$$

а если же все значения имеют частоты n_1, n_2,…,n_k, то

$$\overline{x_в} = \frac{x_1 \cdot n_1 + x_2 \cdot n_2 + x_k \cdot n_k}{n}.$$

Выборочная средняя является несмещенной и состоятельной оценкой генеральной средней. Замечание: Если выборка представлена интервальным вариационным рядом, то за x_i принимают середины частичных интервалов.

Генеральная дисперсия

Для того чтобы охарактеризовать рассеяние значений количественного признака X генеральной совокупности вокруг своего среднего значения, вводят сводную характеристику – генеральную дисперсию.

Генеральной дисперсией $D_г$ называют среднее арифметическое квадратов отклонений значений признака генеральной совокупности от их среднего значения. Если все значения признака генеральной совокупности объема N различны, то

$$D_г = \frac{\sum_{i=1}^{N}(x_i - \overline{x_г})^2}{N}.$$

Если же значения признака имеют соответственно частоты $N_1, N_2, \ldots, N_k$, где $N_1 + N_2 + \ldots + N_k = N$, то

$$D_г = \frac{\sum_{i=1}^{N} N_i(x_i - \overline{x_г})^2}{N}.$$

Кроме дисперсии для характеристики рассеяния значений признака генеральной совокупности вокруг своего среднего значения пользуются сводной характеристикой – средним квадратическим отклонением.

Генеральным средним квадратическим отклонением (стандартом) называют квадратный корень из генеральной дисперсии.

Выборочная дисперсия

Для того, чтобы наблюдать рассеяние количественного признака значений выборки вокруг своего среднего значения, вводят сводную характеристику- выборочную дисперсию.

Выборочной дисперсией называют среднее арифметическое квадратов отклонения наблюдаемых значений признака от их среднего значения.

Если все значения признака выборки различны, то

$$D_в = \frac{\sum_{i=1}^{N}(x_i - \overline{x_в})^2}{n},$$

если же все значения имеют частоты $n_1, n_2, \ldots, n_k$, то

$$Dв = \frac{\sum_{i=1}^{k} n_i (x_i - \overline{x_в})^2}{n}.$$

Для характеристики рассеивания значений признака выборки вокруг своего среднего значения пользуются сводной характеристикой – средним квадратическим отклонением.

Генеральным средним квадратическим отклонением называют квадратный корень из генеральной дисперсии:

$$\sigma_г = \sqrt{D_г}\ .$$

Выборочным средним квадратическим отклонением называют квадратный корень из выборочной дисперсии:

$$\sigma_в = \sqrt{D_в}\,.$$

Вычисление дисперсии – выборочной или генеральной, можно упростить, используя формулу: $D = \overline{x^2} - \left[\overline{x}\right]^2$.

Замечание: если выборка представлена интервальным вариационным рядом, то за x_i принимают середины частичных интервалов.

Исправленная дисперсия

Выборочная дисперсия является смещенной оценкой генеральной дисперсии, т.е. математическое ожидание выборочной дисперсии не равно оцениваемой генеральной дисперсии, а равно

$$M(D_в) = \frac{n-1}{n} \cdot D_г\,.$$

Для исправления выборочной дисперсии достаточно умножить ее на дробь

$$\frac{n}{n-1},$$

получим исправленную дисперсию S^2. Исправленная дисперсия является несмещенной оценкой. В качестве оценки генеральной дисперсии принимают исправленную дисперсию.

Для оценки среднего квадратического генеральной совокупности используют исправленное среднее квадратическое отклонение

$$S = \sqrt{S^2}.$$

Замечание: формулы для вычисления выборочной дисперсии и исправленной дисперсии отличаются только знаменателями. При достаточно больших n выборочная и исправленная дисперсии мало отличаются, поэтому на практике исправленной дисперсией пользуются, если n<30.

Вычислим выборочные характеристики по выборкам, рассмотренным в табл. 16.

Таблица 16

Дискретный вариационный ряд

xi	0	1	2	3	4	5	7
Частота	8	17	16	10	6	2	1
Pi	8/60	17/60	16/60	10/60	6/60	2/60	1/60

Пример 1. Для дискретного вариационного ряда:

Среднее выборочное

$$\bar{x}_в = \frac{1}{60}(0\cdot 8+1\cdot 17+2\cdot 16+3\cdot 10+4\cdot 6+5\cdot 2+7\cdot 1) = 2.$$

Выборочная дисперсия

$$D_в = \left[\begin{array}{l}(0-2)^2 \cdot 8 + (1-2)^2 \cdot 17 + (2-2)^2 \cdot 16 + \\ (3-2)^2 \cdot 10 + (4-2)^2 \cdot 6 + (5-2)^2 \cdot 2 + (7-2)^2 \cdot 1\end{array}\right] \cdot \frac{1}{60} = 2{,}1.$$

Выборочное среднее квадратическое отклонение

$$\sigma_в = \sqrt{2{,}1} \approx 1{,}449.$$

Исправленная дисперсия

$$S^2 = 2{,}1 \cdot \frac{60}{59} \approx 2{,}136.$$

Рассмотрим пример расчета точечных оценок параметров интервального вариационного ряда (табл. 17).

Таблица 17

Данные интервального вариационного ряда

№	$x_i - x_{i+1}$	Частоты	P_i
1	6,67-6,69	2	0,01
2	6,69-6,71	15	0,075
3	6,71-6,73	17	0,085
4	6,73-6,75	44	0,22
5	6,75-6,77	52	0,26
6	6,77-6,79	44	0,22
7	6,79-6,81	14	0,07
8	6,81-6,83	11	0,055
9	6,83-6,85	1	0,005
	$\sum$	200	1

За x_i примем середины частичных интервалов:

$$\bar{x}_в = \frac{1}{200}(6{,}68 \cdot 2 + 6{,}7 \cdot 15 + 6{,}72 \cdot 17 + 6{,}74 \cdot 44 + 6{,}76 \cdot 52 + 6{,}78 \cdot 44 + 6{,}8 \cdot 14 + 6{,}82 \cdot 11 + 6{,}84 \cdot 1) = 6{,}7578.$$

Для вычисления выборочной дисперсии воспользуемся формулой

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

$$\overline{x_в^2} = \frac{1}{200} \cdot \begin{pmatrix} 6{,}68^2 \cdot 2 + 6{,}7^{2^2} \cdot 15 + 6{,}72^2 \cdot 17 + \\ + 6{,}74^2 \cdot 44 + 6{,}76^2 \cdot 52 + 6{,}78^2 \cdot 44 + \\ + 6{,}8^2 \cdot 14 + 6{,}82^2 \cdot 11 + 6{,}84^2 \cdot 1 \end{pmatrix} \approx 45{,}6688.$$

Тогда выборочная дисперсия равна $D_в \approx 45{,}6688 - 6{,}7578 \approx 0{,}001.$

Выборочное среднее квадратическое отклонение:
$\sigma_в = \sqrt{0{,}001} \approx 0{,}0316.$

Интервальные оценки параметров распределения

Интервальной называют оценку, которая определяется двумя числами – концами интервала. Интервальные оценки позволяют установить точность и надежность оценок.

Пусть найденная по данным выборки статистическая характеристика Q* служит оценкой неизвестного параметра Q. Будем считать Q постоянным числом (Q может быть и случайной величиной). Ясно, что Q* тем точнее определяет параметр Q, чем меньше абсолютная величина разности |Q- Q*|. Другими словами, если δ>0 и |Q- Q*| <δ , то чем меньше δ , тем оценка точнее.

Таким образом, положительное число δ характеризует точность оценки. Однако статистические методы не позволяют категорически утверждать, что оценка Q* удовлетворяет неравенству |Q- Q*| <δ; можно лишь говорить о вероятности γ, с которой это неравенство осуществляется.

Надежностью (доверительной вероятностью) оценки называют вероятность γ, с которой осуществляется неравенство $|Q–Q^*| <\delta$.

Обычно надежность оценки задается наперед, причем в качестве γ берут число, близкое к единице. Наиболее часто задают надежность, равную 0,95; 0,99 и 0,999.

Пусть вероятность того, что, $|Q- Q^*| <d$ равна γ:

$$P(|Q- Q^*| <d)= \gamma.$$

Заменив неравенство, равносильным ему двойным неравенством получим:

$$P [Q^* –\delta< Q < Q^* +\delta] = \gamma.$$

Это соотношение следует понимать так: вероятность того, что интервал $Q^* – d< Q < Q^* +d$ заключает в себе (покрывает) неизвестный параметр Q, равна γ.

Интервал $(Q^* – \delta\ Q^* +\delta)$ называется доверительным интервалом, который покрывает неизвестный параметр с надежностью γ.

Интервальные оценки параметров нормального распределения

Доверительный интервал для оценки математического ожидания при известном среднем квадратичном отклонении

Пусть количественный признак генеральной совокупности распределен нормально. Известно среднее квадратическое отклонение этого распределения σ. Требуется оценить математическое ожидание *a* по выборочной средней. Найдем доверительный интервал, покрывающий математическое ожидание *a* с надежностью γ. Выборочную среднюю будем рассматривать как случайную величину (она изменяется от выборки к выборке),

выборочные значения признака – как одинаково распределенные независимые случайные величины с математическим ожиданием *a* и средним квадратическим отклонением σ. Примем без доказательства, что если величина X распределена нормально, то и выборочная средняя тоже распределена нормально с параметрами

$$M(\bar{x}) = a, \sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}}.$$

Потребуем, чтобы выполнялось равенство

$$P(|\bar{x} - a| < \delta) = \gamma.$$

Заменив X и σ, получим

$$P(|\bar{x} - a| < \delta) = 2\Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right), \text{ обозначим } \frac{\delta\sqrt{n}}{\sigma} = t \text{ и выразив } \delta = \frac{t\sigma}{\sqrt{n}},$$

получим $$P\left(\bar{x} - \frac{t\sigma}{\sqrt{n}} < a < \bar{x} + \frac{t\sigma}{\sqrt{n}}\right) = \gamma = 2\Phi(t).$$

Задача решена. Число t находят по таблице функции Лапласа Φ(x).

Пример 1. Случайная величина X распределена нормально и σ =3. Найти доверительный интервал для оценки математического ожидания по выборочным средним, если n = 36 и задана надежность γ =0,95.

Из соотношения 2Φ(t)= 0,95 , откуда Φ(t) = 0,475 по таблице найдем

t =1,96. Точность оценки

$$\delta = \frac{t\sigma}{\sqrt{n}} = \frac{1{,}96 \cdot 3}{\sqrt{36}} = 0{,}98.$$

Доверительный интервал

$$(\bar{x}-0{,}98, \bar{x}+0{,}98)\,.$$

Пример 2. Найти минимальный объем выборки, который обеспечивает заданную точность δ =0,3 и надежность γ = 0,975, если СВХ распределена нормально и σ =1,2.

Из равенства

$$\delta = \frac{t\sigma}{\sqrt{n}}$$

выразим n:

$$n = \frac{t^2\sigma^2}{\delta^2} = \frac{2{,}24^2 \cdot 1{,}44}{0{,}09} = 80,$$

подставим значения и получим минимальный объем выборки n=80.

Доверительный интервал для оценки математического ожидания при неизвестном среднем квадратичном отклонении

Поскольку мы не знакомы с законами распределения случайной величины, которые используются при выводе формулы, то примем ее без доказательства.

В качестве неизвестного параметра σ используют исправленную дисперсию s^2 . Заменяя σ на s, t на величину t_γ. Значение этой величины зависит от надежности γ и объема выборки n и определяется по таблице значений t_γ. Итак:

$$P\left(\bar{x} - \frac{t_\gamma s}{\sqrt{n}} < a < \bar{x} + \frac{t_\gamma s}{\sqrt{n}}\right) = \gamma$$

и доверительный интервал имеет вид

$$\left(\bar{x}-\frac{t_\gamma \sigma}{\sqrt{n}};\ \bar{x}+\frac{t_\gamma \sigma}{\sqrt{n}}\right).$$

Пример 3. Найти доверительный интервал для оценки математического ожидания с надежностью 0,95, если объем выборки n =16, среднее выборочное и исправленная дисперсия соответственно равны 20,2 и 0,8.

По таблице приложения найдем $\mathbf{t_\gamma}$ по заданной надежности γ =0,95 и n= 16: $\mathbf{t_\gamma}$ =2,13. Подставим в формулу s =0,8 и t_γ =2,13 , вычислим границы доверительного интервала:

$$20{,}2-\frac{2{,}13\cdot 08}{4}<a<20{,}2+\frac{2{,}13\cdot 0{,}8}{4},$$

откуда получим доверительный интервал (19,774; 20,626).

Смысл полученного результата: если взять 100 различных выборок, то в 95 из них математическое ожидание будет находиться в пределах данного интервала, а в 5 из них – нет.

Пример 4. Измеряют диаметры 25 корпусов электродвигателей. Получены выборочные характеристики

$$\bar{x}=100\,мм,\ s=16\,мм.$$

$\bar{x}=100$ *мм*, s=16 *мм*.

Необходимо найти вероятность (надежность) того, что интервал: $(0{,}9\,\bar{x};\ 1{,}1\,\bar{x})$ является доверительным интервалом оценки математического ожидания при нормальном распределении. Из условия задачи найдем точность d, составив и решив систему:

$$\begin{cases} 0{,}9\bar{x}=\bar{x}-\delta; \\ 1{,}1\bar{x}=\bar{x}+\delta. \end{cases}$$

Откуда d =10. Из равенства

$$\delta = \frac{t_\gamma s}{\sqrt{n}}$$ выразим

$$t_\gamma = \frac{\delta\sqrt{n}}{s},$$

откуда t_γ =3,125. По таблице для найденного t_γ и n= 25 находим γ =0,99.

Доверительный интервал для оценки дисперсии и среднего квадратического отклонения

Требуется оценить неизвестную генеральную дисперсию и генеральное среднее квадратическое отклонение по исправленной дисперсии, т.е. найти доверительные интервалы, покрывающие параметры D и σ с заданной надежностью γ.

Потребуем выполнения соотношения

$$P(|\sigma - s| < \delta) = \gamma.$$

Раскроем модуль и получим двойное неравенство:

$$s - \delta < \sigma < s + \delta.$$

Преобразуем:

$$s(1 - \frac{\delta}{s}) < \sigma < s(1 + \frac{\delta}{s}).$$

Обозначим d/s = q (величина q находится по "Таблице значений q"и зависит от надежности и объема выборки), тогда доверительный интервал для оценки генерального среднего квадратического отклонения имеет вид:

$$s(1 - q) < \sigma < s(1 + q).$$

Замечание: Так как s >0, то если q >1 , левая граница интервала равна 0:

$$0< s < s (1 + q).$$

Пример 1. По выборке объема n = 25 найдено "исправленное" среднее квадратическое отклонение s = 0,8. Найти доверительный интервал, покрывающий генеральное среднее квадратическое отклонение с надежностью 0,95.

По таблице приложения по данным : $\gamma = 0,95$; n =25 , находим q = 0,32.

Искомый доверительный интервал $0,8(1- 0,32)< s < 0,8(1+ 0,32)$ или $0,544<s <0,056$.

Пример 2. По выборке объема n = 10 найдено s = 0,16. Найти доверительный интервал, покрывающий генеральное среднее квадратическое отклонение с надежностью 0,999.

$$q(n=10, \gamma =0,999) = 1,8>0.$$

Искомый доверительный интервал $0< s <0,16(1+1,8)$ или $0< s <0,448$.

Так как дисперсия есть квадрат среднего квадратического отклонения, то доверительный интервал, покрывающий генеральную дисперсию с заданной надежностью γ, имеет вид:

$$s^2(1-q)^2 < D < s^2(1+q)^2 \text{, если } q < 1;$$
$$0 < D < s^2(1+q)^2 \text{, если } q > 1$$

Контрольные вопросы

1. Определение статистической оценки неизвестного параметра.
2. Какая оценка называется точечной?
3. Каким требованиям должны удовлетворять статистические оценки?

4. Сформулировать определения генеральной средней и генеральной дисперсии.

5. Записать выражения для вычисления выборочной средней, выборочной дисперсии и исправленной дисперсии. Какая из этих оценок не является несмещенной?

6. Методики вычисления границ доверительного интервала для оценки математического ожидания нормально распределенной СВ при известном и неизвестном σ.

7. Методика вычисления границ доверительного интервала для оценки среднего квадратического отклонения нормально распределенной СВ.

8. Доверительный интервал вероятности биноминального распределения по относительной частоте при больших n , при $n<100$.

Линейный корреляционный анализ

Цель: *изучение связей между величинами, носящими случайный характер. Проверка гипотез о линейной и нелинейной корреляции величин.*

Исключительный интерес для широкого класса задач представляет обнаружение взаимных связей между двумя и более случайными величинами. Например, существует ли связь между курением и ожидаемой продолжительностью жизни или между умственными способностями и успеваемостью. В инженерных применениях такие задачи обычно сводятся к установлению связи между некоторым предполагаемым возбуждением и наблюдаемым откликом изучаемой физической системы.

Корреляционный анализ (термин “корреляция” происходит от лат. correlatio – соотношение, связь) измеряет степень взаимосвязи между двумя переменными – например, ценой товара на рынке и объемом спроса на этот товар. Величина, характеризующая наличие связи – коэффициент корреляции показывает, приведут ли изменения одной переменной, например, цены к изменениям другой – спроса.

При корреляционном анализе двух переменных одна из них называется "зависимая", а другая – "независимая". Цель анализа – определить, приведут ли изменения независимой переменной к изменениям зависимой.

Из математики нам известно понятие функции, которая описывает зависимость значения величины Y от значения независимой переменной X, называемой аргументом. Однозначная зависимость между переменными величинами Y и X называется функциональной, т.е. $Y = f(X)$ (“игрек есть функция от икс”). Например, в функции Y = -3X+5 каждому значению X соответствует значение Y. В функции $Y = X^3$ каждому значению X соответствует Y , равный кубу X. Но такого рода однозначные или функциональные связи между переменными величинами встречаются не всегда. Известно, например, что между ростом и массой человека существует положительная связь: более высокие индивиды имеют обычно и большую массу, чем индивиды низкого роста. То же наблюдается и в отношении качественных признаков: блондины, как правило, имеют голубые, а брюнеты – карие глаза. Однако из этого правила имеются исключения, когда сравнительно низкорослые индивиды оказываются тяжелее высокорослых, и среди людей встречаются кареглазые блондины и голубоглазые брюнеты. Причина таких “исключений” в том, что каждый биологический признак, выражаясь математи-

ческим языком, является функцией многих переменных; на его величине сказывается влияние и генетических, и средовых факторов, в том числе и случайных, что вызывает варьирование признаков. Отсюда зависимость между ними приобретает не функциональный, а статистический характер, когда определенному значению одного признака, рассматриваемого в качестве независимой переменной, соответствует не одно и то же числовое значение, а целая гамма распределяемых в вариационный ряд числовых значений другого признака, рассматриваемого в качестве независимой переменной. Такого рода зависимость между переменными величинами называется корреляционной**.** Если функциональные связи одинаково легко обнаружить и на единичных, и на групповых объектах, то этого нельзя сказать о связях корреляционных, которые изучаются только на групповых объектах методами математической статистики.

Задача корреляционного анализа сводится к установлению направления и формы связи между признаками, измерению ее тесноты и к оценке достоверности выборочных показателей корреляции.

Для двух случайных величин ***x*** и ***y*** коэффициент корреляции определяется по формуле:

$$r_{xy} = \frac{s_{xy}}{s_x s_y},$$

где s_{xy} — ковариация x и y, определяемая по формуле, а s_x и s_y – средние квадратичные отклонения по выборкам.

$$s_{xy} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{n-1}.$$

Коэффициент корреляции Браве–Пирсона по выборочным данным можно оценить по формуле:

$$r_{xy} = \frac{\sum_{i=1}^{n} (X_i - \overline{X}) \cdot \left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2 \cdot \sum_{i=1}^{n} (Y_i - \overline{Y})^2}}.$$

Выборочный коэффициент корреляции лежит между -1 и +1 и принимает одно из граничных значений только при наличии идеальной линейной связи между наблюдениями. Нелинейная связь и (или) разброс данных, вызванный ошибками измерения или же неполной коррелированностью случайных величин, приводит к уменьшению абсолютного значения r_{xy}.

Данный коэффициент впервые использовал Карл Пирсон (1857–1936), английский математик, разработавший статистический аппарат для проверки теории Ч.Дарвина. Статистические методики Пирсона широко используются в психологии и педагогике.

Расчет коэффициента корреляции правомерно проводить в том случае, когда случайные величины могут быть измерены в числовой шкале, при этом возможно вычисление средних значений. Приведём пример, когда нахождение коэффициента корреляции некорректно именно по причине измерения случайных величин в качественной шкале. Любые измеряемые величины соотносят с одной из измерительных шкал. Обычно выделяют две качественные шкалы: **номинальную и порядковую.** Номинальная позволяет только качественно отличить один объект от другого, например черное – белое, Марина – Пётр – Саша. Порядковая или ранговая шкала позволяет установить порядок

увеличения или уменьшения какого-либо качества: низкий – средний – высокий, плохо – удовлетворительно – хорошо – отлично и т.д.

Количественные шкалы – **интервалов** и **отношений**, позволяют сравнивать величины между собой и выражать различие числом. Когда исследуется корреляция между количественными признаками, значения которых можно точно измерить в единицах метрических шкал (метры, секунды, килограммы и т.д.), то очень часто принимается модель двумерной нормально распределенной генеральной совокупности. Такая модель отображает зависимость между переменными величинами x_i и y_i графически в виде геометрического места точек в системе прямоугольных координат. Эту графическую зависимость называют также *диаграммой рассеивания* или *корреляционным полем* (рис. 25).

Данная модель двумерного нормального распределения (корреляционное поле) позволяет дать наглядную графическую интерпретацию коэффициента корреляции, т.к. распределение в совокупности зависит от пяти параметров: m_x, m_y – средние значения (математические ожидания); s_x,s_y – стандартные отклонения случайных величин X и Y и p – коэффициент корреляции, который является мерой связи между случайными величинами X и Y.

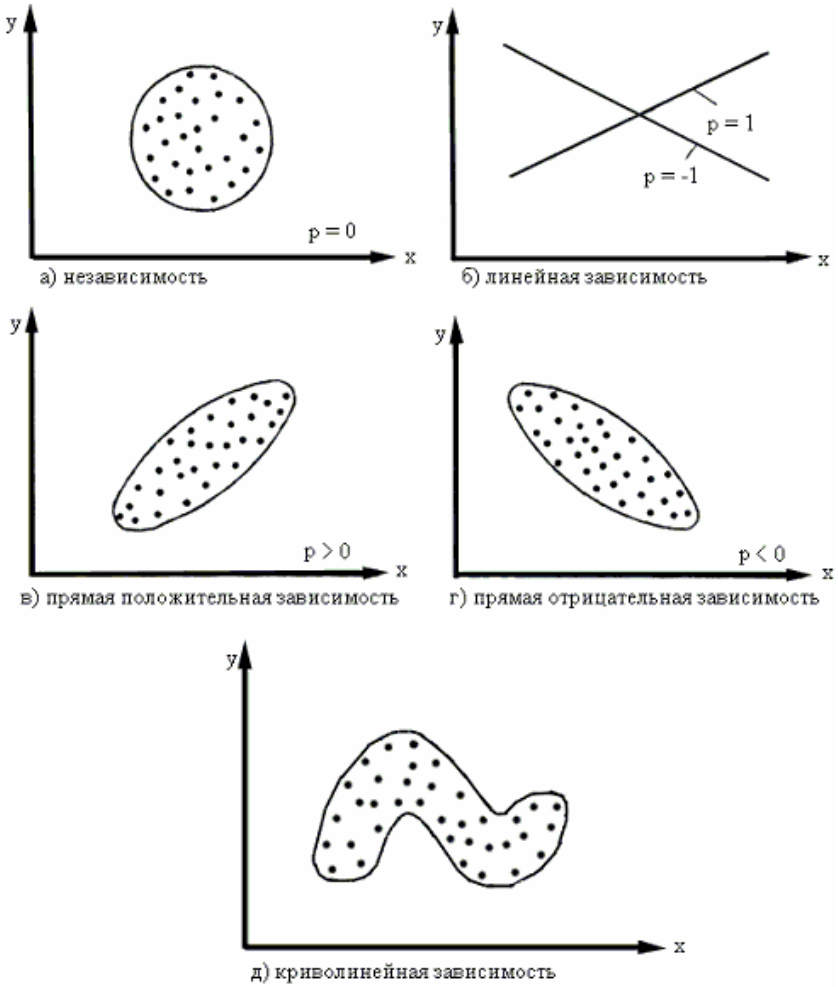


Рис. 25. Диаграммы рассеяния, характеризующие зависимость между двумя случайными величинами

Если p = 0, то значения, x_i, y_i, полученные из двумерной нормальной совокупности, располагаются на графике в координатах x, y в пределах области, ограниченной окружностью. В этом случае между случайными величинами X и Y отсутствует

корреляция и они называются некоррелированными. Для двумерного нормального распределения некоррелированность означает одновременно и независимость случайных величин *X* и *Y*.

Пример. Определить достоверность взаимосвязи между показателями веса и количеством подтягиваний на перекладине у 11 исследуемых с помощью расчета нормированного коэффициента корреляции, если данные выборок таковы:

Таблица 18

Результаты эксперимента

№	1	2	3	4	5	6	7	8	9	10	11
Вес, x_i,	51	50	48	51	46	47	49	60	51	52	56
Кол-во подтягиваний, y_i	13	15	13	16	12	14	12	10	18	10	12

Решение

1.Расчет коэффициента корреляции Пирсона:

$$r_{xy} = \frac{\sum_{i=1}^{n} (X_i - \overline{X}) \cdot \left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2 \cdot \sum_{i=1}^{n} (Y_i - \overline{Y})^2}}.$$

2. Для расчетов создать вспомогательную таблицу

$\mathbf{x_i}$	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$\mathbf{y_i}$	$y_i - \overline{y}$	$(y_i - \overline{y})^2$	$(x_i - \overline{x})(y_i - \overline{y})$
	Σ	Σ162		Σ	Σ60	Σ-34

$$r_{xy} = \frac{-34}{\sqrt{162 \cdot 60}} = -0{,}34.$$

3. Рассчитать число степеней свободы по формуле:

$$K = n - 2 \text{ , } K = 11 - 2 = 9.$$

4. Сравнить рассчитанное значение нормированного коэффициента корреляции ($r_ф$ = -0,34) с табличным значением для К =9 при а = 5 % и сделать вывод.

Вывод

1) так как $r_ф$ = -0,34 < 0, то между данными выборок наблюдается обратная отрицательная взаимосвязь, то есть с увеличением показателей веса у исследуемых снижается их результат в количестве подтягиваний на перекладине;

2) поскольку $r_ф$ = -0,34 < r_{st} = 0,60 для К = 10 при α = 5 %, то с уверенностью γ = 95 % можно говорить о том, что выявленная зависимость недостоверна.

Контрольные вопросы

1. Каковы задачи корреляционного анализа?
2. Что измеряет коэффициент ковариации?
3. Как рассчитывается коэффициент корреляции?
4. Какие виды функциональных связей существуют между величинами?
5. Приведите значения коэффициента корреляции и вид связей, который соответствует каждому значению.
6. Какие существуют шкалы измерений?
7. Что представляет собой диаграмма рассеяния?

Основы дисперсионного анализа

Цель: *познакомиться с теорией дисперсионного анализа.*

Дисперсионный анализ (от латинского Dispersio – рассеивание) – статистический метод, позволяющий анализировать влияние различных факторов на исследуемую переменную. Метод был разработан биологом Р. Фишером в 1925 году и применялся первоначально для оценки экспериментов в растениеводстве. В дальнейшем выяснилась общенаучная значимость дисперсионного анализа для экспериментов в психологии, педагогике, медицине и др.

Целью дисперсионного анализа является проверка значимости различия средних арифметических на основе сравнения дисперсий нескольких групп. Дисперсию измеряемого признака разлагают на независимые слагаемые, каждое из которых характеризует влияние того или иного фактора или их взаимодействия. Последующее сравнение таких слагаемых позволяет оценить значимость каждого изучаемого фактора, а также их комбинации.

Например, проводя опросы по поводу потребления какого-либо товара в различных регионах страны, необходимо сделать выводы на: сколько данные опроса отличаются или не отличаются друг от друга. Сопоставлять отдельные показатели не имеет смысла и поэтому процедура сравнения и последующей оценки производится по некоторым усредненным значениям и отклонениям от этой усредненной оценки. Изучается вариация признака. За меру вариации принимается дисперсия.

На практике часто возникают задачи более общего характера – задачи проверки существенности различий средних выборочных нескольких совокупностей.

Например, требуется оценить влияние различного сырья на качество производимой продукции, решить задачу о влиянии количества удобрений на урожайность с/х продукции.

Дисперсионный анализ включает в себя проверку гипотез, связанных с оценкой выборочной дисперсии. Можно выделить три основных вида гипотез:

1) значимо ли различие между двумя дисперсиями?
2) одна дисперсия значимо больше другой?
3) значимо ли различие между несколькими дисперсиями?

Гипотезой для дисперсионного анализа может служить и такая гипотеза: выборки, по которым определены оценки дисперсии, получены из генеральных совокупностей, обладающих одинаковыми дисперсиями

Иногда дисперсионный анализ применяется, чтобы установить однородность нескольких совокупностей. Дисперсии этих совокупностей одинаковы по предположению; если дисперсионный анализ покажет, что и математические ожидания одинаковы, то в этом смысле совокупности однородны. Однородные же совокупности можно объединить в одну и тем самым получить о ней более полную информацию, следовательно, и более надежные выводы.

В основе дисперсионного анализа лежит разделение дисперсии на части или компоненты. Вариацию, обусловленную влиянием фактора, положенного в основу группировки, характеризует межгрупповая дисперсия $S_{общ}$. Она является мерой вариации частных средних по группам $\overline{x_j}$ вокруг общей средней $\overline{\mathbf{x}}$ и определяется по формуле:

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

где p – число групп; n – число уровней фактора; $\bar{\bar{x}}$ – общая средняя.

Вариацию, характеризующую рассеяние между группами, описывает межгрупповая дисперсия:

$$S_{факт} = q \sum_{j=1}^{з} (\bar{x}_{гр.j} - \bar{x})^2.$$

Остаточная сумма квадратов отклонений наблюдаемых значений группы от своего группового среднего, характеризует рассеяние внутри групп:

$$S_{ост} = \sum_{i=1}^{q} (x_{i1} - \bar{x}_{гр1})^2 + \sum_{i=1}^{q} (x_{i2} - \bar{x}_{гр2})^2 + ... + \sum_{i=1}^{q} (x_{in} - \bar{x}_{гр.n})^2.$$

Между общей дисперсией $S_{общ}$, межгрупповой дисперсией $S_{фак}$ и внутригрупповой дисперсией $S_{ост}$ существует соотношение:

$$S_{общ} = S_{фак} + S_{ост}$$

Внутригрупповая дисперсия объясняет влияние неучтенных при группировке факторов, а межгрупповая дисперсия объясняет влияние факторов группировки на среднее значение по группе. Разделив суммы квадратов на соответствующее число степеней свободы, получим общую, факторную и остаточную дисперсии:

$$s^2_{общ} = \frac{S_{общ}}{n \cdot q - 1}, \quad s^2_{фак} = \frac{S_{фак}}{n-1}, \; s^2_{ост} = \frac{S_{ост}}{n(q-1)}.$$

Таблица 19

План факторного эксперимента для дисперсионного анализа

№ эксп.	Уровни фактора F			
	F_1	F_2	F_j	F_n
1	x_{11}	x_{12}	x_{13}	x_{1n}
2	x_{21}	x_{22}	x_{23}	x_{2n}
…	…	…	…	…
q	x_{q1}	x_{q2}	x_{q3}	x_{qn}
$\overline{x}_{гр}$	$\overline{x_1}$	$\overline{x_2}$	$\overline{x_j}$	$\overline{x_n}$
$S_{вгр}$	$s_{вгр1}$	$s_{вгр1}$	$s_{вгр1}$	$s_{вгр1}$
$S_{межгр}$	$s_{ф1}$	$s_{ф2}$	$s_{фj}$	$s_{фn}$

Если справедлива гипотеза H_0, то все эти дисперсии являются несмещенными оценками генеральной дисперсии. Покажем, что проверка нулевой гипотезы сводится к сравнению факторной и остаточной дисперсии по критерию Фишера-Снедекора.

1. Пусть гипотеза H_0 правильна. Тогда факторная и остаточная дисперсии являются несмещенными оценками неизвестной генеральной дисперсии и, следовательно, различаются незначимо. Поэтому результат оценки по критерию Фишера-Снедекора F покажет, что нулевая гипотеза принимается. Таким образом, если верна гипотеза о равенстве математических ожи-

даний генеральных совокупностей, то верна и гипотеза о равенстве факторной и остаточной дисперсий.

2. Если нулевая гипотеза неверна, то с возрастанием расхождения между математическими ожиданиями увеличивается и факторная дисперсия, а вместе с ней и отношение $F_{экс} = \frac{s^2_{фак}}{s^2_{ост}}$. Поэтому в результате $F_{экс}$ окажется больше $F_{кр}$, и гипотеза о равенстве дисперсий будет отвергнута. Следовательно, если гипотеза о равенстве математических ожиданий генеральных совокупностей ложна, то ложна и гипотеза о равенстве факторной и остаточной дисперсий.

Итак, метод дисперсионного анализа состоит в *проверке по критерию F нулевой гипотезы о равенстве факторной и остаточной дисперсий.*

Если факторная дисперсия окажется меньше остаточной, то гипотеза о равенстве математических ожиданий генеральных совокупностей верна. При этом нет необходимости использовать критерий *F*.

При обработке данных эксперимента наиболее разработанными и поэтому распространенными считаются две модели. Их различие обусловлено спецификой планирования самого эксперимента. В модели дисперсионного анализа с фиксированными эффектами исследователь намеренно устанавливает строго определенные уровни изучаемого фактора. Термин «фиксированный эффект» в данном контексте имеет тот смысл, что самим исследователем фиксируется количество уровней фактора и различия между ними. При повторении эксперимента он или другой исследователь выберет те же самые уровни фактора. В модели со случайными эффектами уровни значения фактора вы-

бираются исследователем случайно из широкого диапазона значений фактора, и при повторных экспериментах, естественно, этот диапазон будет другим.

Таким образом, данные модели отличаются между собой способом выбора уровней фактора, что, очевидно, в первую очередь влияет на возможность обобщения полученных экспериментальных результатов. Для дисперсионного анализа однофакторных экспериментов различие этих двух моделей не столь существенно, однако в многофакторном дисперсионном анализе оно может оказаться весьма важным.

При проведении дисперсионного анализа должны выполняться следующие статистические допущения: независимо от уровня фактора величины отклика имеют нормальный закон распределения и одинаковую дисперсию. Такое равенство дисперсий называется гомогенностью. Таким образом, изменение способа обработки сказывается лишь на положении случайной величины отклика, которое характеризуется средним значением или медианой. Поэтому все наблюдения отклика принадлежат сдвиговому семейству нормальных распределений.

Говорят, что техника дисперсионного анализа является "робастной". Этот термин, используемый статистиками, означает, что данные допущения могут быть в некоторой степени нарушены, но, несмотря на это, технику можно использовать. При неизвестном законе распределения величин отклика используют непараметрические (чаще всего ранговые) методы анализа.

Пример использования однофакторного дисперсионного анализа

В четырех группах испытуемых, по 17 человек в каждой, проводилось изучение времени реакции на звуковой стимул. Интенсивность стимула составила 40, 60, 80 и 100 дБ, причем в

каждой группе предъявлялись стимулы только одной интенсивности.

H_0: Среднее время реакции уменьшается по мере увеличения громкости звука. В этой задаче регулируемым фактором является сила звука, а её уровни рассматриваются как градации фактора. Таким образом, фактор «сила звука» выступает как независимая переменная, а время реакции как результативный признак, или как зависимая переменная. Проверяется гипотеза H_0, согласно которой средние и дисперсии в группах обусловлены случайными влияниями и не зависят от действия регулируемого фактора.

H_1: Среднее время реакции увеличивается по мере увеличения громкости звука.

Представим исходные данные для работы с однофакторным дисперсионным анализом в виде табл. 20, в которую внесены некоторые дополнительные расчетные данные.

Таблица 20

Однофакторный дисперсионный анализ по влиянию уровня звука на время реакции испытуемого

№	Группа 1	Группа 2	Группа 3	Группа 4	
исп.	40 дб.	60 дб.	80 дб.	100 дб.	
1	304	272	202	180	
2	268	264	178	160	
3	272	256	181	157	
4	262	269	183	167	
5	283	285	187	180	
6	265	247	186	167	
7	286	250	190	187	
8	257	245	167	156	
9	279	251	156	159	
10	275	261	183	171	
11	268	250	167	155	
12	254	228	176	158	
13	245	257	186	163	
14	253	214	192	161	
15	235	242	168	157	
16	260	222	176	150	
17	246	234	192	158	
Среднее	265,41	249,82	180,58	163,88	Итого:
Sост_вгр	291,88	336,90	131,25	104,36	864
Sфак_мгр	2548,76	1217,80	1179,11	2605,50	7551
S общ	2 122	8 416		s^2 фак	2517
Ср. общ.	214,92	Степени свободы		s^2 ост	216
F экс	11,64	n-1	3		

F таб	8,56	n(q-1)	64

Сравнивая F экс и F таб, можно сделать вывод, что F экс больше критического табличного значения, а это значит, что нулевую гипотезу H_0 об отсутствии различий следует отвергнуть, а принять гипотезу H_1. Психолог может быть уверенным, что при увеличении силы звука скорость реакции значительно увеличивается. Или регулируемый фактор – сила звука оказывает существенное влияние на независимую переменную – скорость реакции.

Контрольные вопросы

1. Какие цели преследует дисперсионный анализ?
2. Как формулируются гипотезы для дисперсионного анализа?
3. Опишите алгоритм проведения однофакторного дисперсионного анализа?
4. Как использовать критерий Фишера для сравнения выборочных дисперсий?

Факторный анализ

Цель: *освоить методику применения факторного анализа для исследования экспериментальных данных.*

Множество явлений и процессов в окружающем нас мире связаны между собой. Изучение взаимных зависимостей между составляющими явлений и процессов порождает множество вопросов: о силе связей, об их закономерностях, о причинах, породивших определенную структуру связей. Сложные зависимости системы факторов, влияющих на процесс, сложно интерпре-

тировать, поскольку в большинстве ситуаций существуют скрытые параметры, влияющие на коррелированные признаки.

Часто изменения взаимосвязанных признаков происходит согласованно, т.е. признаки дублируются. Стремление объяснить совокупность признаков через введение более глубинных характеристик явления, определяющих его структуру, приводит к модели факторного анализа.

Факторный анализ – многомерный статистический метод, применяемый для изучения взаимосвязей между значениями переменных.

Реализация факторного анализа представляет собой постепенный переход от исходной факторной системы к конечной факторной системе, изучение влияния полного набора прямых, количественно измеряемых факторов, оказывающих влияние на изменение результативного показателя.

Условия выполнения факторного анализа:

- факторный анализ выполняется над взаимосвязанными переменными;
- изучаемые признаки должны быть количественными;
- число признаков должно быть в два раза больше числа переменных;
- выборка должна быть однородна.

По характеру взаимосвязи между показателями различают методы детерминированного и стохастического факторного анализа.

Детерминированный факторный анализ представляет собой методику исследования влияния факторов, связь которых с результативным показателем носит функциональный характер.

Основные свойства детерминированного подхода к анализу:

- построение детерминированной модели путем логического анализа;
- наличие полной (жесткой) связи между показателями;
- невозможность разделения результатов влияния одновременно действующих факторов, которые не поддаются объединению в одной модели;
- изучение взаимосвязей в краткосрочном периоде.

Различают четыре типа детерминированных моделей:

Аддитивные модели представляют собой алгебраическую сумму показателей и имеют вид

$$Y = \sum_{i=1}^{n} x_i = x_1 + x_2 + ...x_i .$$

К таким моделям, например, относятся показатели себестоимости во взаимосвязи с элементами затрат на производство и со статьями затрат; показатель объема производства продукции в его взаимосвязи с объемом выпуска отдельных изделий или объема выпуска в отдельных подразделениях.

Мультипликативные модели в обобщенном виде могут быть представлены формулой

$$Y = \prod_{i=1}^{n} x_i = x_1 x_2 ... x_i .$$

Примером мультипликативной модели является двухфакторная модель объема реализации

$$Q = X \cdot \mathrm{Pr},$$

где X – среднесписочная численность работников;

Pr – средняя выработка на одного работника.

Кратные модели:

$$y = \frac{x_1}{x_2}.$$

Примером кратной модели служит показатель срока оборачиваемости товаров (в днях) . $Т_{ОБ.Т}$:

$$T_{ОБ.Т} = \frac{З_Т}{O_P},$$

где ЗТ – средний запас товаров; ОР – однодневный объем реализации.

Смешанные модели представляют собой комбинацию перечисленных выше моделей и могут быть описаны с помощью специальных выражений:

$$Y = (a + e) \cdot c; \quad Y = \frac{\prod_{i=1}^{n} x_i}{\sum_{j=1}^{m} x_j}; \quad Y = \frac{\sum_{i=1}^{n} x_i}{\sum_{j=1}^{m} x_j}; \quad Y = \frac{\prod_{i=1}^{n} x_i}{\prod_{j=1}^{m} x_j}.$$

Примерами таких моделей служат показатели затрат на 1 руб. товарной продукции, показатели рентабельности и др.

Алгоритмы применения детерминированного факторного анализа для различных моделей

1. Модель вида $y = a \cdot b$:

$$\Delta y\ (a) = e_0 \cdot \Delta a + \frac{1}{2}\Delta a \cdot \Delta e;$$

$$\Delta y\ (e) = a_0 \cdot \Delta e + \frac{1}{2}\Delta a \cdot \Delta e.$$

2. Модель вида $y = a \cdot b \cdot c$

$$\Delta y\ (a) = \frac{1}{2}\Delta a \cdot (e_0 c_1 + e_1 c_0) + \frac{1}{3} \cdot \Delta a \cdot \Delta e \cdot \Delta c;$$

$$\Delta y\ (e) = \frac{1}{2}\Delta e \cdot (a_0 c_1 + a_1 c_0) + \frac{1}{3} \cdot \Delta a \cdot \Delta e \cdot \Delta c;$$

$$\Delta y\ (c) = \frac{1}{2}\Delta c \cdot (a_0 e_1 + a_1 e_0) + \frac{1}{3} \cdot \Delta a \cdot \Delta e \cdot \Delta c.$$

3. Модель вида $y = \frac{a}{b}$:

$$\Delta y\,(a) = \frac{\Delta a}{\Delta e} \cdot \ln\left|\frac{e_1}{e_0}\right|;$$

$$\Delta y\,(e) = \Delta y - \Delta y\,(a).$$

4. Модель вида $y = \frac{a}{b + c}$:

$$\Delta y(a) = \frac{\Delta a}{\Delta a + \Delta c} \cdot \ln \left| \frac{a_1 + c_1}{a_0 + c_0} \right|;$$

$$\Delta y(b) = \frac{\Delta y - \Delta y(a)}{\Delta a + \Delta c} \cdot \Delta b;$$

$$\Delta y(c) = \frac{\Delta y - \Delta y(a)}{\Delta a + \Delta c} \cdot \Delta c.$$

Построение факторной модели – первый этап детерминированного анализа. Далее определяют способ оценки влияния факторов.

Способы оценки влияния факторов

- *Способ цепных подстановок*
- *Способ относительных разниц*
- *Способ абсолютных разниц*

Способ цепных подстановок заключается в определении ряда промежуточных значений обобщающего показателя путем последовательной замены базисных значений факторов на отчетные. Данный способ основан на исключении воздействия всех факторов на величину результативного показателя, кроме одного. При этом исходя из того, что все факторы изменяются независимо друг от друга, т.е. сначала изменяется один фактор, а все остальные остаются без изменения, потом изменяются два при неизменности остальных и т.д.

В общем виде применение способа цепных постановок можно описать следующим образом:

$$y_0 = a_0 \cdot b_0 \cdot c_0;$$
$$y_a = a_1 \cdot b_0 \cdot c_0;$$
$$y_b = a_1 \cdot b_1 \cdot c_0;$$
$$y_1 = a_1 \cdot b_1 \cdot c_1,$$

где a_0, b_0, c_0 – базисные значения факторов, оказывающих влияние на обобщающий показатель y; a_1 , b_1, c_1 – фактические значения факторов; y_a, y_b, – промежуточные изменения результирующего показателя, связанного с изменением факторов a, b, соответственно.

Общее изменение $\Delta y = y_1 - y_0$ складывается из суммы изменений результирующего показателя за счет изменения каждого фактора при фиксированных значениях остальных факторов:

$$\Delta y = \sum \Delta y(a,b,c) = \Delta y_a + \Delta y_b + \Delta y_c;$$
$$\Delta y_a = y_a - y_0;\ \Delta y_b = y_b - y_a;\ \Delta y_c = y_1 - y_b.$$

Проведем факторный анализ влияния на объем товарной продукции количества работников и их выработки описанным выше способом на основе данных табл.21.

Зависимость объема товарной продукции от данных факторов можно описать с помощью мультипликативной модели:

$$Q = X \cdot \mathrm{Pr}$$
$$Q_0 = X_0 \cdot \mathrm{Pr}_0 = 130 * 85 = 11050(тыс.руб.)$$

Таблица 21

Исходные данные для факторного анализа

Показатели	Базисные значения	Факт. значения	Абс. изменение	Относ. изменение
Кол-во работников, X	130	98	-32	75,38
Выработка на одного раб., Pr	85	72	-13	84,71
Объем продукции, Q (тыс. руб.)	11050	7056	-3994	63,86
$Q_{усл1}$	8330			
$\Delta Q_{усл1}$	-2720			
Q_1	7056			
$\Delta Q_{усл2}$	-1274			

Тогда влияние изменения величины количества работников на обобщающий показатель можно рассчитать по формуле:

$$Q_{усл1} = X_1 \cdot \mathrm{Pr}_0 = 98 * 85 = 8330(тыс.руб.)$$
$$\Delta Q_{усл1} = Q_{усл1} - Q_0 = 8330 - 11050 = -2720(тыс.руб.)$$

Далее определим влияние изменения выработки работников на обобщающий показатель

$$Q_1 = X_1 \cdot \mathrm{Pr}_1 = 98 * 72 = 7056(тыс.руб.)$$
$$\Delta Q_{усл2} = Q_1 - Q_{усл1} = 7056 - 8330 = -1274(тыс.руб.)$$

Суммарное влияние двух факторов:
ΔQ= -1274-2720= -3994.

Таким образом, на изменение объема товарной продукции отрицательное влияние оказало изменение на 32 человека численности работников, что вызвало снижение объема продукции на 2720 тыс. руб. и отрицательное влияние оказало снижение выработки на 13 тыс. руб., что вызвало снижение объема на 1274 тыс. руб. Суммарное влияние двух факторов привело к снижению объема продукции на 3994 тыс. руб.

Способ абсолютных разниц является модификацией способа цепной подстановки. Изменение результативного показателя за счет каждого фактора способом разниц определяется как произведение отклонения изучаемого фактора на базисное или отчетное значение другого фактора в зависимости от выбранной последовательности подстановки:

$$y_0 = a_0 \cdot b_0 \cdot c_0;$$
$$\Delta y_a = \Delta a \cdot b_0 \cdot c_0;$$
$$\Delta y_b = \Delta b \cdot a_1 \cdot c_0;$$
$$\Delta y_c = \Delta c \cdot a_1 \cdot b_1;$$
$$y_1 = a_1 \cdot b_1 \cdot c_1;$$
$$\Delta y = \Delta y_a + \Delta y_b + \Delta y_c.$$

Способ относительных разниц применяется для измерения влияния факторов на прирост результативного показателя в мультипликативных и смешанных моделях вида y = (a – b) * c. Он используется в случаях, когда исходные данные содержат

определенные ранее относительные отклонения факторных показателей в процентах.

Для мультипликативных моделей типа y = abc методика анализа следующая: находят относительное отклонение каждого факторного показателя:

$$\Delta a\% = \frac{a_{ф} - a_{пл}}{a_{пл}} \cdot 100\%;$$

$$\Delta b\% = \frac{b_{ф} - b_{пл}}{b_{пл}} \cdot 100\%;$$

$$\Delta c\% = \frac{c_{ф} - c_{пл}}{c_{пл}} \cdot 100\%,$$

затем определяют отклонение результативного показателя ***y*** за счет каждого фактора

$$\Delta y_a = \frac{y_{пл} \cdot \Delta a\%}{100};$$

$$\Delta y_b = \frac{(y_{пл} + \Delta y_a)\Delta b\%}{100};$$

$$\Delta y = \frac{(y_{пл} + \Delta y_a + \Delta y_b) \cdot \Delta c\%}{100}.$$

Пример. Воспользовавшись данными табл. 21, проведем анализ способом относительных разниц. Относительные отклонения рассматриваемых факторов составят:

$$\Delta X = \frac{X_1 - X_0}{X_0} \cdot 100\% = \frac{98 - 130}{130} \cdot 100\% = -24{,}61\% ;$$

$$\Delta \mathrm{Pr} = \frac{\mathrm{Pr}_1 - \mathrm{Pr}_0}{\mathrm{Pr}_0} \cdot 100\% = \frac{72 - 85}{85} \cdot 100\% = -15{,}29\% .$$

Рассчитаем влияние на объем товарной продукции каждого фактора.

Количества работников:

$$\Delta Q_{усл1} = \frac{11050 \cdot (-24{,}61\%)}{100} = -2720(тыс.руб.).$$

Выработки продукции каждым работником:

$$\Delta Q_{усл2} = \frac{(11050 - 2720) \cdot (-15{,}29\%)}{100} = -1274(тыс.руб.)$$

Контрольные вопросы

1. В каких случаях применяется метод факторного анализа?
2. Объясните способы построения детерминированных факторных моделей.
3. Опишите алгоритм применения детерминированного факторного анализа: способа цепных подстановок.
4. Опишите алгоритм применения способа относительных разниц.
5. Приведите примеры задач и факторных моделей, к которым применяется каждый из методов детерминированного факторного анализа.

Линейный регрессионный анализ

Цель: *Используя методы регрессионного анализа, научиться строить прямые регрессии и оценивать полученные данные (прогноз) в заданном доверительном интервале.*

Корреляционный анализ позволяет установить степень взаимосвязи двух и более случайных величин. Однако наряду с этим желательно иметь модель этой связи, которая дала бы возможность предсказывать значения одной случайной величины по конкретным значениям другой. Методы решения подобных задач носят наименование ***регрессионный анализ***.

В линейный регрессионный анализ входит широкий круг задач, связанных с построением (восстановлением) зависимостей между группами числовых переменных

$$X = (x_1, ..., x_p) \text{ и } Y = (y_1, ..., y_m).$$

Предполагается, что X – независимые переменные (факторы, объясняющие переменные) влияют на значения Y – зависимых переменных (откликов, объясняемых переменных). По имеющимся эмпирическим данным
(X_i, Y_i), $i = 1, ..., n$ требуется построить функцию $f(X)$, которая приближенно описывала бы изменение Y при изменении X

Рассмотрим простой случай двух коррелированных случайных величин $\boldsymbol{x}$ и $\boldsymbol{y}$. Линейная связь между двумя случайными величинами означает, что прогноз значения величины $\boldsymbol{y}$ по данному значению $\boldsymbol{x}$ имеет вид

$$\tilde{y} = A + Bx,$$

где A и B – это соответственно отрезок оси ординат, отсекаемой прямой, и ее наклон. Если данные связаны идеальной линейной зависимостью (функциональная или сильная связь – в других терминах) ($r_{xy} = 1$), то предсказанное значение будет в точности

равняться наблюденному значению y_i при любом данном x_i. Однако на практике обычно отсутствует идеальная линейная зависимость между данными. Как правило, внешние случайные воздействия приводят к разбросу данных, и, кроме того, возможны искажения за счет присутствия нелинейных эффектов. Тем не менее, если все же предположить существование линейной связи и наличие неограниченной выборки, то можно подобрать такие значения A и B, которые дадут возможность предсказать ожидаемое значение y_i для любого данного x_i. Это означает, что не обязательно совпадает с наблюдаемым значением y_i, соответствующим данному x_i, однако оно будет равно среднему значению всех таких наблюдаемых значений.

Метод наименьших квадратов

Общепринятая процедура определения коэффициентов уравнения состоит в выборе таких значений A и B, которые минимизируют сумму квадратов отклонений наблюдаемых значений от предсказанного значения y. Эта процедура называется ***методом наименьших квадратов***. Поскольку отклонения наблюдаемых значений от предсказанных равны

$$y_i - \tilde{y} = y_i - (A + Bx_i),$$

то сумма квадратов отклонений имеет вид

$$\boldsymbol{Q = \sum_{i=1}^{n} (y_i - A - Bx_i)^2}.$$

Следовательно, наилучшее согласие в смысле наименьших квадратов обеспечивают значения $\boldsymbol{A}$ и $\boldsymbol{B}$, для которых частные производные равны нулю:

$$\frac{\partial Q}{\partial A} = \frac{\partial Q}{\partial B} = 0.$$

Частные производные по коэффициентам A и B, так как они являются не константами в общем смысле, а некоторыми переменными величинами.

На практике обычно имеется ограниченная выборка из N пар наблюдений значений x и y. Это означает, что уравнение $\frac{\partial Q}{\partial A} = \frac{\partial Q}{\partial B} = 0$ даст всего лишь оценки A и B; обозначим их через a и b соответственно. Для отыскания минимума приравняем к нулю частные производные:

$$\frac{\partial Q}{\partial a} = 2\sum_{i}^{n}(a \cdot x_i + b - y_i) \cdot x_i = 0;$$

$$\frac{\partial Q}{\partial b} = 2\sum_{i}^{n}(ax_i + b - y_i) = 0.$$

Решая систему уравнений относительно оценок величин A и B, получим:

$$(\sum x^2) \cdot a + (\sum x)b = \sum xy$$

$$(\sum x)a + nb = \sum y$$

$$a = \frac{n\sum xy - \sum x \cdot \sum y}{n \cdot \sum x^2 - (\sum x)^2}$$

$$b = \frac{\sum x^2 \cdot \sum y - \sum x \cdot \sum xy}{n\sum x^2 - (\sum x)^2}$$

Оценки A и B можно также подсчитать по формулам:

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

или в обозначениях коэффициента корреляции и выборочных дисперсий:

$$b = \frac{s_{xy}}{s_x^2} = r_{xy} \cdot \frac{s_y}{s_x},$$

$$a = \bar{y} - b\bar{x}.$$

Эти значения можно использовать для построения регрессионной модели, позволяющей предсказывать неизвестное ***y*** по заданному ***x***:

$$\hat{y} = a + bx = (\bar{y} - b\bar{x}) + bx = \bar{y} + b(x - \bar{x}).$$

Прямая линия, задаваемая указанным уравнением называется прямой линейной регрессии у на х. Ясно, что коэффициенты *a* и *b*, определенные формулами, приведенными выше, являются случайными функциями, имеющими свои распределения. Следовательно, необходимо оценить, насколько точно (или по-другому – какой разброс) значения *a* и *b* мы получим.

Доверительные интервалы

Для анализа экспериментальных данных часто рассчитываются значения числовых характеристик случайных величин. Это способ служит для оценки параметров и дает их **точечные оценки**. Числовые характеристики выборки не позволяют судить о степени близости к соответствующим параметрам генеральной совокупности, поскольку каждый вариант выборки дает свои значения параметров. Более содержательны процедуры оценивания параметров, связанные не с получением точечного значения, а с построением интервала, который накрывает оцениваемый параметр с известной степенью достоверности.

Пусть, например, выборочное среднее арифметическое, вычисленное по ***n*** независимым наблюдениям случайной величины

x, используется в качестве оценки среднего μ_x. Обычно представляет интерес оценить μ_x в терминах некоторого интервала $\bar{x} \pm d$, в который μ_x попадает с заданной **степенью достоверности**. Такие интервалы можно построить, если известны выборочные распределения рассматриваемой оценки.

Относительно значения выборочного среднего можно сделать следующее вероятностное утверждение:

$$P\left\{\Phi(1-\frac{\alpha}{2}) < \frac{(\bar{x}-\mu_x)\sqrt{n}}{\sigma_x} \le \Phi\left(\frac{\alpha}{2}\right)\right\} = 1-\alpha,$$

где ***Φ(x)*** – стандартная функция распределения; где ***α*** – называется **вероятностью ошибки**, или **уровнем значимости**. Обычно вероятность ошибки измеряется в пределах от 0,10 до 0,0001 или в процентах от 1 %, 5 % или 10 % .

Значение $S = 1 - \alpha$ – статистическая достоверность**.** S измеряют часто в процентах и говорят, например, о 95%-м доверительном интервале ($S = (1 - \alpha)\cdot 100\%$).

По мере уменьшения α (увеличения интервала, заключенного между Φ(1-α/2) и Φ(α/2)) разумно считать, что вероятность P скорее равна единице, чем нулю. Иначе говоря, если производится много выборок, и для каждой из них вычисляется , то можно ожидать, что она будет попадать в указанный интервал с относительной частотой, примерно равной 1 – α. При таком подходе можно утверждать, что существует интервал, в который величина $\frac{(\bar{x}-\mu_x)\sqrt{n}}{\sigma_x}$ попадает с большой степенью достоверности. Такие утверждения называют доверительными. Интервал, относительно которого делается доверительное утверждение,

называется доверительным интервалом. Степень доверия, сопоставляемая доверительному утверждению, называется уровнем доверия.

При оценивании среднего значения доверительный интервал для среднего μ_x можно построить по выборочному значению $\bar{x}$:

$$\left[\bar{x} - \frac{\sigma_x \Phi(\alpha/2)}{\sqrt{n}} \le \mu_x \le \bar{x} + \frac{\sigma_x \Phi(\alpha/2)}{\sqrt{n}}\right].$$

Если σ_x — неизвестна, то доверительный интервал для μ_x можно построить по выборочным значениям $\bar{x}$ и s (среднеквадратичное отклонение для выборки). В этом случае используется t-распределение Стьюдента:

$$\left[\bar{x} - \frac{st(n,\alpha/2)}{\sqrt{N}} \le \mu_x < \bar{x} + \frac{st(n,\alpha/2)}{\sqrt{N}}\right],$$

где n = N-1 — степени свободы для распределения Стьюдента, N — выборка. Интервалам соответствует уровень доверия 1- α.

Точность оценки параметров линии регрессии

Точность оценок параметров a, b и значения предсказанной функции $\hat{y}$ мы определим в предположении о нормальности распределения y при данном значении x (рис. 26).

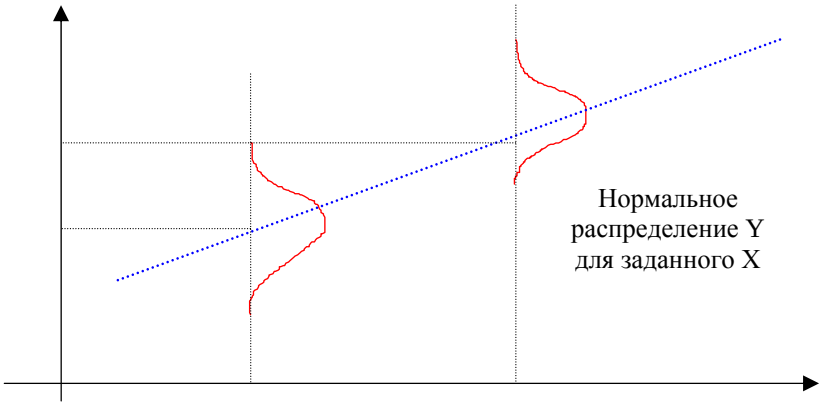


Рис. 26. Графическое представление линии регрессии

Выборочные распределения указанных параметров связаны с ***t***-распределением соотношениями:

$$A = a \pm s_{y|x} t_{N-2} \cdot \sqrt{\frac{1}{N} + \frac{\bar{x}^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2}};$$

$$B = b \pm s_{y|x} t_{N-2} \cdot \sqrt{\frac{1}{\sum_{i=1}^{N} (x_i - \bar{x})^2}}.$$

Распределение $\hat{y}$ при конкретном значении $\boldsymbol{x = x_0}$ представляет особый интерес ($\hat{y}$ -значение, вычисленное по уравнению регрессии, $\tilde{y}$ - оценочное интервальное значение):

$$\Delta y = s_{y|x} t_{N-2} \cdot \sqrt{\frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2}}$$

$$\tilde{y} = \hat{y} \pm s_{y|x} t_{N-2} \cdot \sqrt{\frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2}}.$$

Значение будет определять границы интервала для заданного значения α. В формулах величина $s_{y|x}$ - выборочное стандартное отклонение наблюдаемого значения y_i от предсказанного $\hat{y}_i = a + bx_i$, равное:

$$s_{y|x} = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N-2}} = \sqrt{\left(\frac{n-1}{n-2}\right) s_y^2 \left(1 - r_{xy}^2\right)}.$$

Множественный регрессионный анализ

Цель: *используя множественную регрессионную модель, научиться строить регрессионную зависимость.*

Общее назначение множественной регрессии (этот термин был впервые использован в работе К. Пирсона – Pearson, 1908) состоит в анализе связи между несколькими независимыми переменными (называемыми также регрессорами или предикторами) и зависимой переменной. Например, агент по продаже недвижимости мог бы вносить в каждый элемент реестра размер дома (в квадратных футах), число спален, средний доход населения в этом районе в соответствии с данными переписи и субъективную оценку привлекательности дома. Как только эта информация собрана для различных домов, было бы интересно посмотреть, связаны ли и каким образом эти

характеристики дома с ценой, по которой он был продан. Например, могло бы оказаться, что число спальных комнат является лучшим предсказывающим фактором (предиктором) для цены продажи дома в некотором специфическом районе, чем "привлекательность" дома (субъективная оценка). Могли бы также обнаружиться и "выбросы", т.е. дома, которые могли бы быть проданы дороже, учитывая их расположение и характеристики.

Как только эта так называемая линия регрессии определена, аналитик оказывается в состоянии построить график ожидаемой (предсказанной) оплаты труда и реальных обязательств компании по выплате жалования. Таким образом, аналитик может определить, какие позиции недооценены (лежат ниже линии регрессии), какие оплачиваются слишком высоко (лежат выше линии регрессии), а какие оплачены адекватно.

В общественных и естественных науках процедуры множественной регрессии чрезвычайно широко используются в исследованиях. В общем, множественная регрессия позволяет исследователю задать вопрос (и, вероятно, получить ответ) о том, "что является лучшей посылкой для...". Например, исследователь в области образования мог бы пожелать узнать, какие факторы являются лучшими условиями успешной учебы в средней школе. А психолога мог бы заинтересовать вопрос, какие индивидуальные качества позволяют лучше предсказать степень социальной адаптации индивида. Социологи, вероятно, хотели бы найти те социальные индикаторы, которые лучше других предсказывают результат адаптации новой иммигрантской группы и степень ее слияния с обществом. Заметим, что термин "множественная" указывает на наличие

нескольких предикторов или регрессоров, которые используются в модели.

Множественная корреляция имеет второе название — ***множественное предсказание***. Цель множественного предсказания — оценивание зависимой переменной ***Y*** по линейной (или нелинейной) комбинации ***m*** независимых переменных $X_1, X_2, ..., X_m$.

Термин "множественная регрессия" объясняется тем, что анализу подвергается зависимость одного признака (результирующего) от набора независимых (факторных) признаков. Разделение признаков на результирующий и факторные осуществляется исследователем на основе содержательных представлений об изучаемом явлении (процессе). Все признаки должны быть количественными (хотя допускается и использование дихотомических признаков, принимающих лишь два значения, например 0 и 1). Множественная регрессия применяется в ситуациях, когда из множества факторов, влияющих на результативный признак, нельзя выделить один доминирующий фактор и необходимо учитывать влияние нескольких факторов.

Основная цель множественной регрессии – построить модель с большим числом факторов, определив при этом влияние каждого из них в отдельности, а также совокупное их воздействие на моделируемый показатель.

Различают линейные и нелинейные регрессии.

Линейная регрессия описывается уравнением:

$$y = a + bx + e .$$

Нелинейные регрессии делятся на два класса: регрессии, нелинейные относительно включенных в анализ объясняющих

переменных, но линейные по оцениваемым параметрам, и регрессии, нелинейные по оцениваемым параметрам.

Примеры регрессий, нелинейных по объясняющим переменным, но линейных по оцениваемым параметрам:

· полиномы разных степеней $\hat{y}_x = a + b_1 x + b_2 x^2 + b_3 x^3 + \varepsilon;$

· равносторонняя гипербола $\hat{y}_x = a + \frac{b}{x} + \varepsilon;$

Примеры нелинейных регрессий, по оцениваемым параметрам:

· степенная $\hat{y}_x = a x^b \varepsilon;$

· показательная $\hat{y}_x = a b^x \varepsilon;$

· экспоненциальная $\hat{y}_x = e^{a+bx} \varepsilon.$

Наиболее часто применяются следующие модели регрессий:

- прямой $\hat{y}_x = a + bx;$
- гиперболы $\hat{y}_x = a + \frac{b}{x};$
- параболы $\hat{y}_x = a + bx + cx^2;$
- показательной функции $\hat{y}_x = a b^x;$
- степенной функции $\hat{y}_x = a x^b$ и др.

Чаще всего исследователи ограничиваются линейной регрессией, т.е. зависимостью вида:

$$Y = a + b_1 x_1 + b_2 x_2 + ... + b_m x_m + \varepsilon, \qquad (1)$$

где Y – результирующий признак; x_1, …, x_m – факторные признаки; b_1,…,b_m – коэффициенты регрессии; a – свободный член уравнения; ε – "ошибка" модели.

Это уравнение представляет собой многомерное предсказание переменной Y по аналогии с одномерным случаем. Уравнение (1) называется линейным, поскольку b-коэффициенты входят туда в первой степени. Уравнение (1) само по себе не представляет особой ценности; должна быть установлена процедура, посредством которой для $\boldsymbol{b_i}$ выбирают “хорошие” (близкие к достоверным, наиболее вероятные) значения.

Как и в случае парной регрессии, построение уравнения множественной регрессии осуществляется в два этапа:

- **определение модели;**
- **оценка параметров выбранной модели.**

Определение модели включает в себя решение двух задач:

1) отбор *p* факторов x_j, наиболее влияющих на величину *y*;
2) выбор вида уравнения регрессии $\hat{y} = f(x_1,x_2,...,x_p)$.

Включение в уравнение множественной регрессии того или иного набора факторов связано, прежде всего, с представлением исследователя о природе взаимосвязи моделируемого показателя с другими экономическими явлениями.

1) Факторы, включаемые во множественную регрессию, должны отвечать следующим требованиям.

2) Факторы должны быть количественными. Добавление в модель качественного фактора, требует присвоения ему количественного значения (например, в модели урожайности качество почвы задается в виде баллов; в модели стоимости недвижимости районам присваиваются ранги).

3) Число включаемых факторов должно быть в 6–7 раз меньше объема совокупности, по которой строится регрессия.

Факторы не должны быть взаимозависимы. Если между факторами существует высокая корреляция, то нельзя определить их изолированное влияние на результат, и параметры уравнения регрессии не будут адекватно интерпретироваться.

Включаемые во множественную регрессию факторы должны объяснить вариацию независимой переменной. Если строится модель с набором из p факторов, то для нее рассчитывается показатель детерминации R^2, который фиксирует долю объясненной вариации результативного признака за счет рассматриваемых в регрессии p факторов. Влияние других, не учтенных в модели, факторов оценивается как $1 - R^2$ с соответствующей остаточной дисперсией S^2.

При дополнительном включении в регрессию (p + 1) – фактора x_{p+1} коэффициент детерминации должен возрастать, а остаточная дисперсия уменьшаться, т. е.

$$R^2_{p+1} >= R^2_{p}; S^2_{p+1} <= S^2_{p.}$$

Если же этого не происходит и данные показатели практически мало отличаются друг от друга, то включаемый в анализ фактор x_{p+1} не улучшает модель и является лишним. Насыщение модели лишними факторами не только не снижает величину остаточной дисперсии и не увеличивает показатель детерминации, но и приводит к статистической незначимости параметров регрессии по *t*-критерию Стьюдента.

Отбор факторов производится на основе качественного анализа и обычно осуществляется в две стадии:

- на первой подбираются факторы исходя из сущности проблемы;

- на второй – на основе матрицы показателей корреляции определяют *t*-статистики для параметров регрессии.

Коэффициенты корреляции между объясняющими переменными позволяют исключать из модели дублирующие факторы. Считается, что две переменные находятся между собой в линейной зависимости, если $r_{x_i x_j} \geq 0{,}7.$

Если факторы явно коллинеарны, то они дублируют друг друга и один из них нужно исключить из регрессии. Предпочтение отдается тому фактору, который при достаточно тесной связи с результатом имеет наименьшую тесноту связи с другими факторами.

Пусть, например, при изучении зависимости *y = f (*x, *z, v)* матрица парных коэффициентов корреляции оказалась следующей:

	Y	X	Q	Z
Y	1			
X	0.65	1		
Q	0.7	0.4	1	
Z	0.85	0.35	0.7	1

Очевидно, что факторы q и z дублируют друг друга, поскольку связь между ними r_{zq}=0.7. В анализ целесообразно включить фактор z, а не q, поскольку корреляция z с результатом y сильнее, чем корреляция фактора q и y, и слабее межфакторная корреляция между *z* и x ($r_{zx} < r_{qx}$). Поэтому в данном случае в уравнение множественной регрессии включаются факторы z и x.

Выбор формы уравнения регрессии

Как и в парной зависимости, возможны разные виды уравнений множественной регрессии: линейные и нелинейные. Ввиду четкой интерпретации параметров наиболее широко используются линейная и степенная функции.

В уравнении линейной множественной регрессии

$$Y = a + b_1 x_1 + b_2 x_2 \ + ... + b_m x_m + \varepsilon;$$

параметры при x*i* называются коэффициентами «чистой» регрессии. Они характеризуют среднее изменение результата с изменением соответствующего фактора на единицу при неизменном значении других факторов, закрепленных на среднем уровне.

Предположим, например, что зависимость расходов на продукты питания по совокупности семей характеризуется следующим уравнением:

$$\hat{y}_x = 0{,}45 + 0{,}33x_1 + 0{,}42x_2 + 0{,}25x_3,$$

где у – расходы семьи за месяц, тыс. руб.;

x_1 – месячный доход на одного члена семьи, тыс. руб.;

x_2– размер семьи, человек;

x_3 – коммунальные платежи на одного члена семьи, тыс. руб.

Анализ данного уравнения позволяет сделать выводы – с ростом дохода на одного члена семьи на 1 тыс. руб. расходы на питание возрастут в среднем на 330 руб. при том же среднем размере семьи. Иными словами, 33 % дополнительных семейных расходов тратится на питание. Увеличение размера семьи при тех же ее доходах предполагает дополнительный рост расходов на 420 руб. Увеличение коммунальных платежей при тех же ее доходах предполагает дополнительный рост расходов на 250 руб.

Оценка параметров уравнения множественной регрессии

Для оценки параметров уравнения множественной регрессии применяют метод наименьших квадратов (МНК). Для линейных уравнений регрессии (и нелинейных уравнений, приводимых к линейным) строится система нормальных уравнений, решение которой позволяет получить оценки параметров регрессии. В случае линейной множественной регрессии

$$y = a + b_1 x_1 + b_2 x_2 \ + ... + b_p x_p,$$

система нормальных уравнений имеет следующий вид:

$$\sum y = na + b_1 \sum \ x_1 + b_2 \sum \ x_2 \ + ... + b_p \sum \ x_p;$$
$$\sum yx_1 = a \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_2 x_1 + .. + b_p \sum x_p x_1;$$
$$.....$$
$$\sum yx_p = a \sum x_p + b_1 \sum x_1 x_p + b_2 \sum x_2 x_p + .. + b_p \sum x_2^2.$$

Для определения значимости факторов и повышения точности результата используется **уравнение множественной регрессии в стандартизованном масштабе:**

$$t_y = \beta_1 t_{x1} + \beta_2 t_{x2} + ... + \beta_p t_{xp} + \varepsilon,$$

где t_y, t_{x1}, t_{xp} – **стандартизованные переменные,** рассчитываемые по формулам:

$$t_y = \frac{y - \overline{y}}{\sigma_y},$$
$$t_{xi} = \frac{x_i - \overline{x_i}}{\sigma_{xi}},$$

для которых среднее значение равно нулю $\overline{t_y} = \overline{t_{xi}} = 0,$ а среднее квадратическое отклонение равно единице $\sigma_{t_y} = \sigma_{t_{xi}} = 1.$

Величины βi называются стандартизованными коэффициентами регрессии. К уравнению множественной регрессии в стандартизованном масштабе применим МНК. Стандартизованные коэффициенты регрессии (β-коэффициенты) определяются из следующей системы уравнений:

$$\sum t_y t_{x1} = \beta_1 \sum t_{x1}^2 + \beta_2 \sum t_{x1} t_{x2} + \beta_3 \sum t_{x1} t_{x3} + ... + \beta_p \sum t_{x1} t_{xp};$$
$$\sum t_y t_{x2} = \beta_1 \sum t_{x1} t_{x2} + \beta_2 \sum t_{x2}^2 + \beta_3 \sum t_{x2} t_{x3} + .. + \beta_p \sum t_{x2} t_{xp};$$
$$\sum t_y t_{xp} = \beta_1 \sum t_{x1} t_{xp} + \beta_2 \sum t_{x2} t_{xp} + \beta_3 \sum t_{x3} t_{xp} + .. + \beta_p \sum t_{xp}^2.$$

либо из системы уравнений

$$r_{yx1} = \beta_1 + \beta_2 r_{x2x1} + \beta_3 r_{x3x1} + ... + \beta_p r_{xpx1};$$
$$r_{yx2} = \beta_1 r_{x1x2} + \beta_2 + \beta_3 r_{x3x2} + ... + \beta_p r_{xpx2};$$
$$.....$$
$$r_{yxp} = \beta_1 r_{x1xp} + \beta_2 r_{x2xp} + \beta_3 r_{x3xp} + ... + \beta_p.$$

Стандартизованные коэффициенты регрессии показывают, на сколько сигм (средних квадратических отклонений) изменится в среднем результат, если соответствующий фактор x*i* изменится на одну сигму при неизменном среднем уровне других факторов. В силу того, что все переменные заданы как центрированные и нормированные, стандартизованные коэффициенты регрессии βi сравнимы между собой. Сравнивая их друг с другом, можно ранжировать факторы по силе их воздействия на результат. В этом основное достоинство стандартизованных ко-

эффициентов регрессии в отличие от коэффициентов «чистой» регрессии, которые несравнимы между собой.

В парной зависимости стандартизованный коэффициент регрессии β есть не что иное, как линейный коэффициент корреляции r_{yx}.

Связь коэффициентов множественной регрессии *bi* со стандартизованными коэффициентами β*i* описывается соотношением

$$b_i = \beta_i \cdot \frac{\delta_y}{\delta_{xi}}.$$

Параметр *a* определяется из соотношения:

$$a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 \ - .... - b_p\bar{x}_p.$$

Средние коэффициенты эластичности для линейной множественной регрессии рассчитываются по формуле

$$\overline{Э}_{yxj} = b_j \cdot \frac{\overline{x}_j}{\overline{y}}.$$

и показывают, на сколько процентов в среднем по совокупности изменится результат у от своей величины при изменении фактора х на 1 % от своего значения при неизменных значениях других факторов.

Предположим, что по ряду регионов множественная регрессия величины импорта на определенный товар относительно отечественного его производства x_1, изменения запасов x_2 и потребления на внутреннем рынке x_3 оказалась следующей

$$\hat{y} = -66{,}028 + 0{,}135x_1 + 0{,}476x_2 \ + 0{,}343x_3.$$

При этом средние значения для рассматриваемых признаков составили:

$$\bar{y} = 31{,}5; \bar{x}_1 = 247{,}7; \bar{x}_2 = 3{,}7; \bar{x}_3 = 182{,}5.$$

На основе данной информации могут быть найдены средние по совокупности показатели эластичности. Для данного примера они окажутся равными:

$$\overline{Э}_{yx_1} = 0{,}135 \cdot \frac{247{,}7}{31{,}5} = 1{,}06\ \%,$$

$$\overline{Э}_{yx_2} = 0{,}476 \cdot \frac{3{,}7}{31{,}5} = 0{,}056\ \%,$$

$$\overline{Э}_{yx_3} = 0{,}343 \cdot \frac{182{,}5}{31{,}5} = 1{,}987\ \%.$$

1) С ростом величины отечественного производства на 1 % размер импорта в среднем по совокупности регионов возрастет на 1,06 % при неизменных запасах и потреблении семей; 2) с ростом изменения запасов на 1 % при неизменном производстве и внутреннем потреблении величина импорта увеличивается в среднем на 0,056 %; 3) при неизменном объеме производства и величины запасов с увеличением внутреннего потребления на 1 % импорт товара возрастает в среднем по совокупности регионов на 1,987 %.

Средние показатели эластичности можно сравнивать друг с другом и соответственно ранжировать факторы по силе их воздействия на результат. В рассматриваемом примере наибольшее воздействие на величину импорта оказывает размер внутреннего потребления товара x_3 , а наименьшее – изменение запасов x_2.

Контрольные вопросы

1. Каковы области применения множественного регрессионного анализа?

2. В чем заключается метод наименьших квадратов? Каковы основные условия его применения?
3. Назовите основные условия применения корреляционно-регрессионного метода анализа статистических связей.
4. Приведите примеры различных видов уравнений парной и множественной регрессии.
5. Дайте определение парному и множественному линейным коэффициентам корреляции.
6. Каковы свойства множественного коэффициента корреляции?

Предельные теоремы теории вероятностей

Цель: *Познакомить студентов с законом больших чисел и областями его применения.*

Существует много явлений и ситуаций, когда при проведении подобных испытаний многократно наблюдается одна и та же случайная величина. Практика изучения случайных явлений показывает, что хотя результаты отдельных наблюдений, даже проведенных в одинаковых условиях, могут сильно отличаться, в то же время средние результаты для достаточно большого числа наблюдений устойчивы и слабо зависят от результатов отдельных наблюдений.

Устойчивость испытаний состоит в том, что особенности каждого отдельного случайного явления почти не сказываются на среднем результате большой массы подобных явлений, а характеристики случайных событий и случайных величин, наблюдаемых в испытаниях, при неограниченном увеличении числа испытаний становятся практически не случайными.

Теоретическим обоснованием этого замечательного свойства случайных явлений является **закон больших чисел.** Теоремы закона больших чисел устанавливают зависимость между случайностью и необходимостью. Названием "закон больших чисел" объединена группа теорем, устанавливающих устойчивость средних результатов большого количества случайных явлений и объясняющих причину этой устойчивости.

Простейшая форма закона больших чисел и исторически первая теорема этого раздела – теорема Бернулли. Эта теорема устанавливает связь между вероятностью появления события и его относительной частотой появления и позволяет при этом предсказать, какой примерно будет эта частота в *n* испытаниях.

Теорема Бернулли: если вероятность события A одинакова во всех испытаниях и равна P(A), то при достаточно большом n для произвольного e >0 справедливо неравенство:

$$P\left(\left|\frac{m}{n}-p\right|\le\varepsilon\right)>1-\frac{pq}{n\varepsilon^2},$$

а при переходе к пределу получаем:

$$\lim_{n\to\infty}P(|w-p|<\varepsilon)=1.$$

Из теоремы видно, что с увеличением числа испытаний W – частота события A стремится к вероятности события P(A) и перестает быть случайной.

Иногда (при решении практических задач) требуется оценить вероятность того, что отклонение количества благоприятных исходов испытания m в общем числе испытаний

n от ожидаемого результата np не превысит определенного числа ε. Для данной оценки неравенство переписывают в виде:

$$P\left(\left|m-np\right|\le\varepsilon\right)>1-\frac{pq}{\varepsilon^2}$$

Пример. Монету подбрасывают 1000 раз. Оценить вероятность отклонения частоты появления герба от вероятности его появления меньше чем на ε=0,1.

Решение: вероятность появления герба p=0,5, тогда q = 1-0,5=0,5;

n= 1000, ε = 0,1.

$$P\left(\left|\frac{m}{1000}-0{,}5\right|\le 0{,}1\right)>1-\frac{0{,}5\cdot 0{,}5}{1000\cdot 0{,}1^2}$$

$$\left|\frac{m}{1000}-0{,}5\right|<0{,}1 \qquad\qquad \left|\frac{m}{1000}-0{,}5\right|<0{,}1$$

$$\frac{m}{1000}-05<0{,}1 \qquad\qquad -\frac{m}{1000}+05<0{,}1$$

$$\frac{m}{1000}<0{,}6 \qquad\qquad \frac{m}{1000}>0{,}4$$

$$m<600 \qquad\qquad m>400$$

$$P\left(400<m<600\right)>1-\frac{0{,}5\cdot 0{,}5}{1000\cdot 0{,}1^2}$$

$$P\left(400<m<600\right)>0{,}975$$

Раскрывая модуль и решая неравенство относительно m, получим: 400<m<600. Итак, вероятность небольшого отклонения частоты выпадения герба (±100) от его классической

вероятности 0,5 равна 0,975. Значит, вероятность большего отклонения крайне мала и равна 0,025.

Теорема Пуассона утверждает, что частота события в серии независимых испытаний стремится к среднему арифметическому его вероятностей и перестает быть случайной.

$$W(A) \to \overline{p} = \frac{p_1 + p_2 + ... + p_n}{n}$$

При большом количестве испытаний вычисления по формуле Бернулли становятся затруднительными. Однако в ряде случаев их можно заменить более простыми асимптотическими формулами, например формулой Пуассона (когда npq<9). Если производится n независимых опытов и вероятность появления события в каждом опыте равна p_i, то при увеличении n, 0 частота события n*p= λ , где λ >0 и стремится к среднему арифметическому вероятностей p_i события А.

Пример. В здании 1000 лампочек. Вероятность выхода из строя одной лампочки в течение года p =0.003. Найдем вероятность того, что в течение одного года выйдет из строя более трех ламп. Выполним вычисления используя формулу Бернулли и по теореме Пуассона.

Решение. Для вычисления вероятности используем формулу Пуассона: $P_{n,p}(k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}$

P(m > 3) = 1- P(m <= 3) ,

Параметр лямбда считаем по формуле:

λ = np = 1000 * 0.003=3;

$$P_{n,p}(0) = \frac{3^0 \cdot e^{-3}}{0!} \approx 0{,}05$$

$$P_{n,p}(1) = \frac{3^1 \cdot e^{-3}}{1!} \approx 0{,}14$$

$$P_{n,p}(2) = \frac{3^2 \cdot e^{-3}}{2!} \approx 0{,}22$$

$$P_{n,p}(3) = \frac{3^3 \cdot e^{-3}}{3!} \approx 0{,}22$$

$$P_{n,p}(k > 3) = 1 - (0{,}05 + 0{,}14 + 0{,}22 + 0{,}22) = 0{,}37$$

Ответ: вероятность того, что в течение года выйдет из строя больше трех лампочек 0,37.

Предельные теоремы теории вероятностей объясняют природу устойчивости частоты появлений события. Природа эта состоит в том, что предельным распределением числа появлений события при неограниченном возрастании числа испытаний (если вероятность события во всех испытаниях одинакова) является *нормальное распределение.*

Центральная предельная теорема объясняет широкое распространение нормального закона распределения. Теорема утверждает, что всегда, когда случайная величина образуется в результате сложения большого числа независимых случайных величин с конечными дисперсиями, закон распределения этой случайной величины оказывается практически **нормальным законом.**

Закон больших чисел утверждает, что при большом числе испытаний среднее арифметическое случайной величины стремится к математическому ожиданию и перестает быть случайным.

Теорема Ляпунова объясняет широкое распространение *нормального закона* распределения и поясняет механизм его образования. Теорема позволяет утверждать, что всегда, когда случайная величина образуется в результате сложения большого числа независимых случайных величин, дисперсии которых малы по сравнению с дисперсией суммы, закон распределения этой случайной величины оказывается практически *нормальным* законом. А поскольку случайные величины всегда порождаются бесконечным количеством причин и чаще всего ни одна из них не имеет дисперсии, сравнимой с дисперсией самой случайной величины, то большинство встречающихся в практике случайных величин подчинено нормальному закону распределения.

В основе качественных и количественных утверждений закона больших чисел лежит *неравенство Чебышева*. Оно определяет верхнюю границу вероятности того, что отклонение значения случайной величины от ее математического ожидания больше некоторого заданного числа. Замечательно, что неравенство Чебышева дает оценку вероятности события $|\xi - M\xi| \geq \varepsilon$ для случайной величины, распределение которой неизвестно, известны лишь ее математическое ожидание и дисперсия.

Схема описания всех этих явлений с единых вероятностных позиций выглядит следующим образом: имеется последовательность независимых, одинаково распределенных случайных величин (генеральная совокупность — в терминах математической статистики) и из нее образуется ***среднее арифметическое*** первых ***n*** членов (выборка из генеральной совокупности). Спрашивается, как будет вести себя это среднее арифметическое, если ***n*** велико? Оказывается, что при большом

n оно теряет свойство случайности и приближается к ***математическому ожиданию***. Данный факт мы априорно используем уже давно, перейдя к основным понятиям математической статистики. Этот факт носит название ***закона больших чисел***.

Исторически закон больших чисел доказывается, опираясь на ***неравенство Чебышева***, которое является родоначальником многих других неравенств, широко применяемых в современной теории вероятностей.

Дальнейшее уточнение закона больших чисел происходило в двух направлениях. Первое связано с динамикой поведения средних арифметических. К основным результатам этого направления следует отнести усиленный закон больших чисел и закон повторного логарифма, полученные А.Н.Колмогоровым. Исходным пунктом второго направления, называемого иногда центральной предельной проблемой, являются теоремы Муавра-Лапласа.

Теорема Муавра-Лапласа:

Локальная теорема: Если в схеме Бернулли число испытаний n "велико", то для всех m справедлива приближенная формула (локальная формула Муавра-Лапласа)

$$\sqrt{npq}P_n(m) \approx \varphi(x), \text{ где } x = (m - np) / \sqrt{npq}, \text{ а}$$

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}.$$

Интегральная теорема: Если в схеме Бернулли число испытаний n "велико", то для вероятности $P\{m_1 \le \mu \le m_2\}$ того, что число "успехов" μ заключено в пределах от m_1 до m_2, справедливо приближенное соотношение (интегральная формула Муавра-Лапласа)

$\mathbf{P\{m_1 \le \mu \le m_2\} \approx \Phi(x_2) - \Phi(x_1)}$, где $x_1 = (m_1 - np)/\sqrt{npq}$,

$x_2 = (m_2 - np)/\sqrt{npq}$, а

$\Phi(x) = \int_{-\infty}^{x} \varphi(y)dy = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}}$ – функция стандартного нормального распределения.

Решение центральной предельной проблемы позволило описать класс всех распределений, которые могут выступать в качестве предельных для функций распределения сумм независимых случайных величин в том случае, когда вкладом каждого слагаемого можно пренебречь, найти необходимые и достаточные условия сходимости к каждому распределению этого класса, оценить скорость сходимости (скорость сходимости, как и сама вероятность сходимости ряда наблюдений к некоторой постоянной величине — является важным критерием ***устойчивости*** случайного процесса).

Неравенство Чебышёва

Рассмотрим случайную величину ξ, имеющую дисперсию $D\xi = \sigma^2$. Дисперсия является показателем разброса ξ вокруг математического ожидания $M\xi$. Однако с точки зрения исследователя, разброс естественнее характеризовать вероятностью $P\{|\xi - M\xi| \ge \varepsilon\}$ случайной величины ξ от $M\xi$ на величину, большую некоторого заданного ε. Следующее неравенство позволяет оценить эту вероятность через дисперсию σ^2.

Неравенство Чебышёва. *Для каждой случайной величины* ξ, *имеющей дисперсию* $D\xi = \sigma^2$, *при любом* $\varepsilon > 0$ *справедливо неравенство*

$$P\left\{\left|\xi - M\xi\right| \geq \varepsilon\right\} \leq \frac{\sigma^2}{\varepsilon^2}.$$

Доказательство для непрерывной случайной величины ξ с плотностью распределения ***p(x).*** По определению

$$D\xi = M(\xi - M\xi)^2 = \int_{-\infty}^{\infty} (x - M\xi)^2 p(x)dx.$$

Поскольку подынтегральное выражение неотрицательно, то при уменьшении области интегрирования интеграл может только уменьшиться. Поэтому

$$D\xi \geq \int_{-\infty}^{M\xi-\xi} (x-M\xi)^2 p(x)dx + \int_{M\xi-\xi}^{\infty} (x-M\xi)^2 p(x)dx = \int_{|x-M\xi|\geq\varepsilon} (x-M\xi)^2 p(x)dx.$$

Учитывая теперь, что $(x - Mx)^2 \geq \varepsilon^2$, если $|x - M\xi| \geq \varepsilon$, получаем

$$\int_{|x-M\xi|\geq\varepsilon} (x - M\xi)^2 p(x)dx \geq \varepsilon^2 \cdot \int_{|x-M\xi|\geq\varepsilon} p(x)dx.$$

Последний интеграл представляет собой вероятность события $|\xi - M\xi| \geq \varepsilon$, и, значит $D\xi \geq \varepsilon^2 P\left\{\left|\xi - M\xi\right| \geq \varepsilon\right\}$, откуда и следует неравенство Чебышёва. Аналогично неравенство доказывается и для дискретного случая, при этом нужно только заменить интеграл на сумму.

Ясно, что применять неравенство Чебышёва имеет смысл только тогда, когда $\varepsilon > \sigma$, в противном случае оно дает тривиальную оценку. Неравенство Чебышёва дает ***грубую оценку*** того, что исследуемая величина примет некоторое значение в заданном диапазоне.

Рассмотрим последовательность $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ независимых одинаково распределенных случайных величин (так как случайные величины ξ_i одинаково распределены, то все

их числовые характеристики, в частности математические ожидания и дисперсии, равны между собой). Эта последовательность удовлетворяет (слабому) **закону больших чисел**, если для некоторого **a** и любого $\varepsilon > 0$

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i - a\right| \geq \varepsilon\right\} \xrightarrow[n\to\infty]{} 0.$$

Иными словами, выполнение закона больших чисел отражает предельную устойчивость средних арифметических случайных величин: при большом числе испытаний они практически перестают быть случайными и с большой степенью достоверности могут быть предсказаны.

Иногда вместо выражения "последовательность $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ удовлетворяет закону больших чисел" говорят "*среднее арифметическое случайных величин* $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ ***сходится по вероятности*** *к некоторой предельной постоянной* a".

Теорема (закон больших чисел)

Если последовательность $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ *независимых одинаково распределенных случайных величин такова, что существуют* $M\xi = m$ *и* $D\xi = \sigma^2$, *то для любого* $\varepsilon > 0$

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i - m\right| \geq \varepsilon\right\} \xrightarrow[n\to\infty]{} 0.$$

Доказательство является элементарным следствием неравенства Чебышёва: по свойствам математического ожидания и дисперсии

$$M\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i\right) = m, \qquad D\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i\right) = \sigma^2.$$

Воспользовавшись теперь неравенством Чебышёва, получаем, что для любого $\varepsilon > 0$

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i - m\right| \geq \varepsilon\right\} \leq \frac{\sigma^2}{n\varepsilon^2} \xrightarrow[n\to\infty]{} 0.$$

Таким образом, показано, что для последовательности $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ выполняется закон больших чисел, причем постоянная a совпадает с математическим ожиданием $M\xi_i = m$.

Контрольные вопросы

1. Что такое устойчивость испытаний?
2. Какова простейшая форма закона больших чисел (теорема Бернулли)?
3. О чем гласит теорема Пуассона?
4. В чем суть центральной предельной теоремы Муавра-Лапласа?
5. Приведите закон больших чисел и объясните его смысл.
6. Сформулируйте неравенство Чебышёва.
7. Почему закон больших чисел и центральная предельная теорема занимают центральное место в вероятностно-статистических методах?

Лабораторный практикум

Лабораторная работа № 1

Основы статистической обработки информации

Цель: *Формирование навыков определения числовых характеристик экспериментальных данных. Построение полигона частот выборки.*

Исследователем получены экспериментальные данные, медицинского обследования 100 студентов (табл. 22). Необходимо оценить числовые характеристики выборки студентов, а также проанализировать форму распределения частот.

Таблица 22

Результаты измерения веса студентов

61	57	61	85	48	41	73	66	91	70
50	45	64	46	55	82	69	75	82	72
68	43	81	71	47	50	54	75	81	68
80	67	64	76	61	57	62	57	66	53
79	56	63	88	65	74	67	54	65	80
86	40	59	64	65	71	72	78	70	61
39	63	89	59	61	75	67	51	65	55
62	60	75	73	91	72	54	46	52	55
78	67	94	60	44	49	88	74	44	60
52	61	66	74	56	52	71	73	75	60

1. Используя данные выборки студентов, рассчитать числовые характеристики выборки:
   - среднее арифметическое;
   - медиану;

- моду;
- дисперсию;
- среднее квадратичное отклонение;
- эксцесс;
- асимметрию распределения.

Таблица 23

Образец выполнения работы

Числовые характеристики		Частотный анализ	
max	85,4	Признаки	Частоты
min	54,6	59,4	2
R	30,8	64,2	12
k	6,4	68,9	13
Ср. арифм.	69,30	73,7	13
Медиана	68,53	78,5	4
Мода	68,55	83,3	4
Дисперсия	45,77	88,1	2
Ср. кв. откл	6,76		
Эксцесс	0,057		
Ассиметрия	0,385		

2. Определить распределение выборки по частотам.
   - Найти **min** и **max** значения в выборочной совокупности (с помощью статистических функций Excel).
   - Размах варьирования: $\mathbf{R_x = max - min}$.
   - Число интервалов: k **≈ (1+3,2 lg(n))**, (где n – количество данных в выборке). Количество интервалов следует округлить вверх до целого числа.

- Определить размер одного интервала (цену деления):

$$\mathbf{c = R_x/k.}$$

- Создать массив признаков (интервалов) и посчитать для них частоту. Первый признак определяется как минимальное значение плюс длина одного интервала (цена деления).

3. Построить в Excel гистограмму распределения признаков по частотам и полигон частот. Определить форму распределения выборки.

Рекомендации к выполнению:

Для построения гистограммы и полигона частот используется функция Excel **ЧАСТОТА (массив данных; массив интервалов**). Эта функция относится к классу статистических и производит операции над массивами.

Массив данных — ячейки с данными выборки.

Массив интервалов — ячейки, содержащие значения интервалов.

Результатом выполнения функции ЧАСТОТА является массив, содержащий частоты вариантов, попадающие в указанные интервалы. На основе этого результирующего массива частот и строятся гистограммы и полигоны.

1. Скопировать массив данных из таблицы, расположенной в лабораторной работе.

2. Создать массив интервалов (количество интервалов будет вами рассчитано). Первый интервал определяется как сумма минимального элемента выборки и цена деления, последний элемент не должен существенно превышать максимального элемента выборки.

3. Выделить ячейки под массив частот (пометить доступными способами). Этих ячеек должно быть столько же, сколько ячеек отведено под массив интервалов.

4. Запустить Мастер Функций. Под двоичным массивом здесь понимается массив интервалов**.** Ввести координаты массива данных (вариант) и массива интервалов.

5. После указания всех аргументов функции нажать комбинацию: **Ctrl+Shift+Enter**. После этого функция ЧАСТОТА заполнит весь выделенный массив.

Контрольные вопросы

1. Что называется генеральной совокупностью?
2. Приведите пример генеральной совокупности, исследуемого признака и варианта.
3. Дайте понятие частоты.
4. Что представляет собой полигон частот? Какую информацию можно получить, исследуя полигон частот?
5. Какие формы распределений существуют и чем они отличаются друг от друга? В чем разница между теоретическими и экспериментальными распределениями?
6. Что называется медианой и как ее определяют?
7. Что такое мода?
8. Как определить дисперсию экспериментального распределения?
9. Что характеризует асимметрия выборки?
10. Как рассчитывается эксцесс выборки?
11. При каком значении эксцесса полигон частот наиболее заострен?

Лабораторная работа № 2

Распределения непрерывных случайных величин

Цель: *познакомиться с распределениями непрерывных случайных величин. Сформировать представления о виде функции и плотности непрерывных распределений.*

1. Построить функции и плотности для распределений:

- нормального (табл. 24);
- экспоненциального распределения;
- Вейбулла с параметрами β и λ (табл. 25);
- Гамма-распределения.

Плотности для каждого распределения представить на одном графике. Функции для каждого распределения представить на одном графике. Встроенные функции Excel для искомых распределений: нормальное – НОРМРАСП(); экспоненциальное – ЭКСПРАСП(); распределение Вейбулла – ВЕЙБУЛЛ(); гамма-распределение – ГАММАРАСП().

2. В лабораторной работе №1 были получены полигон частот, среднее арифметическое, дисперсия.

- Построить ***теоретическое*** распределение нормальной плотности, используя значения (min,max, средне арифметическое, стандартное отклонение) из первой лабораторной работы. Первое значение в выборки взять равным минимальному варианту. Шаг изменения Δx_i — выбрать 2,75. Число значений в выборке n должно быть около 20.
- Сравнить внешний вид полученной кривой с полигоном относительных частот.

Таблица 24

Расчет значений функции и плотности нормального распределения

X	39	41,75	44,5	47,25	…	94
F(x)	0,021	0,034	0,054	0,082	…	0,989
P(x)	0,004	0,006	0,009	0,012	…	0,002

Таблица 25

Расчет значений функции и плотности распределения Вейбулла

	λ	β	λ	β	λ	β	λ	β
	0,5	0,5	1	1	1,5	1,5	2	2
X	**F(x)**	**P(x)**	**F(x)**	**P(x)**	**F(x)**	**P(x)**	**F(x)**	**P(x)**
0	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
0,2	0,469	0,840	0,181	0,819	0,048	0,348	0,010	0,099
0,4	0,591	0,457	0,330	0,670	0,129	0,450	0,039	0,192
0,6	0,666	0,305	0,451	0,549	0,224	0,491	0,086	0,274
0,8	0,718	0,223	0,551	0,449	0,323	0,495	0,148	0,341
1	0,757	0,172	0,632	0,368	0,420	0,474	0,221	0,389

Контрольные вопросы

1. Дать определение функции и плотности непрерывной случайной величины.
2. Каковы свойства плотности непрерывной случайной величины.
3. Описать функцию и плотность нормального распределения.
4. Описать функцию и плотность распределения Вейбулла.
5. Сравнить теоретическую плотность нормального распределения и экспериментальную и объяснить различия в их форме.

Лабораторная работа № 3

Выборочные распределения

Цель работы: *познакомиться с различными способами выборки объектов из генеральной совокупности. Оценить репрезентативность различных выборок.*

Для изучения свойств генеральной совокупности исследователи создают выборки, анализируют их числовые характеристики, строят полигоны частот, а затем результаты подобных экспериментов используют на практике.

В настоящее время выделяются следующие типы выборок:

- собственно-случайная;
- механическая;
- типическая;
- серийная.

Собственно-случайная выборка

Простая случайная выборка с возвращением — объект извлекается из генеральной совокупности случайным образом, и, перед извлечением следующего, возвращается обратно.

Например, на заводе, выпускающем гвозди, выполняется проверка длины гвоздя на соответствие определенному размеру. Для проверки с конвейера случайным образом выбираются несколько элементов, затем их измеряют и снова возвращают назад в партию. При следующей операции контроля объекты, попавшие в выборку, могут снова оказаться в ней.

Выборка без возвращения — извлеченный объект не возвращается в генеральную совокупность, а значит, может появиться в выборке только один раз.

Например, производится проверка качества конфет на кондитерской фабрике. Для проверки конфеты выбираются случайным образом, а затем выполняется химический анализ

составляющих компонентов изделия, целостность конфеты при этом нарушается и возврат элемента в совокупность уже невозможен.

Механическая выборка

Механической называется выборка, в которую объекты из генеральной совокупности отбираются через определенный интервал. Если объем выборки должен составлять 5% объема генеральной совокупности, то отбирается каждый двадцатый объект генеральной совокупности.

Недостатком механической выборки является возможность попасть в период периодически изменяющейся случайной величины и выбрать элементы с одинаковыми характеристиками, что будет неадекватно отображать свойства генеральной совокупности.

Типическая выборка

Если генеральную совокупность предварительно разбить на непересекающиеся группы, а затем образовать собственно-случайные выборки из каждой группы и все отобранные объекты считать попавшими в выборку, то получим выборочную совокупность, называемую типической выборкой.

Например, если исследователь изучает успеваемость учащихся восьмых классов в средней школе, а в параллели учатся 5 таких классов, то типической выборкой может быть часть восьмиклассников, например по 7 человек, случайно отобранных из каждого восьмого класса параллели. Объем такой выборки составит 35 человек.

Серийная выборка

Если генеральную совокупность предварительно разбить на непересекающиеся серии, а затем, образовать из них собственно-случайную выборку (выбирается часть серий или одна серия) и все объекты отобранных серий считаются попавшими в серийную выборку.

Например, в продуктовый магазин поступает партия товаров от четырех различных производителей. Продукция каждого производителя может считаться серией. Случайным образом выбираются серии, допустим, первого и третьего производителя и все элементы этих двух серий попадают в выборку

В работе студенты должны создать нормально распределенную генеральную совокупность из 1000 чисел. Затем, необходимо сформировать три выборки объемом по 100 элементов, используя различные методы отбора и оценить репрезентативность различных выборок.

ЗАДАНИЕ

Для генерации нормально распределенной совокупности из 1000 случайных чисел воспользуйтесь специальным инструментом меню Сервис\Анализ данных\Генерация случайных чисел. Чтобы сформировался набор случайных данных с *нормальным* распределением некоторого признака, нужно указать в качестве параметров: Число переменных -1, число случайных чисел -1000, Среднее – 0, Стандартное отклонение -1, Выходной интервал A1: A1000.

1. Рассчитать числовые характеристики выборки: минимум, максимум, размах вариации, количество интервалов, частоту, среднее арифметическое, среднее квадратичное отклонение.

2. Построить полигон частот для заданной генеральной совокупности.

3. Выполнить анализ репрезентативности различных видов выборок. Создать три выборки: случайную, механическую, серийную. Для каждой выборки выполнить пункты 1-3. Построить полигон частот для каждой выборки. Для подведения итогов по оценке репрезентативности выборок построить на

общем графике полигоны частот трех выборок и генеральной совокупности. Для качественного представления на общем графике рекомендуется частоты генеральной совокупности соотнести с частотами выборок (можно поделить их на 8).

Собственно-случайная выборка

Пронумеровать данные любым доступным способом. Создать выборку – массив из 100 элементов, поместив в нее значения из генеральной совокупности, которые соответствуют номерам, сгенерированным случайным образом в диапазоне от 1 до 1000.

Механическая выборка

Создать 10% выборку, то есть интервал для выбора элементов установить равным 10, и отобрать 100 элементов из генеральной совокупности.

Серийная выборка

Разделить генеральную совокупность на 10 серий. Выбрать самостоятельно одну серию данных и все ее значения считать серийной выборкой.

Справка по Excel

Для формирования выборок можно воспользоваться функциями:

СЛЧИС() или СЛУЧМЕЖДУ() — генерация случайного числа. Первая функция возвращает значение в диапазоне [0,1]. Вторая — в диапазоне, заданном пользователем.

СЦЕПИТЬ() — конкатенация (склейка) двух текстовых строк.

ДВССЫЛ() — ссылка на ячейку, указанную в аргументе, как на адрес, содержащий искомое значение. Например, если мы хотим получить значение, содержащееся в ячейке С2, то необходим вызов функции: =ДВССЫЛ (с2), где с2 — текст.

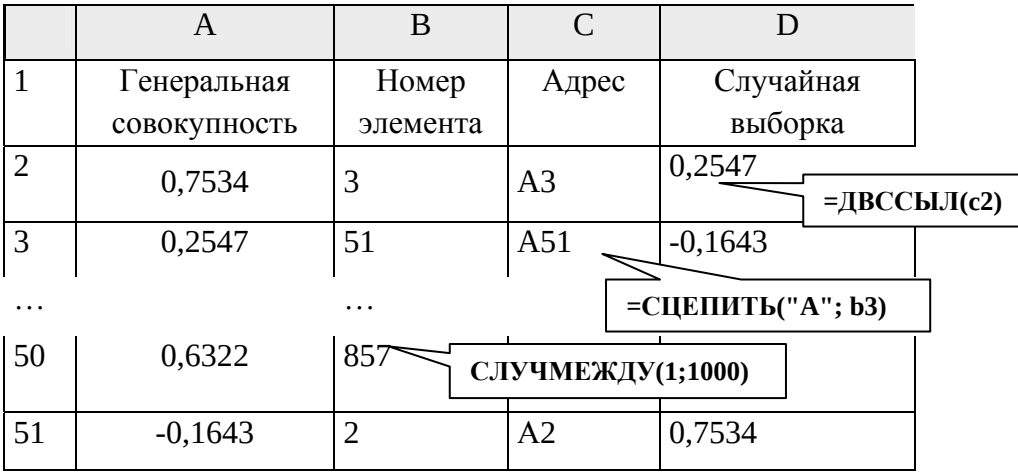


Рис. 27. Пример использования функций

Контрольные вопросы:

1. Чем отличаются случайная выборка с возвращением и без возвращения? Привести примеры.
2. Какой метод отбора элементов используется для механической выборки? Каковы недостатки механической выборки?
3. Объяснить принципы отбора элементов для типической и серийной выборок.
4. Сравнить параметры распределений и виды графиков для разных выборок и генеральной совокупности.
5. Сделать вывод о репрезентативности различных методов получения выборочных данных.

Лабораторная работа № 4

Проверка гипотез на основе критерия согласия Пирсона

Цель: *освоить алгоритм проверки непараметрических гипотез* χ^2

Критерий Пирсона или $\chi 2$ — наиболее часто используемый статистический критерий для проверки гипотезы о законе распределения случайной величины. Во многих практических задачах закон распределения неизвестен и требует определения. Для достоверного выбора того или иного закона формулируется гипотеза, которая требует подтверждения.

По выборочным данным строится полигон частот и рассчитываются параметры распределения. Гипотеза о предполагаемом законе распределения изучаемого признака выдвигается на основе исследования выборки.

Нулевая гипотеза несет информацию о законе распределения. Например: $\mathbf{H_0: F(x)=F_0(x);}$ **где** $\mathbf{F_0(x)=\Phi(x;\ \mu_0,\ \sigma_0^2).}$ Выборочная совокупность имеет нормальное распределение.

Тогда к**онкурирующая гипотеза:** выборочная совокупность имеет распределение, отличное от нормального.

Критерий Пирсона является алгоритмом, позволяющим сделать вывод о достоверности выдвинутой гипотезы. Последовательность действий для определения критерия $\chi 2$ описана ниже.

1. Построить таблицу частот опытного распределения в выбранных интервалах (см. лаб. работу 1). Если среди опытных частот имеются малочисленные $\boldsymbol{(n_i < 5),}$ то объединить их с соседними. Это будет выбор ***групп***.

2. Определить теоретические частоты при помощи выбранного закона распределения (например, нормального):

Теоретическая частота для ***i***-го интервала (группы) определяется по формуле: $n_i^o = n \cdot \left[\Phi\left(\frac{\beta_i - \overline{x}}{\sigma} \right) - \Phi\left(\frac{\alpha_i - \overline{x}}{\sigma} \right) \right]$, где n — объем выборки; β_i, α_i — границы интервала, **Φ(t)** — нормированная (стандартная) функция. Например, имеется ряд интервалов: 25, 28, 31, 34, ...

Для i = 2, α_i = 28, β_i = 31. Значение Φ(t) вычисляется (если использовать Excel), как функция нормального распределения, с μ_n = 0, σ_n = 1, а значение x — вычисляется по формуле: $x_2 = \frac{\beta_i - \overline{x}}{\sigma}$ и $x_1 = \frac{\alpha_i - \overline{x}}{\sigma}$.

3. По формуле $\chi^2 = \sum_{i=1}^{m} \frac{(n_i - n_i^o)^2}{n_i^0}$ вычислить величину χ^2. Это будет χ^2_0.

4. Определить число степеней свободы ***k***.

5. Воспользовавшись специальной таблицей, по полученным значениям χ^2 и ***k,*** найти вероятность α того, что случайная величина, имеющая χ^2 -распределение, примет какое-либо значение, не меньшее χ^2_0: $P(\chi^2 \geq \chi^2_0) = \alpha$.

6. Сформулировать вывод, руководствуясь общим принципом применения критериев согласия: если вероятность α больше 0.01, то имеющиеся расхождения между теоретическими и эмпирическими частотами следует считать несущественными, а опытное распределение — согласующимся с теоретическим. В противном случае ($\alpha \leq 0.01$), указанные расхождения признаются ***неслучайными***, а закон распределения, избранный в качестве предполагаемого теоретического — отвергается.

Задание:

1. Используя набор данных из лабораторной работы №1, провести оценку по критерию χ^2. В качестве гипотезы выбрать: «Экспериментальные данные подчиняются закону нормального распределения».
2. Рассчитать необходимые параметры для выбранной гипотезы.
3. Построить таблицу для расчета χ^2. Примерный вид таблицы для анализа (табл. 26.
4. Рассчитать критерий согласия Пирсона. Для вероятности $\alpha = 0.05$, сделать вывод подтверждении или отрицании гипотезы нормального распределения данных измерений. Воспользоваться функцией Excel — **ХИ2ОБР(),** которая выдает значения ***таблицы вероятностей P для критерия χ^2 (Пирсона).***

Если табличное значение оказалось меньше рассчитанного экспериментальным путемχ^2, то в этом случае нулевая гипотеза принимается, поскольку отклонения экспериментальных частот от теоретических являются несущественными.

Таблица 26

Расчетная таблица

Интервал α_i-β_i	Частота	$\beta_i - \overline{x}$	$\alpha_i - \overline{x}$	$\Phi(x_1)$	$\Phi(x_2)$	Теоретич частота	Разности $(n_i - n_i^0)$	$\frac{(n_i - n_i^o)^2}{n_i^0}$
							χ^2	

Контрольные вопросы

1. Объясните, чем отличаются непараметрические методы проверки гипотез от параметрических.

2. К какому из методов проверки гипотез относится критерий Пирсона?
3. Что называется теоретической частотой?
4. Опишите алгоритм проверки гипотезы по критерию χ^2.
5. Как определить число связей и число степеней свободы?
6. Что такое доверительный интервал и как он определяется?
7. Какие данные позволяют сделать вывод об истинности или ложности гипотезы при расчетах критерия Пирсона?

Лабораторная работа № 5

Основы корреляционного анализа

Цель: *формирование навыков изучения тесноты и вида связи между случайными величинами.*

Исследуйте зависимость между официальным курсом доллара США за 12 месяцев 2009 г и ценой за баррель нефти сорта «Юралс». Статистические данные получены с сайта Банка России: http://www.cbr.ru/statistics и представлены в табл. 27.

Таблица 27

Курс доллара и цена нефти за 2009 год

№	1	2	3	4	5	6	7	8	9	10	11	12
X	41,9	42,2	45,4	48,7	56,5	68,2	64,3	72	67	72,4	76	73,7
Y	35,4	35,7	34,0	33,3	31,0	31,3	31,8	31,6	30,1	29,1	29,8	30,2

Обозначения данных в таблице: N – номер месяца 2009 года, Y – курс доллара в рублях, X – цена 1 барреля нефти в долларах. Выборочное уравнение прямой регрессии Y на X определяется по формуле:

$$Y_i - \overline{Y} = \frac{r_{xy} \cdot \delta_y \cdot (X_i - \overline{X})}{\delta_x}.$$

Задание

1. По данным выборки найти корреляционную зависимость между случайными величинами и построить линейное приближение методом наименьших квадратов.
2. Построить диаграмму рассеяния.
3. Рассчитать коэффициент корреляции r_{xy} и оценить вид связи (прямая, обратная, сильная, слабая, нет связи).
4. По внешнему виду диаграммы рассеяния определить, является связь прямой или обратной.
5. Построить однофакторную линейную регрессионную модель связи признаков X и Y, используя инструмент Регрессия надстройки Пакет анализа, и оценить тесноту связи признаков X и Y на основе линейного коэффициента корреляции r.
6. Добавить на диаграмму рассеяния линию тренда (зависимость определить линейной) и вывести на диаграмму уравнение регрессии (во вкладке Параметры опция «Показывать уравнение на диаграмме»).

Контрольные вопросы

1. Что такое корреляция?
2. Как измерить связь между двумя случайными величинами?
3. Какие виды зависимостей существуют между величинами?
4. Что показывает коэффициент корреляции?
5. Приведите значения коэффициента корреляции и вид связей, который соответствует каждому значению.
6. Какие выводы можно сделать по внешнему виду диаграммы рассеяния?

Лабораторная работа № 6

Линейный регрессионный анализ

Цель: *получение навыков исследования данных с помощью линейного регрессионного анализа. Знакомство с методом наименьших квадратов.*

Пусть имеются две коррелированные случайные величины X и Y. Если связь между двумя величинами линейная, то ее можно представить зависимостью:

$$\tilde{y} = A + Bx,$$

где ***A*** и ***B*** — это соответственно отрезок оси ординат, отсекаемой прямой, и ее наклон. Если данные связаны идеальной линейной зависимостью, то можно предсказать значение $\tilde{y}_i$ по известному значению x_i. Если предположить существование линейной связи и наличие неограниченной выборки, то можно подобрать значения параметров ***A*** и ***B,*** которые обеспечат расчет прогнозируемого $\tilde{y}_i$.

Общепринятая процедура определения коэффициентов уравнения состоит в выборе таких значений ***A*** и ***B***, которые минимизируют сумму квадратов отклонений наблюдаемых значений от предсказанного значения ***y***. Эта процедура называется ***методом наименьших квадратов***. Поскольку отклонения наблюдаемых значений от предсказанных равны

$y_i - \tilde{y} = y_i - (A + Bx_i)$, то сумма квадратов отклонений имеет вид

$$Q = \sum_{i=1}^{n} (y_i - A - Bx_i)^2.$$

Следовательно, наилучшее согласие в смысле наименьших квадратов обеспечивают значения ***A*** и ***B***, для которых

$$\frac{\partial Q}{\partial A} = \frac{\partial Q}{\partial B} = 0.$$

Мы имеем ограниченную выборка из ***N*** пар наблюдений значений ***x*** и ***y***. Это означает, что данное уравнение даст всего лишь оценки ***A*** и ***B***; обозначим их через ***a*** и ***b*** соответственно. Решая систему уравнений относительно оценок величин A и B, получим

$$a = \bar{y} - b\bar{x},$$

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

или в обозначениях коэффициента корреляции и выборочных дисперсией:

$$b = \frac{s_{xy}}{s_x^2} = r_{xy} \cdot \frac{s_y}{s_x}.$$

Эти оценки можно использовать для построения модели, позволяющей предсказывать ***y*** по данному ***x***:

$$\hat{y} = a + bx = \bar{y} + b(x - \bar{x}).$$

Прямая линия, задаваемая указанным уравнением, называется прямой линейной регрессии y на x.

Доверительные интервалы

В лабораторной работе № 1 использовались выборочные значения для оценки параметров распределений случайных величин. Такие процедуры дают только ***точечные оценки*** интересующих параметров. Они не позволяют судить о степени близости выборочных значений к оцениваемому параметру. Более содержательны процедуры оценивания параметров, связанные не с получением точечного значения, а с построением

интервала, который накрывает оцениваемый параметр с известной степенью достоверности.

Пусть, например, выборочное среднее арифметическое $\bar{x}$, вычисленное по n независимым наблюдениям случайной величины x, используется в качестве оценки среднего μ_x. Обычно представляет интерес оценить μ_x в терминах некоторого интервала $\bar{x} \pm d$, в который μ_x попадает с заданной ***степенью достоверности***. Такие интервалы можно построить, если известны выборочные распределения рассматриваемой оценки.

Относительно значения выборочного среднего $\bar{x}$ можно сделать следующее вероятностное утверждение:

$$P\left\{\Phi(1-\alpha/2) < \frac{(\bar{x}-\mu_x)\sqrt{n}}{\sigma_x} \le \Phi(\alpha/2)\right\} = 1-\alpha,$$

где Ф() — стандартная функция распределения; α — называется вероятностью ошибки, или уровнем значимости. А значение S = 1 – α — статистической достоверностью. S измеряют часто в процентах и говорят, например, о 95%-м доверительном интервале (S = (1 – α)·100%). В этом случае $\alpha = 0.05$ (см. лабораторную работу о критерии оценки хи-квадрат).

По мере уменьшения α (увеличения интервала, заключенного между Ф(1-α/2) и Ф(α/2)) разумно считать, что вероятность Р скорее равна единице, чем нулю. Иначе говоря, если производится много выборок, и для каждой из них вычисляется $\bar{x}$, то можно ожидать, что она будет попадать в указанный интервал с относительной частотой, примерно равной 1 – α. При таком подходе можно утверждать, что существует интервал, в который величина $\frac{(\bar{x}-\mu_x)\sqrt{n}}{\sigma_x}$ попадает

с большой степенью достоверности. Такие утверждения называют ***доверительными***. Интервал, относительно которого делается доверительное утверждение, называется ***доверительным интервалом***. Степень доверия, сопоставляемая доверительному утверждению, называется ***уровнем доверия***.

Распределение $\hat{y}$ при конкретном значении $x = x_0$ представляет особый интерес ($\hat{y}$ — значение, вычисленное по полученному уравнению регрессии, $\tilde{y}$ — оценочное интервальное значение):

$$\tilde{y} = \hat{y} \pm s_{y|x} t_{N-2} \cdot \left(\frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \right)^{1/2}.$$

Значение $\Delta y = s_{y|x} t_{N-2} \cdot \left(\frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \right)^{1/2}$ будет определять границы интервала для заданного значения α.

В формулах величина $s_{y|x}$ — выборочное стандартное отклонение наблюдаемого значения y_i от предсказанного $\hat{y}_i = a + bx_i$, равное

$$s_{y|x} = \left[\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N-2} \right]^{1/2} = \left[\left(\frac{n-1}{n-2} \right) s_y^2 \left(1 - r_{xy}^2 \right) \right]^{1/2}.$$

Задание

В табл. 28 приведены значения двух величин, которые являются характеристиками массы и расхода электроэнергии поездов. Пусть X- масса поезда, выраженная в тыс. т., Y – удельный расход электроэнергии, кВт/ч на10 тыс. км.

Определить: выборочное уравнение прямой регрессии **Y** на **X**. Сделать вывод о характере и тесноте связи между массой поезда **X** и удельным расходом электроэнергии **Y**.

Таблица 28

Данные о массе поезда и расходе электроэнергии

№	Масса поезда, X, тыс. т.	Электроэнергия, Y, кВт/ч на 10000 км
1	2,5	85
2	2,5	105
3	3	85
4	3	95
5	3	105
6	3,5	75
7	3,5	85
8	3,5	95
9	4	75
10	4	85
11	4	95
12	4,5	75
13	4,5	85

1. Построить ***диаграмму рассеяния.***
2. Рассчитать коэффициент корреляции для данной популяции. Сделать вывод о виде ***связи***.
3. Рассчитать ***коэффициенты*** прямой регрессии **a** и **b**. Составить уравнение регрессии.

4. Построить по *заданным x* и *вычисленным* в уравнении регрессии *y* — прямую регрессии на том же графике, что и диаграмма рассеяния.

5. Рассчитать прогнозируемые затраты электроэнергии для поездов с массой: 2, 3, 4 тыс. тонн.

6. Определить доверительные интервалы для рассчитанных данных. Добавить несколько данных о массе и расходе электроэнергии поездами в исходную таблицу.

7. Выполнить задание заново на **новом листе.** Сделать вывод о размере доверительного интервала при увеличении выборки.

8. Выбрать значение α — вероятность ошибки, которая задает границы доверительного интервала. Например, для 95%-го интервала $\alpha = 0.05$.

9. Определить по таблице $t_{N-2} = t_{N-2;\alpha/2}$. Половина α для доверительного интервала берется, так как t-распределение симметричное.

10. Пример таблицы для t-распределения Стьюдента представлен в табл. 29.

11. При расчете этого коэффициента можно воспользоваться функцией Excel:

СТЬЮДРАСПОБР (Вероятность, степени_свободы), где **вероятность** — значение α, **степени_свободы** — N-2, объем выборки минус количество связей.

Рекомендуется проверить данные, полученные при помощи этой функции.

7. Рассчитать значение интервала Δy и получить интервальную оценку в виде: $\tilde{y} = \hat{y}_i \pm \Delta y$

Таблица 29

Пример t-распределения Стьюдента

n	α				
	0.100	0.050	0.025	0.010	0.005
1	3.078	6.314	12.706	31.831	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
…	…	…	…	…	…
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
…	…	…	…	…	…

Контрольные вопросы

1. Каковы этапы построения регрессионной модели?
2. Что представляет собой линейный регрессионный анализ?
3. В чем суть метода наименьших квадратов?
4. Как определяются коэффициенты линейной регрессии?
5. Каким образом определяется уравнение линейной регрессии?
6. Что называется доверительным интервалом и как он определяется?
7. Как определяются границы доверительных интервалов?

Лабораторная работа № 7

Множественный регрессионный анализ

Цель: *получение навыков выполнения множественного регрессионного анализа, построения уравнения регрессии в статистическом пакете Statistica.*

В результате исследования бюджетов домохозяйств за десять месяцев прошедшего года были получены сведения о средней месячной стоимости четырех продуктов питания (X_1, X_2, X_3, X_4), а также данные о затратах на продукты домохозяйства в целом (Y). Необходимо выполнить множественный регрессионный анализ для изучения функциональной зависимости между продуктовыми затратами домохозяйства и стоимостями продуктов питания.

Таблица 30

Затраты домохозяйства за 10 месяцев текущего года и стоимости продуктов

№	Y	X1	X2	X3	X4
1	4765,3	26,2	18,5	175	170
2	3851,7	23,8	16,5	160	140
3	5117,9	27,6	18,5	170	185
4	6205	28	21	195	220
5	6099,2	25,5	22	216	210
6	6483,1	27,5	17,9	190	230
7	4619,4	26	18,6	185	160
8	5102,65	28,5	18,85	190	175
9	6247,6	27,5	18,4	215	210
10	4253,4	24,8	17,8	175	150

1. Получить дескриптивные статистики по каждому признаку (среднее арифметическое, дисперсию, среднее

квадратичное отклонение).

2. Рассчитать коэффициенты парной корреляции для всех признаков. Оценить показатели вариации каждого признака и сделать вывод о возможностях применения метода наименьших квадратов для их изучения. Проанализировать линейные коэффициенты парной и частной корреляции.

3. Составить уравнение множественной регрессии, оценить его параметры.

4. С помощью F-критерия Фишера оценить статистическую надежность уравнения регрессии в целом.

Рекомендации к выполнению:

1. С помощью соответствующих статистических функций в Excel определить: среднее арифметическое, дисперсию, среднее квадратичное отклонение по каждому признаку (Y, X_1, X_2, X_3, X_4).

2. Постройте матрицу парных коэффициентов корреляции:

	Y	X_1	X_2	X_3	X_4
Y	1	0,508	0,494	0,754	0,987
X_1	0,508	1	0,307	0,394	0,465
X_2	0,494	0,307	1	0,689	0,447
X_3	0,754	0,394	0,689	1	0,645
X_4	0,987	0,465	0,447	0,645	1

Очевидно, что все четыре изучаемые фактора имеют линейную связь с результирующим показателем Y, поскольку их коэффициенты корреляции с Y достаточно высокие. Межфакторная корреляция признаков не превышает 0,7, поэтому все факторы рекомендуется включить в уравнение множественной регрессии.

3. Для составления уравнения множественной регрессии нужно подставить рассчитанные коэффициенты корреляции в уравнения:

$$r_{yx1} = \beta_1 + \beta_2 r_{x2x1} + \beta_3 r_{x3x1} + ... + \beta_p r_{xpx1};$$
$$r_{yx2} = \beta_1 r_{x1x2} + \beta_2 + \beta_3 r_{x3x2} + ... + \beta_p r_{xpx2};$$
$$.....$$
$$r_{yxp} = \beta_1 r_{x1xp} + \beta_2 r_{x2xp} + \beta_3 r_{x3xp} + ... + \beta_p.$$

Получим такую систему линейных уравнений:

$$0{,}508 = \beta_1 + 0{,}307 \cdot \beta_2 + 0{,}394 \cdot \beta_3 + 0{,}465\beta_4;$$
$$0{,}494 = 0{,}307 \cdot \beta_1 + \beta_2 + 0{,}689\beta_3 + 0{,}447 \cdot \beta_4;$$
$$0{,}754 = 0{,}394 \cdot \beta_1 + 0{,}689 \cdot \beta_2 + \beta_3 + 0{,}645 \cdot \beta_4$$
$$0{,}987 = 0{,}465 \cdot \beta_1 + 0{,}447 \cdot \beta_2 + 0{,}645 \cdot \beta_3 + \beta_4.$$

Для решения системы уравнений и определения стандартизованных коэффициентов β_i можно использовать матричный метод (создать обратную матрицу, а затем для определения неизвестных найти сумму произведений элементов каждой строки матрицы со свободными членами). Для построения обратной матрицы можно воспользоваться функцией МОБР(), для которой агрументом будут элементы исходной матрицы, а для определения стандартизованных коэффициентов β_i нужно применить функцию СУММПРОИЗВ.

Например, для рассчета коэффициента β_1 нужно ввести функцию:

=СУММПРОИЗВ (H19:K19;H27:K27),

для рассчета коэффициента β_2 нужно ввести функцию:

=СУММПРОИЗВ (H19:K19;H28:K28) и т.д.

Образцы рассчетных матриц приведены ниже.

	H	I	J	K
№	Свободные члены матрицы			
19	0,512	0,571	0,784	0,979
	Исходная матрица			
	1	0,307	0,394	0,465
	0,307	1	0,689	0,447
	0,394	0,689	1	0,645
	0,465	0,447	0,645	1
	Обратная матрица			
27	1,305	-0,085	-0,152	-0,472
	-0,085	1,911	-1,299	0,023
	-0,152	-1,299	2,631	-1,047
	-0,472	0,023	-1,047	1,885
	Коэффициенты			
	β_1	β_2	β_3	β_4
	0,039022	0,052289	0,218454	0,7960302

Затем, на основе стандартизованных коэффициентов β_i рассчитать коэффициенты линейной регресси и b_i по формуле: $b_i = \beta_i \cdot \frac{\delta_y}{\delta_{xi}}$.

$$b_1 = 0{,}039022 \cdot \frac{938{,}68}{1{,}53} = 23{,}96 \approx 24$$

$$b_2 = 0{,}052289 \cdot \frac{938{,}68}{1{,}58} = 31{,}065 \approx 31$$

$$b_3 = 0{,}218454 \cdot \frac{938{,}68}{18{,}31} = 11{,}19 \approx 11{,}2$$

$$b_4 = 0{,}7960302 \cdot \frac{938{,}68}{29{,}89} = 24{,}99 \approx 25$$

После этого рассчитывается свободный член регрессионного уравнения а:

$$a = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 - .... - b_p \bar{x}_p,$$

$$a = 6465{,}44 - 24 * 26{,}54 - 31 * 18{,}81 - 11{,}2 * 187{,}1 - 25 * 186 \approx -1500.$$

Затем составляется уравнение линейной регрессии:

$$\hat{y} = -1500 + 24x_1 + 31x_2 + 11{,}2x_3 + 25x_4.$$

Контрольные вопросы

1. Как оценивается значимость коэффициента корреляции?
2. Что характеризуют параметры регрессионного уравнения? Объясните сущность коэффициента парной линейной регрессии.
3. Как оценивается значимость параметров регрессионного уравнения?
4. Дайте определение стандартизованному коэффициенту регрессии. Что он характеризует?
5. Что позволяет оценить множественный коэффициент детерминации?
6. Как оценить статистическую надежность регрессионного уравнения в целом?

Список рекомендуемой литературы

1. Вентцель Е.С. Теория вероятностей и ее инженерные приложения / Е.С. Вентцель, Л.А. Овчаров – М.: Высш. шк, 2007. – 496с.
2. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике: учеб. пособие.– 11-е изд. – М.: Высшее образование, 2009. – 416 с.
3. Гмурман В.Е. Теория вероятностей и математическая статистика: Учеб. пособие.– 12-е изд., перераб. – М.: Высшее образование, 2009. – 480 с.
4. Дубров А.М. Многомерные статистические методы : учебник А.М. Дубров, В.С. Мхитарян, Л.И. Трошин. – М. : Финансы и статистика, 2000. – 352 с.
5. Калинина В. Н. Теория вероятностей и математическая статистика. Компьютерно-ориентированный курс / В.Н. Калинина. – М.: Дрофа, 2009. – 480 с.
6. Кремер Н.Ш. Теория вероятностей и математическая статистика : учебник для вузов / Н.Ш. Кремер. – М. : ЮнитиДана, 2007. – 551 с.
7. Соколов Г.А. Теория вероятностей : учебник / Г.А Соколов, Н.А. Чистякова.- М. : Изд-во «Экзамен», 2005.– 416 с.
8. Чистяков В.П. Курс теории вероятностей : учебник для вузов / В.П. Чистяков. – 7-е изд. – М. : Дрофа, 2007. – 256 с.
9. Шведов А.С. Теория вероятностей и математическая статистика : учеб. пособие / А.С. Шведов. – М. : Изд. Дом ГУ ВШЭ, 2005. – 254 с.
10. Шмойлова Р.А. Практикум по теории статистики : учеб. пособие / Р.А. Шмойлова, В.Г. Минашкин, Н.А. Садовникова;

под ред. Р.А. Шмойловой. – М. : Финансы и статистика, 2005. – 416 с.
11. Шмойлова Р.А. Теория статистики : учебник / Р.А. Шмойлова, В.Г. Минашкин, Н.А. Садовникова, Е.Б. Шувалова; под ред. Р.А. Шмойловой. – 4-е изд., перераб. и доп.– М. : Финансы и статистика, 2005. – 656 с.

Учебное издание

Гусева Елена Николаевна

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Учебное пособие


МАГНИТОГОРСК, 2009