ИДА Кривой Рог IBM

Частное Учебное Заведение

Институт Делового Администрирования

Private Educational Institution

Institute of Business Managment

*Кафедра информационных систем*

*и*

*высшей математики*

*Математическая cтатистика*



*Конспект лекций*

для специальностей УА, ФК 1995

© Г.И. Корнилов

1997

# Введение в курс

## Основные определения

Несмотря на многообразие используемых в литературе определений термина “статистика”, суть большинства из них сводится к тому, что статистикой чаще всего называют науку, изучающую методы сбора и обработки фактов и данных в области человеческой деятельности и природных явлений.

В нашем курсе, который можно считать введением в курс “Экономическая статистика”, речь будет идти о так называемой прикладной статистике, − т.е. только о сущности специальных методов сбора, обработки и анализа информации и, кроме того, о практических приемах выполнения связанных с этим расчетов.

Великому американскому сатирику О’Генри принадлежит ироническое определение статистики: “Есть три вида лжи − просто ложь, ложь злостная и …статистика!”. Попробуем разобраться в причинах, побудивших написать эти слова.

Практически всему живому на земле присуще воспринимать окружающую среду как непрерывную последовательность фактов, событий. Этим же свойством обладают и люди, с той лишь разницей, что только им дано анализировать поступающую информацию и (хотя и не всем из них это удается) делать выводы из такого анализа и учитывать их в своей сознательной деятельности. Поэтому можно смело утверждать, что во все времена, все люди занимались и занимаются статистическими “исследованиями”, даже не зная иногда такого слова − “статистика”.

Все наши наблюдения над окружающем нас миром можно условно разделить на два класса:

• наблюдения за фактами − событиями, которые могут произойти или не произойти;

• наблюдения за физическими величинами, значения которых в момент наблюдения могут быть различными.

И атеист и верующий в бога человек, скорее всего, согласятся с несколько необычным заявлением − в окружающем нас мире происходят только случайные события, а наблюдаемые нами значения всех показателей внешней среды являются случайными величинами (далее везде – СВ) . Более того, далее будет показано, что иногда можно использовать только одно понятие − случайное событие.

Не задерживаясь на раскрытии философской сущности термина “случайность” (вполне достаточно обычное, житейское представление), обратимся к чрезвычайно важному понятию − вероятность. Этот термин обычно используют по отношению к событию и определяют числом (от 0 до 1), выражающим степень нашей уверенности в том, что данное событие произойдет. События с вероятностью 0 называют невозможными, а события с вероятностью 1 − достоверными (хотя это уже ­– неслучайные, детерминированные события).

Иногда в прикладной статистике приходится иметь дело с так называемыми редкими (маловероятными) событиями. К ним принято относить события, значение вероятности которых не превышает определенного уровня, чаще всего – 0.05 или 5 %.

В тех случаях, когда профессионалу−статистику приходится иметь дело со случайными величинами, последние часто делят на две разновидности:

• дискретные СВ, которые могут принимать только конкретные, заранее оговоренные значения (например, − значения чисел на верхней грани брошенной игральной кости или порядковые значения текущего месяца);

• непрерывные СВ (чаще всего − значения некоторых физических величин: веса, расстояния, температуры и т.п.), которые по законам природы могут принимать любые значения, хотя бы и в некотором интервале.

## Вероятности случайных событий

Итак, основным “показателем” любого события (факта) А является численная величина его вероятности P(A), которая может принимать значения в диапазоне [0…1] − в зависимости от того, насколько это событие случайно. Такое, смысловое, определение вероятности не дает, однако, возможности указать путь для вычисления ее значения.

Поэтому необходимо иметь и другое, отвечающее требованиям практической работы, определение термина “вероятность”. Это определение можно дать на основании житейского опыта и обычного здравого смысла.

Если мы интересуемся событиемA, то, скорее всего, можем наблюдать, фиксировать факты его появления. Потребность в понятии вероятности и ее вычисления возникнет, очевидно, только тогда, когда мы наблюдаем это событие не каждый раз, либо осознаем, что оно может произойти, а может не произойти. И в том и другом случае полезно использовать понятие частоты появления события fA − как отношения числа случаев его появления (благоприятных исходов или частостей) к общему числу наблюдений.

Интуиция подсказывает, что частота наступления случайного события зависит не только от степени случайности самого события. Если мы наблюдали за событием  всего пять раз и в трех случаях это событие произошло, то мало кто примет значение вероятности такого события равным 0.6 или 60 %. Скорее всего, особенно в случаях необходимости принятия каких–то важных, дорогостоящих решений любой из нас продолжит наблюдения. Здравый смысл подсказывает нам, что уж если в 100 наблюдениях событие  произошло 14 раз, то мы можем с куда большей уверенностью полагать его вероятность равной 14 % .

Таким образом, мы (конечно же, − не первые) сформулировали второе определение понятия вероятности события − как предела, к которому стремится частота наблюдения за событием при непрерывном увеличении числа наблюдений. Теория вероятностей, специальный раздел математики, доказывает существование такого предела и сходимость частоты к вероятности при стремлении числа наблюдений к бесконечности. Это положение носит название центральной предельной теоремы или закона больших чисел.

Итак, первый ответ на вопрос − как найти вероятность события, у нас уже есть. Надо проводить эксперимент и устанавливать частоту наблюдений, которая тем точнее даст нам вероятность, чем больше наблюдений мы имеем.

Ну, а как быть, если эксперимент невозможен (дорог, опасен или меняет суть процессов, которые нас интересуют)? Иными словами, нет ли другого пути вычисления вероятности событий, без проведения экспериментов?

Такой путь есть, хотя, как ни парадоксально, он все равно основан на опыте, опыте жизни, опыте логических рассуждений. Вряд ли кто либо будет производить эксперименты, подбрасывая несколько сотен или тысячу раз симметричную монетку, чтобы выяснить вероятность появления герба при одном бросании! Вы будете совершенно правы, если без эксперимента найдете вероятность выпадения цифры 6 на симметричной игральной кости и т.д., и т.п.

Этот путь называется статистическим моделированием – использованием схемы случайных событий и с успехом используется во многих приложениях теоретической и прикладной статистики. Продемонстрируем этот путь, рассматривая вопрос о вероятностях случайных величин дальше. Обозначим  величину вероятности того, что событие A не произойдет. Тогда из определения вероятности через частоту наступления события следует, что

P(A)+ = 1, {1–1}

что полезно читать так − вероятность того, что событие произойдет или не произойдет, равна 100 %, поскольку третьего варианта попросту нет.

Подобные логические рассуждения приведут нас к более общей формуле − сложения вероятностей. Пусть некоторое случайное событие может произойти только в одном из 5 вариантов, т.е. пусть имеется система из трех несовместимых событий A, B и C .

Тогда очевидно, что:

P(A) + P(B) + P(C) = 1; {1–2} и столь же простые рассуждения приведут к выражению для вероятности наступления одного из двух несовместимых событий (например, A или B):

P(A∪B) = P(A) + P(B); {1–3} или одного из трех:

P(A∪B∪C) = P(A) + P(B) + P(C); {1-4} и так далее.

Рассмотрим чуть более сложный пример. Пусть нам надо найти вероятность события C, заключающегося в том, что при подбрасывании двух разных монет мы получим герб на первой (событие A) и на второй (событие B). Здесь речь идет о совместном наступлении двух независимых событий, т.е. нас интересует вероятность P(C) = P(A∩ B).

И здесь метод построения схемы событий оказывается чудесным помощником − можно достаточно просто доказать, что

P(A∩B) =P(A)•P(B). {1-5} Конечно же, формулы {1-4} и {1-5} годятся для любого количества событий: лишь бы они были несовместными в первом случае и независимыми во втором.

Наконец, возникают ситуации, когда случайные события оказываются взаимно зависимыми. В этих случаях приходится различать условные вероятности:

P(A / B) – вероятность A при условии, что B уже произошло;

P(A / ) – вероятность A при условии, что B не произошло,

называя P(A) безусловной или полной вероятностью события A .

Выясним вначале связь безусловной вероятности события с условными. Так как событие A может произойти только в двух, взаимоисключающих вариантах, то, в соответствии с {1–3} получается, что

P(A) = P(A/B)•P(B) + P(A/)• P(). {1–6}

Вероятности P(A/B) и P(A/) часто называют апостериорными (“a posteriopri” – после того, как…), а безусловную вероятность P(A) – априорной (“a priori” – до того, как…).

Очевидно, что если первым считается событие B и оно уже произошло, то теперь наступление события A уже не зависит от B и поэтому вероятность того, что произойдут оба события составит

P(A∩B) = P(A/B)•P(B). {1–7} Так как события взаимозависимы, то можно повторить наши выводы и получить

P(B) = P(B/A)•P(A) + P(B/)•P(); {1–8}

а также P(A∩B) = P(B/A)•P(A). {1–9}

Мы доказали так называемую теорему Байеса

P(A/B)•P(B) = P(B/A)•P(B); {1–10} – весьма важное средство анализа, особенно в области проверки гипотез и решения вопросов управления на базе методов прикладной статистики.

Подведем некоторые итоги рассмотрения вопроса о вероятностях случайных событий. У нас имеются только две возможности узнать что либо о величине вероятности случайного события A:

• применить метод статистического моделирования − построить схему данного случайного события и (если у нас есть основания считать, что мы правильно ее строим) и найти значение вероятности прямым расчетом;

• применить метод статистического испытания − наблюдать за появлением события и затем по частоте его появления оценить вероятность.

На практике приходится использовать оба метода, поскольку очень редко можно быть абсолютно уверенным в примененной схеме события (недостаток метода моделирования) и столь же редко частота появления события достаточно быстро стабилизируется с ростом числа наблюдений (недостаток метода испытаний).

# Распределения вероятностей случайных величин

## Шкалирование случайных величин

Как уже отмечалось, дискретной называют величину, которая может принимать одно из счетного множества так называемых “допустимых” значений. Примеров дискретных величин, у которых есть некоторая именованная единица измерения, можно привести достаточно много.

Прежде всего, надо учесть тот факт что все физические величины (вес, расстояния, площади, объемы и т.д.) теоретически могут принимать бесчисленное множество значений, но практически − только те значения, которые мы можем установить измерительными приборами. А это значит, что в прикладной статистике вполне допустимо распространить понятие дискретных СВ на все без исключения численные описания величин, имеющих единицы измерения.

Вместе с тем надо не забывать, что некоторые СВ просто не имеют количественного описания, естественных единиц измерения (уровень знаний, качество продукции и т. п.).

Покажем, что для решения вопроса о “единицах измерения” любых СВ, с которыми приходится иметь дело в прикладной статистике, достаточно использовать четыре вида шкал.

• Nom. Первой из них рассмотрим так называемую номинальную шкалу — применяемую к тем величинам, которые не имеют природной единицы измерения. В ряде случаев нам приходится считать случайными такие показатели предметов или явлений окружающего нас мира, как марка автомобиля; национальность человека или его пол, социальное положение; цвет некоторого изделия и т.п.

В таких ситуациях можно говорить о случайном событии − "входящий в магазин посетитель оказался мужчиной", но вполне допустимо рассматривать пол посетителя как дискретную СВ, которая приняла одно из допустимых значений на своей номинальной шкале.

Итак, если некоторая величина может принимать на своей номинальной шкале значения X, Y или Z, то допустимыми считаются только выражения типа: X # Y, X=Z , в то время как выражения типа X ≥ Z, X + Z не имеют никакого смысла.

• Ord. Второй способ шкалирования – использование порядковых шкал. Они незаменимы для СВ, не имеющих природных единиц измерения, но позволяющих применять понятия предпочтения одного значения другому. Типичный пример: оценки знаний (даже при числовом описании), служебные уровни и т. п. Для таких величин разрешены не только отношения равенства (= или #), но и знаки предпочтения (> или <). Очень часто порядковые шкалы называют ранговыми и говорят о рангах значений таких величин.

• Int. Для СВ, имеющих натуральные размерности (единицы измерения в прямом смысле слова), используется интервальная шкала. Для таких величин, кроме отношений равенства и предпочтения, допустимы операции сравнения – т. е. все четыре действия арифметики. Главная особенность таких шкал заключается в том, что разность двух значений на шкале (36 и 12) имеет один смысл для любого места шкалы (28 и 4). Вместе с тем на интервальной шкале не имеют никакого смысла отрицательные значения, − если это веса предметов, возраст людей и подобные им показатели.

• Rel. Если СВ имеет естественную единицу измерения (например, − температура по шкале Цельсия) и ее отрицательные значения столь же допустимы, как и положительные, то шкалу для такой величины называют относительной.

Методы использования описанных шкал относится к специальному разделу – так называемой непараметрической статистике и обеспечивают, по крайней мере, два неоспоримых преимущества.

• Появляется возможность совместного рассмотрения нескольких СВ совершенно разной природы (возраст людей и их национальная принадлежность, марка телевизора и его стоимость) на единой платформе − положения каждой из величин на своей собственной шкале.

• Если мы сталкиваемся с СВ непрерывной природы, то использование интервальной или относительной шкалы позволит нам иметь дело не со случайными величинами, а со случайными событиями — типа “вероятность того, что вес продукции находится в интервале 17 Кг”. Появляется возможность применения единого подхода к описанию всех интересующих нас показателей при статистическом подходе к явлениям окружающего нас мира.

## Законы распределений дискретных случайных величин.

Пусть некоторая СВ является дискретной, т.е. может принимать лишь фиксированные (на некоторой шкале) значения X i. В этом случае ряд значений вероятностей P(X i)для всех (i=1…n) допустимых значений этой величины называют её законом распределения.

В самом деле, − такой ряд содержит всю информацию о СВ, это максимум наших знаний о ней. Другое дело, − откуда мы можем получить эту информацию, как найти закон распределения? Попытаемся ответить на этот принципиально важный вопрос, используя уже рассмотренное понятие вероятности.

Точно также, как и для вероятности случайного события, для закона распределения СВ есть только два пути его отыскания. Либо мы строим схему случайного события и находим аналитическое выражение (формулу) вычисления вероятности (возможно, кто–то уже сделал или сделает это за нас!), либо придется использовать эксперимент и по частотам наблюдений делать какие–то предположения (выдвигать гипотезы) о законе распределения.

Заметим, что во втором случае нас будет ожидать новый вопрос, − а какова уверенность в том, что наша гипотеза верна? Какова, выражаясь языком статистики, вероятность ошибки при принятии гипотезы или при её отбрасывании?

Продемонстрируем первый путь отыскания закона распределения.

Пусть важной для нас случайной величиной является целое число, образуемое по следующему правилу: мы трижды бросаем симметричную монетку, выпадение герба считаем числом 1 (в противном случае 0) и после трех бросаний определяем сумму S.

Ясно, что эта сумма может принимать любое значение в диапазоне 0…3, но всё же − каковы вероятности P(S=0), P(S=1), P(S=2), P(S=3); что можно о них сказать, кроме очевидного вывода − их сумма равна 1?

Попробуем построить схему интересующих нас событий. Обозначим через p вероятность получить 1 в любом бросании, а через q=(1–p) вероятность получить 0. Сообразим, что всего комбинаций ровно 8 (или 23), а поскольку монетка симметрична, то вероятность получить любую комбинацию трех независимых событий (000,001,010…111) одна и та же: q3 = q2•p=…= p3 = 0.125 . Но если p # q , то варианты все тех же восьми комбинаций будут разными:

Таблица 1−1

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Первое бросание | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| Второе бросание | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 |
| Третье бросание | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| Сумма S | 0 | 1 | 1 | 2 | 1 | 2 | 2 | 3 |
| Вероятность P(S) | q3 | q2•p | q2•p | q•p2 | q2•p | q•p2 | q•p2 | p3 |

Запишем то, что уже знаем − сумма вероятностей последней строки должна быть равна единице:

p3 +3•qp2 + 3•q2•p + q3 = (p + q)3  = 1.

Перед нами обычный бином Ньютона 3-й степени, но оказывается − его слагаемые четко определяют вероятности значений случайной величины **S** !

Мы “открыли” закон распределения СВ, образуемой суммированием результатов n последовательных наблюдений, в каждом из которых может появиться либо 1 (с вероятностью p), либо 0 (с вероятностью 1– p).

Итог этого открытия достаточно скромен:

• возможны всего N = 2 n вариантов значений суммы;

• вероятности каждого из вариантов определяются элементами разложения по

степеням бинома (p + q) n ;

• такому распределению можно дать специальное название − биномиальное.

Конечно же, мы опоздали со своим открытием лет на 300, но, тем не менее, попытка отыскания закона распределения с помощью построения схемы событий оказалась вполне успешной.

В общем случае биномиальный закон распределения позволяет найти вероятность события S = k в виде

P(S=k)=•pk•(1– p)n-k, {2–1} где  − т.н. биномиальные коэффициенты, отыскиваемые из известного “треугольника Паскаля” или по правилам комбинаторики − как число возможных сочетаний из n элементов по k штук в каждом:

= n•(n –1)• ...•(n – k + 1)/ (1•2• .... • k). {2–2}

Многие дискретные СВ позволяют построить схему событий для вычисления вероятности каждого из допустимых для данной случайной величины значений.

Конечно же, для каждого из таких, часто называемых "классическими", распределений уже давно эта работа проделана ­– широко известными и очень часто используемыми в прикладной статистике являются биномиальное и полиномиальное распределения, геометрическое и гипергеометрическое, распределение Паскаля и Пуассона и многие другие.

Для почти всех классических распределений немедленно строились и публиковались специальные статистические таблицы, уточняемые по мере увеличения точности расчетов. Без использования многих томов этих таблиц, без обучения правилам пользования ими последние два столетия практическое использование статистики было невозможно.

Сегодня положение изменилось – нет нужды хранить данные расчетов по формулам (как бы последние не были сложны!), время на использование закона распределения для практики сведено к минутам, а то и секундам.

Уже сейчас существует достаточное количество разнообразных пакетов прикладных компьютерных программ для этих целей. Кроме того, создание программы для работы с некоторым оригинальным, не описанным в классике распределением не представляет серьезных трудностей для программиста “средней руки”.

Приведем примеры нескольких распределений для дискретных СВ с описанием схемы событий и формулами вычисления вероятностей. Для удобства и наглядности будем полагать, что нам известна величина p – вероятность того, что вошедший в магазин посетитель окажется покупателем и обозначая (1– p) = q.

• Биномиальное распределение

Если X – число покупателей из общего числа n посетителей, то вероятность P(X= k) = •pk•qn-k .

• Отрицательное биномиальное распределение (распределение Паскаля)

Пусть Y – число посетителей, достаточное для того, чтобы k из них оказались покупателями. Тогда вероятность того, что n–й посетитель окажется k–м покупателем составит P(Y=n)=•pk•qn–k.

• Геометрическое распределение

Если **Y** – число посетителей, достаточное для того, чтобы один из них оказался

покупателем, то P(Y=1)= p•qn–1.

• Распределение Пуассона

Если ваш магазин посещают довольно часто, но при этом весьма редко делают покупки, то вероятность k покупок в течение большого интервала времени, (например, – дня) составит P(Z=k) = λk •Exp(-λ) / k! , где λ – особый показатель распределения, так называемый его параметр.

## Односторонние и двухсторонние значения вероятностей

Если нам известен закон распределения СВ (пусть – дискретной), то в этом случае очень часто приходится решать задачи, по крайней мере, трех стандартных типов:

• какова вероятность того, что случайная величина X окажется равной (или наоборот – не равной) некоторому значению, например – Xk ?

• какова вероятность того, что случайная величина **X** окажется больше (или наоборот – меньше) некоторого значения, например –Xk ?

• какова вероятность того, что случайная величина X окажется не меньше Xi и при этом не больше Xk ?

Первую вероятность иногда называют "точечной", ее можно найти из закона распределения, но только для дискретной случайной величины. Разумеется, что вероятность равенства задана самим законом распределения, а вероятность неравенства составляет

P(X#Xk) = 1 – P(X=Xk).

Вторую вероятность принято называть "односторонней". Вычислять ее также достаточно просто – как сумму вероятностей всех допустимых значений, равных и меньших Xk . Для примера "открытого" нами закона биномиального распределения при p=0.5 и m=4 одностороння вероятность того, что X окажется менее 3 (т.е.0, 1 или 2), составит точно 0.0625+0.25+0.375=0.6875.

Вероятность третьего типа называют "двухсторонней" и вычисляют как сумму вероятностей значений X внутри заданного интервала. Для предыдущего примера вероятность того, что X менее 4 и более 1 составит 0.375+0.25=0.625.

Односторонняя и двухсторонняя вероятности являются универсальными понятиями – они применимы как для дискретных, так и для непрерывных случайных величин.

## Моменты распределений дискретных случайных величин.

Итак, закон распределения вероятностей дискретной СВ несет в себе всю информацию о ней и большего желать не приходится.

Не будет лишним помнить, что этот закон (или просто – распределение случайной величины) можно задать тремя способами:

• в виде формулы: например, для биномиального распределения при n=3 и p=0.5 вероятность значения суммы S=2 составляет 0.375;

• в виде таблицы значений величины и соответствующих им вероятностей:

• в виде диаграммы или, как ее иногда называют, гистограммы распределения:

Таблица 2–1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Сумма | 0 | 1 | 2 | 3 |
| Вероятность | 0.125 | 0.375 | 0.375 | 0.125 |

Рис. 2–1 Гистограмма распределения



Необходимость рассматривать вопрос, поставленный в заглавии параграфа, не так уж и очевидна, поскольку непонятно, что же еще нам надо знать?

Между тем, все достаточно просто. Пусть, для какого–то реального явления или процесса мы сделали допущение (выдвинули гипотезу), что соответствующая СВ принимает свои значения в соответствии с некоторой схемой событий. Рассчитать вероятности по принятой нами схеме ­– не проблема!

Вопрос заключается в другом – как проверить свое допущение или, на языке статистики, оценить достоверность гипотезы?

По сути дела, кроме обычного наблюдения за этой СВ у нас нет иного способа выполнить такую проверку. И потом – в силу самой природы СВ мы не можем надеяться, что через достаточно небольшое число наблюдений их частоты превратятся в “теоретические” значения, в вероятности. Короче – результат наблюдения над случайной величиной тоже … случайная величина или, точнее, – множество случайных величин.

Так или примерно так рассуждали первые статистики–профессионалы. И у кого–то из них возникла простая идея: сжать информацию о результатах наблюдений до одного, единственного показателя!

Как правило, простые идеи оказываются предельно эффективными, поэтому способ оценки итогов наблюдений по одному, желательно “главному”, “центральному” показателю пережил все века становления прикладной статистики и по ходу дела обрастал как теоретическими обоснованиями, так и практическими приемами использования.

Вернемся к гистограмме рис. 2–1 и обратим внимание на два, бросающихся в глаза факта:

• “наиболее вероятными” являются значения суммы S=1 и S=2 и эти же значения лежат “посредине” картинки;

• вероятность того, что сумма окажется равной 0 или 1, точно такая же, как и вероятность 2 или 3, причем это значение вероятности составляет точно 50 %.

Напрашивается простой вопрос – если СВ может принимать значения 0, 1, 2 или 3, то сколько в среднем составляет ее значение или, иначе – что мы ожидаем, наблюдая за этой величиной?

Ответ на такой вопрос на языке математической статистики состоит в следующем. Если нам известен закон распределения, то, просуммировав произведения значений суммы S на соответствующие каждому значению вероятности, мы найдем математическое ожидание этой суммы как дискретной случайной величины –

M(S) = Σ S i •P(S i). {2–3}

В рассматриваемом нами ранее примере биномиального распределения, при значении p=0.5, математическое ожидание составит

M(S) = 0•0.125+1•0.375+2•0.375+3•0.125= 1.5 .

Обратим внимание на то, что математическое ожидание дискретной величины типа Int или Rel совсем не обязательно принадлежит к множеству допустимых ее значений. Что касается СВ типа Nom или Ord, то для них понятие математического ожидания (по закону распределения), конечно же, не имеет смысла. Но так как с номинальной, так и с порядковой шкалой дискретных СВ приходится иметь дело довольно часто, то в этих случаях прикладная статистика предлагает особые, непараметрические методы.

Продолжим исследование свойств математического ожидания и попробуем в условиях нашего примера вместо S рассматривать U= S – M(S). Такая замена СВ (ее часто называют центрированием) вполне корректна: по величине U всегда можно однозначно определить S и наоборот.

Если теперь попробовать найти математическое ожидание новой (не обязательно дискретной) величины M(U) , то оно окажется равным нулю, независимо от того считаем ли мы конкретный пример или рассматриваем такую замену в общем виде.

Мы обнаружили самое важное свойство математического ожидания – оно является “центром” распределения. Правда, речь идет вовсе не о делении оси допустимых значений самой СВ на две равные части. Поистине – первый показатель закона распределения “самый главный” или, на языке статистики, – центральный.

Итак, для СВ с числовым описанием математическое ожидание имеет достаточно простой смысл и легко вычисляется по законам распределения. Заметим также, что математическое ожидание – просто числовая величина (в общем случае не дискретная, а непрерывная) и никак нельзя считать ее случайной.

Другое дело, что эта величина зависит от внутренних параметров распределения (например, – значения вероятности р числа испытаний n биномиальном законе).

Так для приведенных выше примеров дискретных распределений математическое ожидание составляет:

|  |  |
| --- | --- |
| Тип распределения | Математическое ожидание |
| Биномиальное | n•p |
| Распределение Паскаля | k •q / p |
| Геометрическое распределение | q / p |
| Распределение Пуассона | λ |

Возникает вопрос – так что же еще надо? Ответ на этот вопрос можно получить как из теории, так и из практики.

Один из разделов кибернетики – теория информации (курс “Основы теории информационных систем” у нас впереди) в качестве основного положения утверждает, что всякая свертка информации приводит к ее потере. Уже это обстоятельство не позволяет допустить использование только одного показателя распределения СВ – ее математического ожидания.

Практика подтверждает это. Пусть мы построили (или использовали готовые) законы распределения двух случайных величин X и Y и получили следующие результаты:

Таблица 2–2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Значения | 1 | 2 | 3 | 4 |
| P(X) % | 12 | 38 | 38 | 12 |
| P(Y) % | 30 | 20 | 20 | 30 |



Y

Y

Y

Y

Рис. 2–2

Простое рассмотрение табл.2–2 или соответствующих гистограмм рис.2–2 приводит к выводу о равенстве M(X) = M(Y) = 0.5 , но вместе с тем столь же очевидно, что величина X является заметно “менее случайной”, чем Y.

Приходится признать, что математическое ожидание является удобным, легко вычислимым, но весьма неполным способом описания закона распределения. И поэтому требуется еще как–то использовать полную информацию о случайной величине, свернуть эту информацию каким–то иным способом.

Обратим внимание, что большие отклонения от M(X) у величины X маловероятны, а у величины Y – наоборот. Но при вычислении математического ожидания мы, по сути дела “усредняем” именно отклонения от среднего, с учетом их знаков. Стоит только “погасить” компенсацию отклонений разных знаков и сразу же первая СВ действительно будет иметь показатель разброса данных меньше, чем у второй. Именно такую компенсацию мы получим, усредняя не сами отклонения от среднего, а квадраты этих отклонений.

Соответствующую величину

D(X) = Σ (X i – M(X))2 • P(X i); {2–4} принято называть дисперсией распределения дискретной СВ.

Ясно, что для величин, имеющих единицу измерения, размерность математического ожидания и дисперсии оказываются разными. Поэтому намного удобнее оценивать отклонения СВ от центра распределения не дисперсией, а квадратным корнем из нее – так называемым среднеквадратичным отклонением σ, т.е. полагать

σ2 = D(X).  {2–5}

Теперь оба параметра распределения (его центр и мера разброса) имеют одну размерность, что весьма удобно для анализа.

Отметим также, что формулу {2–3} часто заменяют более удобной

D(X) = Σ (X i)2 •P(X i) – M(X)2. {2–6}

Весьма полезно будет рассмотреть вопрос о предельных значениях дисперсии.

Подобный вопрос был бы неуместен по отношению к математическому ожиданию ­– мало ли какие значения может иметь дискретная СВ, да еще и со шкалой Int или Rel.

Но дословный перевод с латыни слова “дисперсия” означает “рассеяние”, “разброс” и поэтому можно попытаться выяснить – чему равна дисперсия наиболее или наименее “разбросанной” СВ? Скорее всего, наибольший разброс значений (относительно среднего) будет иметь дискретная случайная величина X, у которой все n допустимых значений имеют одну и ту же вероятность 1/n. Примем для удобства Xmin и Xmax (пределы изменения данной величины), равными 1 и n соответственно.

Математическое ожидание такой, равномерно распределенной случайной величины составит M(X) = (n+1)/2 и остается вычислить дисперсию, которая оказывается равной D(X) = Σ (Xi)2/n – (n+1)2/4= (n2–1)/ 12.

Можно доказать, что это наибольшее значение дисперсии для дискретной СВ со шкалой Int или Rel .

Последнее выражение позволяет легко убедиться, что при n =1 дисперсия оказывается равной нулю ­ – ничего удивительного: в этом случае мы имеем дело с детерминированной, неслучайной величиной.

Дисперсия, как и среднеквадратичное отклонение для конкретного закона распределения являются просто числами, в полном смысле показателями этого закона.

Полезно познакомиться с соотношениями математических ожиданий и дисперсий для упомянутых ранее стандартных распределений:

Таблица 2–3

|  |  |  |  |
| --- | --- | --- | --- |
| Тип  распределения | Математическое ожидание | Дисперсия | Коэффициент  вариации |
| Биномиальное | np | npq | Sqrt(q/n•p) |
| Паскаля | kq/p | kq/p2 | Sqrt(1/ kq) |
| Геометрическое | q/p | q/p2 | Sqrt(1/q) |
| Пуассона | λ | λ | Sqrt(1/λ) |

Можно ли предложить ещё один или несколько показателей – сжатых описаний распределения дискретной СВ? Разумеется, можно.

Первый показатель (математическое ожидание) и второй (дисперсия) чаще всего называют моментами распределения. Это связано со способами вычисления этих параметров по известному закону распределения – через усреднение значений самой СВ или усреднение квадратов ее значений.

Конечно, можно усреднять и кубы значений, и их четвертые степени и т.д., но что мы при этом получим? Поищем в теории ответ и на эти вопросы.

Начальными моментами k-го порядка случайной величины X обычно называют суммы:

νk = Σ(X i)k• P(X i); ν0 = 0; {2–7}

а центральными моментами – суммы:

μk= Σ (X i –ν1)k• P(X i), {2–8} при вычислении которых усредняются отклонения от центра распределения – математического ожидания.

Таким образом,

• μ1 = 0;

• ν1 = M(X) является параметром центра распределения;

• μ2 = D(X)  является параметром рассеяния; {2-9}

• ν3 и μ3 – описывают асимметрию распределения;.

• ν4 и μ4 – описывают т.н. эксцесс (выброс) распределения и т.д.

Иногда используют еще один показатель степени разброса СВ – коэффициент вариации V= σ/ M(X), имеющий смысл при ненулевом значении математического ожидания.

## Распределения непрерывных случайных величин

До этого момента мы ограничивались только одной “разновидностью” СВ – дискретными, т.е. принимающими конечные, заранее оговоренные значения на любой из шкал Nom, Ord, Int или Rel .

Но теория и практика статистики требуют использовать понятие непрерывной СВ ­– допускающей любые числовые значения на шкале типа Int или Rel . И дело здесь вовсе не в том, что физические величины теоретически могут принимать любые значения – в конце концов, мы всегда ограничены точностью приборов их измерения. Причина в другом…

Математическое ожидание, дисперсия и другие параметры любых СВ практически всегда вычисляются по формулам, вытекающим из закона распределения. Это всего лишь числа и далеко не всегда целые.

Так обстоит дело в теории. На практике же, мы имеем только одно – ряд наблюдений над случайной (будем далее полагать – всегда дискретной) величиной. По этим наблюдениям можно строить таблицы или гистограммы, используя значения соответствующих частот (вместо вероятностей). Такие распределения принято называть выборочными, а сам набор данных наблюдений – выборкой.

Пусть мы имеем такое выборочное распределение некоторой случайной величины X – т.е. для ряда ее значений (вполне возможно неполного, с “пропусками" некоторых допустимых) у нас есть рассчитанные нами же частоты f i .

В большинстве случаев нам неизвестен закон распределения СВ или о его природе у нас имеются догадки, предположения, гипотезы, но значения параметров и моментов (а это неслучайные величины!) нам неизвестны.

Разумеется, частоты fi суть непрерывные СВ и, кроме первой проблемы ­– оценки распределения X, мы имеем ещё одну ­– проблему оценки распределения частот.

Существование закона больших чисел, доказанность центральной предельной теоремы поможет нам мало:

• во-первых, надо иметь достаточно много наблюдений (чтобы частоты “совпали” с вероятностями), а это всегда дорого;

• во-вторых, чаще всего у нас нет никаких гарантий в том, что условия наблюдения остаются неизменными, т.е. мы наблюдаем за независимой случайной величиной.

Теория статистики дает ключ к решению подобных проблем, предлагает методы “работы” со случайными величинами. Большинство этих методов появилось на свет как раз благодаря теоретическим исследованиям распределений непрерывных величин.

### Нормальное распределение

Первым, фундаментальным по значимости, является т.н. нормальный закон распределения непрерывной случайной величины X, для которой допустимым является любое действительное числовое значение. Доказано, что такой закон распределения имеет величина, значение которой обусловлено достаточно большим количеством факторов (причин).

Для вычисления вероятности того, что X лежит в заранее заданном диапазоне, получено выражение, которое называют интегралом вероятности:

P(a ≤ X ≤ b) = 

Обратим внимание на то, что в это выражение входят две константы (параметра) μ и σ. Как и для любой (не обязательно дискретной) СВ, здесь также имеют смысл понятия моментов распределения и оказывается, что

M(X) = μ , а D(x) = σ2 .  {2–10}

Для непрерывно распределенных величин не существует понятия вероятности конкретного значения. Вопрос ­– “какова вероятность достижения температурой воздуха значения 14 градусов?” – некорректен. Все зависит от прибора измерения, его чувствительности, ошибок измерения. Но вместе с тем функция под интегралом вероятности существует, она однозначно определена:

ϕ(X) = ,

ее график (аналог гистограммы) имеет вид:

ϕ (X)

99.73 %

μ – 3σ μ μ+3σ

а площадь под кривой на заданном интервале X определяет вероятность попадания в этот интервал.

Чаще всего закон нормального распределения используется для нормированной случайной величины

Z = (X – μ) /σ, {2–11} у которой M(Z)=0; D(Z)=1.  {2–12}

Отметим ряд других особенностей этого распределения, полагая его нормированным.

• Доказано, что целый ряд “классических” распределений (как дискретных, так и непрерывных) стремятся к нормальному при непрерывном изменении их внутренних параметров.

• Симметрия нормального распределения позволяет достаточно просто оценивать вероятность “попадания” случайной нормированной величины в заданный диапазон. Очень часто в прикладной статистике приходится использовать понятие “маловероятного” значения. Для нормированной величины с нормальным распределением вероятность попадания в диапазон ± 3σ составляет 0.9973 (правило “трех сигм”).

• Особую роль играет нормальное распределение при решении вопросов о “представительности” наблюдений. Оказывается, что работа с выборочными распределениями в большинстве случаев позволяет решить проблему оценки наших предварительных выводов, предположений, гипотез – с использованием разработанных и теоретически обоснованных приемов на базе нормального закона.

### *Распределения выборочных значений параметров нормального распределения*

Пусть у нас имеется некоторая непрерывная случайная величина **X ,** распределенная нормально с математическим ожиданием μ и среднеквадратичным отклонением σ. Если мы имеем n наблюдений над такой величиной (имеем выборку объемом n из генеральной совокупности) , то выборочные значения Mx иSx  являются также случайными величинами и нам крайне важно знать их законы распределения. Это необходимо как для оценки доверия к этим показателям, так и для проверки принадлежности исходного распределения к нормальному. Существует ряд теоретически обоснованных выводов по этой проблеме:

• величина  имеет нормированное нормальное распределение, что позволяет оценивать Mx при заранее известной дисперсии;

• величина  имеет так называемое распределение Стьюдента, для которого также имеется выражение плотности вероятности и построены таблицы;

• величина  имеет распределение "хи–квадрат", также с аналитической функцией плотности и рассчитанными по ней таблицами.

Отметим, что распределения Стьюдента и "хи–квадрат" имеют свой внутренний параметр, который принято называть числом степеней свободы. Этот параметр полностью определяется объемом выборки (численностью наблюдений) и выбирается обычно равным m =(n – 1).

# Взаимосвязи случайных величин

## Парная корреляция

Прямое толкование термина "корреляция" — стохастическая, вероятная, возможная связь между двумя (парная) или несколькими (множественная) случайными величинами.

Выше говорилось о том, что если для двух случайных величин X и Y имеет место равенство P(X ∩Y) = P(X)•P(Y), то эти величины считаются независимыми. Ну, а если это не так!?

Ведь всегда важно знать: насколько зависит одна СВ от другой? Дело не только в присущем людям стремлении анализировать что-либо обязательно в числовом измерении. Уже понятно, что прикладная статистика требует непрерывных вычислений, что использование компьютера вынуждает нас работать с числами, а не с понятиями.

Для числовой оценки взаимосвязи между двумя СВ: Y – с известными M(Y) и σy

и X – с M(X) и σx принято использовать так называемый коэффициент корреляции

. {3–1}

Обратим внимание на способ вычисления коэффициента корреляции. В числителе находится математическое ожидание произведения отклонений величин X и Y от собственных математических ожиданий.

Этот коэффициент может принимать значения от –1 до +1 — в зависимости от тесноты и характера связи между данными СВ.

Если коэффициент корреляции равен нулю, то X и Y называют некоррелированными. Считать их независимыми обычно нет оснований — оказывается, что существуют такие, как правило — нелинейные связи величин, при которых коэффициент корреляции равен нулю, хотя величины зависят друг от друга.

Обратное всегда верно — если величины независимы, то R(XY) = 0. Но, если модуль R(XY) равен 1, то есть все основания предполагать наличие линейной связи между Y и X. Именно поэтому часто говорят о линейной корреляции при использовании такого способа оценки связи между СВ.

Если у нас имеется ряд наблюдений за двумя случайными величинами, то можно оценить выборочное значение коэффициента корреляции –

 {3–2}

Оценку корреляционной связи двух СВ можно производить и без учета их дисперсий.

Числитель коэффициента корреляции

. {3–3}

называют ковариацией случайных величин, которая также служит мерой связи, но без непосредственного учета дисперсий.

Различие между такими двумя показателями парной связи СВ достаточно существенное.

• Коэффициент корреляции определяет степень, тесноту линейной связи между величинами и является безразмерной величиной.

• Ковариация двух СВ определяет эту связь безотносительно к ее виду и является величиной размерной.

## Множественная корреляция

В ряде случаев статистического анализа приходится решать вопрос о связях нескольких (более 2) СВ или вопрос о множественной корреляции.

Пусть X, Y и Z – случайные величины, имеющие математические ожидания M(X), M(Y), M(Z) и среднеквадратичные отклонения σx ,σy, σz соответственно. Тогда можно найти парные коэффициенты корреляции Rxy, Rxz, Ryz по приведенной выше формуле.

Но этого явно недостаточно – ведь мы на каждом из трех этапов попросту забывали о наличии третьей СВ! Поэтому в случаях множественного корреляционного анализа иногда требуется отыскивать т. н. частные коэффициенты корреляции — например, оценка виляния Z на связь между X и Y производится с помощью коэффициента



И, наконец, можно поставить вопрос — а какова связь между данной СВ и совокупностью остальных? Ответ на такие вопросы дают коэффициенты множественной корреляции RX.YZ, RY.XZ, RZ.XY, формулы для вычисления которых построены по тем же принципам — учету связи одной из величин со всеми остальными в совокупности.

# Проверка статистических гипотез

## Понятие статистической гипотезы

Как уже отмечалось, основным занятием статистика–прикладника является чаще всего решение вопроса о том, что и как можно извлечь из наблюдений над случайной величиной (выборочных её значений) для последующего использования в практике.

Скажем, для некоторой экономической задачи требуется знание длины очереди автомашин, ожидающих технического обслуживания, а эта величина хоть и выражается целым числом, но является случайной.

Очень редко задачи такого рода имеют “теоретическую платформу” – хотя бы в части закона распределения СВ, не говоря уже о внутренних параметрах этого распределения или его моментах. Чаще всего в нашем распоряжении нет практически ничего, кроме некоторого количества наблюдений за значениями СВ и … необходимости решать задачу.

Выражаясь чисто научным языком, современный подход к статистическим задачам в последние два десятилетия заключается в использовании непараметрической статистики, а не традиционных, классических методов, которые применимы только при заранее известных законах распределений.

Но и в первом, и во втором случаях одной из важнейших задач профессионального статистика является проверка выдвинутых им же предположений или гипотез.

Чем же отличаются статистические гипотезы от обычных, житейских предположений? Прежде всего, тем, что статистических гипотез всегда две и они взаимоисключающие. Одна из них (обычно та, которую предполагают отклонить) носит название нулевой гипотезы **Њ0,** вторая – альтернативная гипотеза **Њ1**  всегда отрицает нулевую, противостоит ей.

Вся “хитрость” заключается именно в нулевой гипотезе – её надо построить, сформулировать так, чтобы иметь возможность найти интересующие нас вероятности в условиях истинности этой гипотезы.

Пусть мы исследуем игральную кость ­– “проверяем” ее симметричность. Ясно, что в качестве нулевой гипотезой надо считать предположение о полной симметрии кости.

Ведь если **Њ0** верна, то вероятности выпадения всех шести цифр на гранях будут одинаковы – по 1/6 . А вот выдвижение в качестве нулевой гипотезы предположения об асимметрии кости ничего бы не дало – в этом случае мы ничего не можем сказать о значениях вероятностях выпадения цифр.

С процедурами проверки статистических гипотез неразрывно связано еще одно, непривычное для обычных расчетных работ, понятие уровня значимости результатов наблюдений.

В самом начале курса уже упоминался метод выделения редких событий ­– вероятность которых не превышает 5 %. Конечно, это значение является чисто условным – в некоторых случаях редкими считают события с вероятностью не более 1 %.

Теория вероятностей позволяет обосновать деление случайных событий на три класса ­– обычные, редкие и исключительные. При этом наблюдение события исключительного дает основания считать, что причины его наступления являются уже неслучайными – имеет место влияние некоторого фактора.

Будем далее использовать 5 % уровень значимости, как это принято почти во всех прикладных направлениях статистики, в том числе и в экономике.

Итак, если наблюдения относятся к событиям редким (с вероятностью до 5 %), то такие наблюдения и результаты их обработки будем называть статистически значимыми. Как же так, спросите вы, ­– вероятность мала, а предлагается считаться с ней. Все очень просто ­– если мы вычислили вероятность некоторого результата наблюдения в условиях основной гипотезы и она (априорная вероятность) оказалась очень малой, то чем она меньше, тем больше у нас оснований отвергнуть Њ0. С другой стороны, если мы увидели очень редкое событие – выпадение 10 гербов при 15 подбрасываниях монетки, то значимость такого наблюдения чрезвычайно высока – гипотезу о симметрии вполне можно отбросить.

## Критерии статистических гипотез

Если мы пытаемся решить некоторую статистическую задачу, то в большинстве случаев нам придется заниматься не столько математическими выкладками и числовыми расчетами, сколько принимать решение – какую из выдвинутых нами же статистических гипотез принять (или – какую из них отвергнуть).

Так вот, решающее правило, согласно которому мы будем действовать, принято называть статистическим критерием. К сожалению, не существует единого, универсального критерия значимости ­– их приходится разрабатывать в теории и использовать на практике применительно к особенностям конкретных задач.

Вместе с тем, любому критерию значимости присуще одно и то же свойство – во всех случаях мы не получим категоричного указания на “истинную” гипотезу, прямого ответа на вопрос ­– какую из гипотез нам принять.

Еще более непривычным для человека с навыками искать и находить ответы в расчетных задачах, будет сама форма ответа на вопрос о сравнении гипотез Њ0 и Њ1 – например, в таком виде "если отбросить нулевую гипотезу, то вероятность ошибки такого действия не превосходит 3 % ".

Дальше уже наше дело, принять или отвергнуть ту или иную гипотезу – теория большего дать не в состоянии. Надо понять различие между выделенным утверждением и вроде бы аналогичным ­– "вероятность верности гипотезы Њ1 составляет 97%" . Все между тем очень просто – вычислить возможно только вероятность ошибочности Њ0 и не более того!

Пусть мы интересуемся симметрией обычной монетки и собираемся проводить эксперименты ­– подбрасывать её и фиксировать результаты. Выдвинем гипотезу – монета симметрична. Если мы собираемся произвести N подбрасываний и по их итогам проверить гипотезу, должны просчитать вероятности выпадения 0, 1, 2 и т.д. до N “гербов”. Конечно, можно выполнить расчеты и после окончания опыта ­– всё равно это будут априорные вероятности по своей сути.

Проиллюстрируем это на рассмотренной ранее ситуации 8 экспериментов с монеткой. Предположим, что частости появления возможных исходов уже вычислены – в таких случаях говорят о наличии выборочного распределения вероятностей. Для нашего эксперимента такое распределение имеет вид:

Таблица 4–1

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Число наблюдений гербов k | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Вероятность P(X =k) в 1 / 256 | 1 | 8 | 28 | 56 | 70 | 56 | 28 | 8 | 1 |
| Вероятность P(X ≤k) в 1 / 256 | 1 | 9 | 37 | 93 | 163 | 219 | 247 | 255 | 256 |

Если мы в результате эксперимента получили сумму гербов S = 1, то вероятность наблюдать такую сумму (и менее вероятное значение S=0) составляет для симметричной монетки P(S<2) = (1+8) / 256 ≅ 0.036. Можно, однако, рассуждать и иначе. Ведь мы наблюдали в том же опыте 7 появлений “решки”. Вероятность наблюдать такое и менее вероятное число 8 составляет точно столько же – P(S>6) = (1+8) / 256 ≅ 0.036. Осталось построить решающее правило ­– критерий для принятия окончательного решения в отношении выдвинутых гипотез (основной Њ0 и альтернативной Њ1).

Заметим, что при выдвинутой нами основной гипотезе Њ0:(p=q) альтернативную гипотезу можно выдвигать по разному:

Њ1: (p#q)– монета несимметрична, ненаправленная гипотеза, требующая использования двухсторонних вероятностей;

Њ1: (p<q)– монета несимметрична и при этом “герб” легче, направленная гипотеза, достаточно односторонних вероятностей.

Применим оба приема построения критерия в условиях нашего примера.

• Нулевая гипотеза Њ0: (p=q). Альтернативная гипотеза Њ1: (p#q).

Уровень значимости α=0.05. Итог наблюдений при N=8: S= 1 .

Вероятность такого итога при условии, что нулевая гипотеза верна составляет

P(S<2)+P(S>6) ≅ 0.072, т.е. больше порогового значения

Решение: нулевую гипотезу не отвергаем, монетку считаем симметричной.

• Нулевая гипотеза Њ0: (p=q). Альтернативная гипотеза Њ1: (p<q).

Уровень значимости α=0.05. Итог наблюдений при N=8: S= 1 .

Вероятность такого итога при условии, что нулевая гипотеза верна составляет P(S<2) ≅ 0.036, т.е. меньше уровня значимости.

Решение**:** нулевую гипотезу отвергаем, монетку считаем направленно несимметричной.

Возможно у вас возникло сомнение в части первого способа оценки статистических гипотез – ведь герб наблюдался всего один раз из восьми и, тем не менее, гипотеза о симметрии монетки не отбрасывается.

На самом деле всё правильно и обосновано ­– смысл нулевых гипотез Њ0 в первом и втором случае, несмотря на формальную тождественность, не одинакова. Суть дела заключена в формулировке альтернативных гипотез Њ1.

В первом случае Њ1 охватывает два события (p>q) или (p<q), а значит это более жесткое предположение. Во втором случае Њ1 связана только с одним событием (p<q), а значит она мягче, требует меньшего количества информации для признания ее истинной.

## Ошибки при проверке статистических гипотез

Выбирая окончательно в качестве рабочей одну из гипотез ­– нулевую или альтернативную, мы используем следующую логическую схему (алгоритм):

Рис.4–1

Њ0:

Неверна

Верна

Принята

Принята

Отброшена

Отброшена

Нет ошибки

Ошибка

2 рода

Ошибка

1 рода

Не забудем, что отвергая Њ0, мы принимаем альтернативную Њ1 и наоборот. Пусть у нас уже есть правило, в соответствии с которым мы либо принимаем основную гипотезу Њ0,  либо отвергаем её.

Как уже говорилось, контрольной цифрой является уровень значимости – вероятность α наблюдать то, что мы имеем после эксперимента, в случае если гипотеза Њ0 верна.

Пусть, к примеру, мы знаем вероятность данного наблюдения при истинности основной гипотезы и она равна 0.04. Мы вправе принять эту гипотезу – вероятность ошибиться меньше, чем α=0.05.

Конечно, приняв нулевую гипотезу, мы рискуем ошибиться. Степень риска можно найти очень просто – вероятность отбросить верную нулевую гипотезу (совершить ошибку первого рода или **α**–ошибку) составляет 5 %.

Но ведь можно совершить и другую ошибку ­– принять нулевую гипотезу, когда она на самом деле неверна (ошибка второго рода или **β**–ошибка). Величина эта зависит, прежде всего, от решающего правила ­– критерия принятия гипотез. Поэтому величину (1 –**β**) принято называть мощностью критерия.

С определением вероятности ошибки второго рода дело обстоит не так просто – ее приходится вычислять. В первом приближении можно считать, что нам одинаково “вредны” ошибки как первого, так и второго рода. Более актуальным является вопрос ­ – а как их избежать или хотя бы снизить вероятность их появления? К сожалению, в задачи курса не входит рассмотрение таких вопросов.

Достаточно знать, что в прикладной статистике существуют методы повышения эффективности критериев проверки статистических гипотез.

Кроме того, нельзя упускать из виду и "простой рецепт" снижения вероятностей ошибок как первого, так и второго рода – надо иметь побольше наблюдений.

Так, например, имеются достаточно надежные методы определения так называемых “критических” значений СВ. Эти значения для задач рассмотренных выше типов (с биномиальным распределением вероятностей) позволяют сразу же оценить возможность отбрасывания нулевой гипотезы ­– по данным о числе испытаний и числе наблюдений данного события.

Если число испытаний монетки на симметрию составляет N=12 и выдвинуты гипотезы Њ0: (p=q); Њ1: (p#q), то критическими значениями наблюдений при граничной вероятности α=0.05 являются S=2 и S=10. Это означает, что при наблюдаемом числе гербов ≤ 2 или ≥ 10 нулевая гипотеза может быть отвергнута.

Обратим также внимание на явную зависимость наших решений от числа наблюдений – нам не удалось отвергнуть гипотезу о симметрии монетки при всего одном гербе (из восьми бросаний), но вполне обосновано удается сделать это при 0, 1 и даже 2 – при увеличении числа наблюдении или, на языке статистики, увеличении объема выборки.

# Выборочные распределения на шкалах Int и Rel

## Оценка наблюдений при неизвестном законе распределения

Какова цель наблюдений над случайной величиной; для чего используются результаты наблюдений; где, как и для чего применить возможности теории вероятностей и прикладной статистики? Ответы на эти, простые с виду, вопросы зависят от многих факторов, обстоятельств и не всегда оказываются конкретными.

Попытаемся всё же сформулировать ответ применительно к конкретной обстановке – при статистических расчетах в экономических системах.

В таких системах основные числовые показатели “жизни” системы в целом и отдельных её элементов можно свести к трем разновидностям:

• продукция, с конкретными ее показателями (вес, объем, количество и т.д.), ­– величинами на шкале Int или Rel;

• деньги, с единицей измерения по шкале Int или Rel (отрицательные величины обычно означают убытки или долги);

• информация, с несколькими шкалами измерений – в битах (байтах) для количественного описания по шкале Int или в виде сообщений о событиях на шкалах Nom или Ord.

Простые размышления приводят к мысли о возможности допустить, что все эти величины являются, во-первых, случайными и, во-вторых, дискретными. Ясно также, что без учета всех этих величин эффективной экономики быть не может – только знание всех этих показателей позволит управлять экономикой.

Конечно, у многих из вас уже готово решение проблемы – раз уж мы не знаем точно значение величины (скажем – суммы прибыли), так воспользуемся её математическим ожиданием! Это верная мысль…

Но для вычисления математического ожидания надо знать закон распределения вероятностей, т.е. иметь информацию

• обо всех допустимых (возможных) значениях прибыли;

• о соответствующих им значениях вероятностей.

Рассмотрим простейший пример. Пусть у нас есть всего четыре наблюдения над суммой G дневной выручки в 196, 208, 210 и 214 гривен. Легко подсчитать среднее значение ­– оно составит 207 гривен. Какое доверие к этой цифре? Ведь мы совершенно ничего не знаем о законе распределения СВ, кроме того, что эта величина дискретная и имеет относительную шкалу. Тем не менее, кое–что полезное из таких скудных наблюдений (малой выборки) можно извлечь.

Поступим следующим образом – вместо случайной величины G будем рассматривать другую величину U= (G–M(G)). Математическое ожидание новой СВ будет всегда равно нулю – какие бы гипотезы о значении M(G) мы ни выдвигали,

Теперь подумаем о том, как сформулировать нулевую гипотезу. Вроде бы это надо делать так:

Њ0: дневная выручка имеет некоторый закон распределения

с математическим ожиданием в M(G)=207 гривен.

Теперь результаты наблюдений над выручкой G можно представить в виде четырех наблюдений над U: –11,+1,+3,+7. Теория математической статистики предлагает следующий, т.н. биномиальный критерий проверки гипотез в подобных ситуациях.

Предполагается, что распределение вероятностей наблюдаемой величины U симметрично относительно значения математического ожидания, т.е. относительно нуля.

Далее предлагается рассматривать N имеющихся у нас значений U как совокупность случайных величин, принимающих с вероятностью 0.5 значения по итогам наблюдения или противоположные им по знаку. В нашем примере это приводит к

P(U1=11)=P(U1= –11)= 1/ 2; P(U2=1)=P(U2= –1)= 1/ 2;

P(U3=3)=P(U3= –3)= 1/ 2; P(U4=7)=P(U4= –7)= 1/ 2;

Теперь рассматривается сумма этих случайных величин S – она может принимать 2N различных значений, с одинаковой вероятностью 1/2N.

Таблица 5-1

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| U1 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | -11 | -11 | -11 | -11 | -11 | -11 | -11 | -11 |
| U2 | 1 | 1 | 1 | 1 | -1 | -1 | -1 | -1 | 1 | 1 | 1 | 1 | -1 | -1 | -1 | -1 |
| U3 | 3 | 3 | -3 | -3 | 3 | 3 | -3 | -3 | 3 | 3 | -3 | -3 | 3 | 3 | -3 | -3 |
| U4 | -7 | 7 | 7 | -7 | 7 | -7 | 7 | -7 | 7 | -7 | 7 | -7 | 7 | -7 | 7 | -7 |
| S | 8 | 22 | 16 | 2 | 20 | 6 | 14 | 0 | 0 | -14 | -6 | -20 | -2 | -16 | -8 | -22 |

Отклонения от гипотетического математического ожидания в сумме составляют в нашем примере точно 0 и нам необходимо определить количество вариантов, в которых сумма **S>**0. Всего вариантов 16, а вариантов с нулевой или положительной суммой 9. Вероятность ошибки при отклонении Њ0оказалась равной 9/16≅0.57, что намного больше контрольных 5 % . Как и следовало ожидать, нам нет смысла отбрасывать нулевую гипотезу – слишком велика ошибка первого рода.

Все было бы хорошо, но если мы выдвинем другую нулевую гипотезу о математическом ожидании выручки, например – Њ0: M(G)= 196 гривен, то после аналогичных расчетов получим результат ­– и эту гипотезу нет оснований отбрасывать, правда вероятность ошибки первого рода теперь будет иной – “всего лишь” 0.125. Столько же составит вероятность этой ошибки и при Њ0: M(G)= 214. Таким образом, все нулевые гипотезы со значениями от 196 до 214 можно не отвергать (не достигнуто пороговое значение 0.05). Можно ли рекомендовать принятие альтернативной гипотезы и, если – да, то при каком значении гипотетического математического ожидания?

Теория прикладной статистики отвечает на этот вопрос однозначно ­– нет, рекомендовать нам это она не вправе!

Вспомним “неудобное” свойство статистических выводов или рекомендаций – они никогда не бывают однозначными, конкретными. Поэтому наивно ожидать решения задачи об оценке математического ожидания по данным наблюдений в виде одного, конкретного числа.

Еще раз продумаем, чего мы добиваемся, меняя значение в нулевой гипотезе? Ведь самая большая ошибка первого рода была как раз тогда, когда мы выдвинули такое понятное предположение ­– математическое ожидание равно среднему.

Более того, проверка нулевой гипотезы такого вида была совершенно бессмысленным делом. Практически всегда в этих случаях альтернативная гипотеза окажется самой вероятной, но практически никогда вероятность ее истинности не достигнет желанных 95 %.

Всё дело в том, что просчитать последствия своего решения мы умеем только отвергая нулевую гипотезу, но, принимая ее, последствия просчитать не можем.

Вот если бы, передвигая воображаемый указатель по шкале СВ мы получили сигнал “СТОП, достаточно! Достигнут уровень ошибки 5 %”, то мы бы запомнили данное значение как левую (или правую) границу интервала, в котором почти “наверняка” лежит искомое нами математическое ожидание. В нашем примере этого не произошло и, оказывается и не могло произойти.

Дело в том, что у нас всего 4 наблюдения (196,208,210,214) со средним значением 207 и среднеквадратичным отклонением около 13.5 гривен (т.е. более 6 % от среднего). И получить значимые статистические выводы в этом случае просто невозможно – надо увеличить объем выборки, число наблюдений.

А вот на вопрос – а сколько же надо наблюдений, каково их достаточное число, прикладная статистика имеет ответ: для “преодоления 5 % барьера” достаточно 5 наблюдений.

Попробуем решить другую задачу об оценке математического ожидания СВ на интервальной шкале, но будем решать её не “по чувству”, а “по разуму”.

• Наблюдения над случайной величиной X: 19,17,15,13,12,11,10,8,7.

• Количество наблюдений: 9, возможных исходов 512.

• **Њ0:** M(X)= 9, **Њ1:** M(X)# 9.

Найдем сумму отклонений от гипотетического среднего, S = 31.

Из 512 возможных вариантов суммы отклонений выберем только те, в которых эта сумма составляет 31 и более. Таких вариантов всего 11, значит при принятии нулевой гипотезы **Њ0:** M(X)= 9 вероятность наблюдать такие суммы P(S ≥31) составляет 11/512 ≅ 0.02 , что меньше порогового значения в 5 % .

Вывод: гипотезу Њ0 следует отвергнуть и считать приемлемым по надежности неравенство M(X) # 9.

До сих пор мы выдвигали гипотезу о значении математического ожидания на “левом крае” распределения наблюдений и могли бы повторять проверки, задаваясь значениями M(X) в 10, 11 и т.д., до тех пор, пока вероятность ошибки первого рода не достигла бы порогового значения.

Можно также исследовать правый край распределения ­– проверять гипотезы при больших значениях математического ожидания.

Например:

• Наблюдения над случайной величиной X: 19,17,15,13,12,11,10,8,7.

• Количество наблюдений: 9, возможных исходов 512.

• **Њ0:** M(X)= 17, **Њ1:** M(X)# 17.

Теперь сумма отклонений от гипотетического среднего окажется S = – 41.

Из 512 возможных вариантов суммы отклонений выберем только те, в которых эта сумма составляет –41 и менее. Таких вариантов всего 3, значит при принятии нулевой гипотезы Њ0: M(X)= 17 вероятность наблюдать такие суммы составляет P(S ≤ – 31) = 3/512 ≅ 0.006 , что намного меньше порогового значения в 5 % . Следовательно, можно попробовать гипотезы с меньшим M(X), сужая диапазон или так называемый доверительный интервал для неизвестного нам математического ожидания.

## Оценка наблюдений при известном законе распределения

Не всегда закон распределения СВ представляет для нас полную тайну. В ряде случаев у нас могут быть основания предполагать, что случайные события, определяющие наблюдаемые нами значения этой величины, подчиняются определенной вероятностной схеме.

В таких случаях использование методов выдвижения и проверки гипотез даст нам информацию о параметрах распределения, что может оказаться вполне достаточно для решения конкретной экономической задачи.

### Оценка параметров нормального распределения

Нередки случаи, когда у нас есть некоторые основания считать интересующую нас СВ распределенной по нормальному закону. Существуют специальные методы проверки такой гипотезы по данным наблюдений, но мы ограничимся напоминанием природы этого распределения ­– наличия влияния на значение данной величины достаточно большого количества случайных факторов.

Напомним себе также, что у нормального распределения всего два параметра – математическое ожидание μ и среднеквадратичное отклонение σ.

Пусть мы произвели 40 наблюдений над такой случайной величиной **X** и эти наблюдения представили в виде:

Таблица 5-2

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Xi | 85 | 105 | 125 | 145 | 165 | 185 | 205 | 225 | Всего |
| ni | 4 | 3 | 3 | 2 | 4 | 7 | 12 | 5 | 40 |
| f i | 0.100 | 0.075 | 0.075 | 0.050 | 0.100 | 0.175 | 0.300 | 0.125 | 1 |

Если мы усредним значения наблюдений, то формула расчета выборочного среднего

Mx = Σ Xi • ni =Σ Xi • fi {5–1} будет отличаться от выражения для математического ожидания μ только использованием частот вместо вероятностей.

В нашем примере выборочное среднее значение составит Mx = 171.5 , но из этого пока еще нельзя сделать заключение о равенстве μ = 171.5.

• Во-первых, Mx – это непрерывная СВ, следовательно, вероятность ее точного равенства чему-нибудь вообще равна нулю.

• Во-вторых, нас настораживает отсутствие ряда значений **X.**

• В-третьих, частоты наблюдений стремятся к вероятностям при бесконечно большом числе наблюдений, а у нас их только 40. Не мало ли?

Если мы усредним теперь значения квадратов отклонений наблюдений от выборочного среднего, то формула расчета выборочной дисперсии

Dx = (Sx)2 = Σ (Xi – Mx)2• ni =Σ (Xi)2• fi – (Mx)2 {5–2} также не будет отличаться от формулы, определяющей дисперсию σ2 .

В нашем примере выборочное значение среднеквадратичного отклонения составит Sx= 45.5 , но это совсем не означает, что σ =45.5.

И всё же ­– как оценить оба параметра распределения или хотя бы один из них по данным наблюдений, т.е. по уже найденным Mx и Sx?

Прикладная статистика дает следующие рекомендации:

• значение дисперсии σ2 считается неизвестным и решается первый вопрос ­– достаточно ли число наблюдений N для того, чтобы использовать вместо величины σ ее выборочное значение Sx;

• если это так, то решается второй вопрос ­ ­– как построить нулевую гипотезу о величине математического ожидания μ и как ее проверить.

Предположим вначале, что значение σ каким–то способом найдено. Тогда формулируется простая нулевая гипотеза **Њ0:** μ=Mx и осуществляется её проверка с помощью следующего критерия. Вычисляется вспомогательная функция (**Z**–критерий)

 , {5-3} значение и знак которой зависят от выбранного нами предполагаемого μ.

Доказано, что значение **Z** является СВ с математическим ожиданием 0 , дисперсией 1 и имеет нормальное распределение.

Теперь важно правильно построить альтернативную гипотезу **Њ1.** Здесь чаще всего применяется два подхода.

Выбор одного из них зависит от того ­– большое или малое (по модулю) значение **Z** у нас получилось. Иными словами ­– как далеко от расчетного Mx  мы выбрали гипотетическое μ..

• При малых отличиях между Mx и μ разумно строить гипотезы в виде

**Њ0:** μ= Mx;

**Њ1:** неизвестное нам значение μ лежит в пределах

Mx – •Z 2k ≤ μ ≤ Mx + •Z 2k {5–4}

Критическое (соответствующее уровню значимости в 5%) значение критерия составляет при этом = 1.96 (двухсторонний критерий). Если оказывается, что выборочное значение критерия ⏐Z⏐ < 1.96, то гипотезаЊ0: μ=Mx принимается, данные наблюдений не противоречат ей.

Если же это не так, то мы “в утешение” получаем информацию другого вида ­– где, на каком интервале находится искомое значение μ.

• При больших отличиях (в большую или меньшую сторону) между μ и Mx гипотезы строятся иначе Њ**0:** μ= Mx; **Њ1:** неизвестное нам значение μ лежит вне пределов, указанных в {5–4}.

Теперь критическое (соответствующее уровню значимости в 5%) значение критерия составляет Z 1k = 1.645 (односторонний критерий). Если оказывается, что выборочное значение критерия⏐Z⏐ ≥ 1.645, то гипотеза Њ0: μ =Mx отвергается, данные наблюдений противоречат ей.

Если же это не так, то мы получаем информацию другого вида ­– где, на каком крае интервале находится искомое значение μ. Разумеется, для других (не 5%) значений уровня значимости Z1k и Z 2k являются другими.

Чуть сложнее путь проверки гипотез о математическом ожидании μ в случаях, когда σ нам неизвестна и приходится довольствоваться выборочным значением среднеквадратичного отклонения по данным наблюдений.

В этом случае вместо “**z** –критерия” используется т.н. “t–критерий” или критерий Стьюдента

 , {5–5} в котором используется значение “несмещенной” оценки для дисперсии σ2

**(Sx)2**= Σ (Xi – Mx)2• ni . {5–6}

Далее используется доказанное в теории положение ­– случайная величина t имеет специальное распределение Стьюдента с m=N–1 степенями свободы.

Существуют таблицы для этого распределения по которым можно найти вероятность ошибки первого рода или, что более удобно, – граничное значение этой величины при заданных заранее α и m. Таким образом, если вычисленное нами значение ⏐t⏐≥ t(α,m), то Њ0 отвергается, если же это не так – Њ0 принимается. Конечно, при большом количестве наблюдений (N>100…120) различие между z– и t–критериями несущественно. Значения критерия Стьюдента для α=0.05 при разных количествах наблюдений составляют:

Таблица 5–3

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| m | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 20 | 30 | 40 | 120 |
| t | 12.7 | 4.30 | 3.18 | 2.78 | 2.57 | 2.45 | 2.36 | 2.31 | 2.26 | 2.23 | 2.09 | 2.04 | 2.02 | 1.98 |

### Оценка параметров дискретных распределений

В ряде случаев работы с некоторой дискретной СВ нам удается построить вероятностную схему событий, приводящих к изменению значений данной величины. Иными словами ­– закон распределения нам известен, но неизвестны его параметры. И наша задача ­– научиться оценивать эти параметры по данным наблюдений.

Начнем с наиболее простого случая. Пусть у нас есть основания считать, что случайная величина X может принимать целочисленные значения на интервале [0…k…n] с вероятностями

P(X=k)=pk(1– p)n-k,

т.е. распределена по биномиальному закону. Так вот, – единственный параметр p этого распределения нас как раз и интересует.

Примером подобной задачи является чисто практический вопрос о контроле качества товара.

Пусть мы решили оценить качество одной игральной кости из партии, закупленной для казино. Проведя n=200 бросаний мы обнаружили появлений цифры 6 в X = 25 случаях.

Выдвинем нулевую гипотезу Њ0: кость симметрична, то есть p= 1/6.

Вроде бы по наблюдениям частота выпадения цифры 6, составившая 25/200 не совпадает с гипотетическим значением вероятности 1/6. Но это чисто умозрительное, дилетантское заключение.

Теория прикладной статистики рекомендует вычислить значение непрерывной СВ

, {5–7} т.е. использовать **z**–критерий (см. {5–3}).

В нашем примере наблюдаемое значение Z составит около –1.58. Следовательно, при пороговой вероятности в 5% условие ⏐Z⏐< 1.96 выполняется и у нас нет оснований отбрасывать нулевую гипотезу о симметрии игральной кости.

Отметим, что **z**–критерий позволяет решать еще одну важную задачу – о достаточном числе испытаний.

Пусть нам требуется проверить качество товара – некоторых изделий, каждое из которых может быть годным или негодным (бракованным). Пусть допустимый процент брака составляет p=5%. Ясно, что чем больше испытаний мы проведем, тем надежнее будет наш статистический вывод ­– браковать партию товара (например, – 10000 штук) или считать её пригодной.

Если мы провели n=500 проверок и обнаружили X=30 бракованных изделий, то выдвинув гипотезу Њ0: p=5% , мы найдем выборочное значение критерия по {5–7}. Оно составит около 1.03, что меньше “контрольного” 1.96 . Значит, у нас нет оснований браковать всю партию.

Но возникает вопрос – сколько проверок достаточно для принятия решения с уровнем значимости в 5%? Для этого достаточно учесть допустимый процент брака (т.е. задать p), указать допустимое расхождение между ним и наблюдаемым процентом брака в выборке (d= p–X/n) и воспользоваться выражением

**** {5–8}

Если мы примем d=±0.02, то получим ответ – вполне достаточно 456 проверок, чтобы убедиться в том, что реальный процент брака отличается от допустимого не более чем на 2%.

# Выборочные распределения на шкале Nom

Напомним, что случайная величина X, принимающая одно из n допустимых значений A, B, C и т.д. имеет номинальную шкалу тогда, когда для любой пары этих значений применимы только понятия “равно” или “неравно”.

Для подобных СВ не существует понятий математического ожидания, как и других моментов распределения. Но понятие закона распределения имеет смысл ­– это ряд вероятностей PA = P(X=A) для каждого из допустимых значений. Соответственно, итоги наблюдения над такой СВ дадут нам частоты fA. Если у нас имеется всего N наблюдений за такой величиной, то иногда имеется возможность выдвинуть и проверить гипотезы о природе такой случайной величины, ее законе распределения и параметрах этого закона. Ситуации, когда это возможно сделать, не так уж и редки – всё зависит от понимания нами природы, сути случайных событий, от многозначности случайной величины и, конечно же, от количества наблюдений.

## Случай двухзначной случайной величины, N<50

Пусть нам крайне важно оценить "симметричность" некоторой случайной величины на номинальной двухпозиционной шкале со значениями "+" и "–" по наблюдениям за этой величиной. Если таких наблюдений было N+ =15и N– = 25 соответственно, то это вся информация, которая у нас есть. Что же можно узнать из нее? Оказывается – достаточно много и иногда … даже надёжно!

В конце концов, мы можем полагать вероятность значения "+" на данной номинальной шкале равной p и тогда q = (1 – p) даст нам вероятность положения "–" на этой же шкале. Таким образом, мы уже построили закон распределения и дело остается за оценкой его единственного параметра p.

По сути дела у нас есть одна дискретная случайная величина – число появлений X на "первой" позиции своей номинальной шкалы и это число составляет S= N+ .

Но совершенно ясно, что новая случайная величина S имеет биномиальный закон распределения и вероятность наблюдения N+ =15вполне можно вычислить, если знать или задаться значением p.

Выдвинем вначале нулевую гипотезу о симметрии распределения X и альтернативную ненаправленную гипотезу –

Њ0: p=q= 0.5; Њ1: p#q# 0.5.

Как обычно, оценим вероятность имеющегося наблюдения при верной нулевой гипотезе. Используя формулы расчета вероятности P(S≤15) или специальные таблицы биномиального распределения находим для 5%–го уровня значимости, что критическое значение **S** составляет 27, т.е. заметно больше наблюдаемого N+ =15. Следовательно, наши наблюдения статистически значимы – можно отвергнуть гипотезу Њ0, рискуя при этом ошибиться только в пяти случаях из 100.

Рассмотрим теперь несколько иной пример. Пусть нам необходимо проверить партию изделий в 50 штук при следующем правиле ­– вся партия бракуется, если доля бракованных изделий превышает 10%.

Выдвигаем гипотезы

Њ0: p ≤ 0.10 и q ≥ 0.90; Њ1: q ≤ 0.90 и p ≥ 0.10.

Можно сразу решить вопрос о количестве проверок N, достаточном для обоснованном решении об отбрасывании нулевой гипотезы. Поскольку мы имеем биномиальное распределение числа бракованных изделий в выборке из N наблюдений, то нам надо, прежде всего, установить порог значимости наблюдений ­– примем его традиционно, равным 0.05.

Теперь можно начинать наблюдения, накапливая результаты и по мере роста числа наблюдений контролировать их значимость. Покажем, как это делать в ситуации, когда N=48, а число бракованных изделий к этому времени составило 4.

По сути дела, нам надо вычислить вероятность появления 4 отрицательных исходов и всех еще менее вероятных в серии из 48 испытаний. Правда сделать это вручную слишком сложно – придется работать с биномом 48 степени. Поэтому при отсутствии компьютерной программы можно использовать специальные таблицы биномиального распределения.

В них можно найти значение числа событий с вероятностью 0.10 каждое, достаточное для отбрасывании нулевой гипотезы с вероятностью ошибки первого рода в 5%. В наших условиях это число равно 9, значит при наблюдаемом меньшем числе бракованных изделий (всего 4) гипотезу Њ0 следует принять и всю партию не браковать.

## Случай двухзначной случайной величины, N>50

При достаточно больших выборках можно поступать и иначе. В качестве правила проверки гипотез используют так называемый критерий "хи–квадрат”

χ2 = ∑. {6–1}

Эта непрерывная случайная величина была предложена видным статистиком Р.Фишером для проверки гипотез о соответствии выборочного распределения некоторому заданному закону. Для этого используются экспериментальные частости NE и вычисленные в соответствии Њ0 “теоретические” NH . Разумеется, суммирование ведется по всем допустимым значениям СВ. В нашем примере у нее всего лишь два значения (изделие годно или бракованное), поэтому в числителе надо иметь т.н. поправку на непрерывность. Она корректирует влияние природы распределений: дискретное у наблюдаемой величины и непрерывное у критерия Фишера.

Изменим условия предыдущего примера ­– пусть N= 100, число бракованных изделий составило NE–=12. Нетрудно определить NE+=88, но что касается "гипотетических" частостей NH– и NH+, то эти величины зависят от того, как мы сформулируем гипотезы. Если их оставить без изменения, то эти частости составят NH+ = 90 и NH– = 10. Вычисление выборочного значения χ2–критерия не вызывает проблем, важнее знать – как использовать результат расчета. В нашем примере расчетное значение критерия составит 0.25. Кроме конкретного значения критерия надо учесть так называемое число степеней свободы. В нашем случае это 1, а в общем случае надо уменьшить число допустимых значений n на единицу. Ну, а далее требуется взять стандартные статистические таблицы, учесть пороговое значение ошибки первого рода ­– и получить ответ. Для примера приведем часть такой таблицы при α=0.05

Таблица 6–1

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Степеней свободы | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Критическое χ2 | 3.84 | 5.99 | 7.82 | 9.49 | 11.1 | 12.6 | 14.1 | 15.5 | 16.9 |

Если наблюдаемое значение χ2 меньше критического, гипотеза Њ0 может быть принята.

В условиях нашего примера расчетное значение критерия χ2 составляет всего лишь 0.25, что меньше критического 3.48 (для одной степени свободы) и отвергать гипотезу Њ0 (браковать всю партию) нет оснований. Но, если бы мы наблюдали не 12, а 17 случаев брака, то расчетное значение критерия составило бы около 4.62 и гипотезу Њ0 пришлось бы отвергнуть.

## Случай многозначной случайной величины

Существует достаточно обширный класс задач со случайными величинами, распределенными на номинальной шкале с тремя и более допустимыми значениями.

В таких задачах обычно используется все тот же критерий χ2 с числом степеней свободы более одной. По сути дела, используют почти ту же формулу –

χ2 = ∑, {6–2} в которой просто не используется поправка на непрерывность.

Так, например, наблюдая численности покупок четырех категорий некоторого товара, мы могли зафиксировать следующие данные:

Таблица 6–1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Товары | A | B | C | D | Всего |
| Число покупок | 30 | 55 | 27 | 48 | 160 |

Выдвинем гипотезы:

Њ0: Все товары одинаково популярны или РА=РB=РC=РD=0.25

Њ1: Популярности товаров значимо различны.

Несложный расчет дает расчетную величину критерия около 14, т.е. ощутимо больше критического значения 7.8 для 3–х степеней свободы по табл. 6–1. Это дает нам основание отвергнуть гипотезу о равной популярности этих видов товара.

# Выборочные распределения на шкале Ord

Случайные величины с порядковой шкалой измерения – это дискретные, для всех допустимых значений которых, кроме отношений“=" или "#”, разрешены отношения “<" или ">”. Классическим примером порядковых величин являются оценки знаний, успеваемости, приоритета. Для таких СВ, как и для номинальных, не имеют смысла понятия моментов распределений.

Продемонстрируем ряд задач, возникающих при оперировании такими величинами и рассмотрим специальные методы непараметрической статистики в применении к этим задачам.

Следует различать ситуации, связанные с величинами на порядковой шкале:

• случайная величина имеет всего два допустимых значения (одно из них больше, предпочтительнее второго);

• случайная величина имеет более двух допустимых значений.

В первом случае мы имеем по сути дела двух позиционную номинальную шкалу и все сказанное выше о распределениях на шкале Nom вполне приемлемо для решения задач на такой шкале Rel. К примеру ­– задачи о проверке симметрии монеты или о допустимом количестве бракованных изделий вполне могут рассматриваться с использование порядковой шкалы, если считать герб “старше” решки, бракованное изделие “хуже” исправного.

Второй тип СВ предполагает наличие нескольких фиксированных значений, упорядоченных по некоторому признаку, свойству или нашему предпочтению. В этих случаях говорят, что случайная величина (например – оценка знаний, сорт товара) может быть величиной “первого ранга”, “второго ранга” и т.д.

В принципе корректная постановка задач о распределении СВ на порядковых (ранговых) шкалах ничем не отличается от рассмотренных ранее методов статистики для интервальных, относительных и номинальных шкал.

Пусть мы наблюдали, зафиксировали оценки знаний 100 обучаемых по четырех ранговой шкале (“отлично”, “хорошо”, “удовлетворительно” и “плохо”)

Таблица 7–1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Оценка знаний | Отл. | Хор. | Удовл. | Плохо | Всего |
| Ранг оценки по смыслу | 1 | 2 | 3 | 4 |  |
| Количество наблюдений | 25 | 45 | 20 | 10 | 100 |
| Ранг по итогам наблюдений | 2 | 1 | 3 | 4 |  |

Как обычно, далее приходится строить гипотезы и подбирать критерии для их проверки. При выдвижении нулевой гипотезы надо, прежде всего, помнить о необходимости с её помощью рассчитать распределение СВ – в нашем случае это означает расчет количества оценок в условиях истинности Њ0.

Конечно, без “технологических” представлений о природе СВ выдвижение и проверка гипотез (а затем использование статистических выводов) ­– пустая трата времени.

Пусть мы осознаем зависимость оценки знаний от предварительной подготовки обучаемых (она может быть одинакова у всех или значимо отличаться), от эффективности системы обучения и, наконец, от способа проверки знаний. Тогда результаты наблюдений могут оказаться полезными при решении задач управления обучением и, по крайней мере, контроля процесса обучения.

Если у нас есть основания считать предварительную подготовку обучаемых одинакового уровня для всех и способ проверки знаний достаточно объективным, то тогда можно выдвинуть нулевую гипотезу Њ0: система обучения эффективна. Конечно, мы не можем теоретически предсказать количество оценок каждого из рангов. Но этого и не нужно ­– оценки не числа, и частота наблюдения оценки “отлично” не может быть умножена на значение этой оценки. Другое дело, если мы договоримся считать систему обучения эффективной только в том случае, если она по отношению к одинаково подготовленным обучаемым дает большие числа более высоких оценок.

Тогда, в соответствии с Њ0 ранги 2–й строки табл.7–1 могут рассматриваться как гипотетические, а ранги 4-й строки ­– как выборочные, наблюдаемые. Осталось установить ­– какой же критерий принять для проверки нашей гипотезы. Один из часто используемых в подобных задачах критериев носит название коэффициента ранговой корреляции Спирмэна

 , {7–1}

в котором di – разности гипотетических и наблюдаемых рангов; n – число рангов.

Величина коэффициента ранговой корреляции имеет непрерывное распределение на интервале [–1…+1] с математическим ожиданием 0 – если, конечно, гипотеза Њ0 верна. Поэтому значение вычисленного Rs можно использовать в качестве критерия проверки гипотез. В нашем примере сумма квадратов разностей рангов равна S=2 и для n=4 коэффициент Спирмэна по итогам наблюдений составит Rs = 0.8. Обратимся теперь к статистическим таблицам и рассмотрим ту, которая рассчитана для числа рангов n=4.

Таблица 7–2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Наблюдаемое значение суммы S | 2 | 4 | 6 | 8 | 10 |
| Вероятность S при ошибочности Њ0 | 0.042 | 0.167 | 0.208 | 0.375 | 0.458 |

Для нашего примера предположение о полной эффективности системы обучения вполне обосновано.

Мы ознакомились только с одним из существующих методов статистического анализа СВ со шкалой Ord. Существуют и другие, обоснованные и апробированные методы (коэффициент ранговой корреляции Кэндалла). Отличие между ними только в способе расчета критерия принятия или отбрасывания нулевой гипотезы.Вместе с тем мы не затронули вопроса о проблемах, возникающих при наличии нескольких величин с ранговой шкалой измерения. Эти проблемы связаны с множественной ранговой корреляцией или конкордацией (согласованностью рангов).

Пусть у нас имеются ранжировки m=4 экспертов по отношению к n=6 факторам, которые определяют эффективность некоторой экономической системы:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Эксперты / Факторы | F1 | F2 | F3 | F4 | F5 | F6 | ∑ |
| A | 5 | 4 | 1 | 6 | 3 | 2 | 21 |
| B | 2 | 3 | 1 | 5 | 6 | 4 | 21 |
| C | 4 | 1 | 6 | 3 | 2 | 5 | 21 |
| D | 4 | 3 | 2 | 5 | 1 | 6 | 21 |
| Сумма рангов | 15 | 11 | 10 | 19 | 12 | 17 | 84 |
| Суммарный ранг | 4 | 2 | 1 | 6 | 3 | 5 |  |
| Отклонение суммы рангов от 84/6 =14 | +1 | -3 | -4 | +5 | -2 | +3 |  |
| Квадраты этих отклонений | 1 | 9 | 16 | 25 | 4 | 9 | 64 |

Заметим, что полная сумма рангов составляет 84, что дает в среднем по 14 на фактор. Для общего случая n факторов и m экспертов среднее значение суммы рангов для любого фактора определится выражением

**Δ ** 0.5•m•(n+1) {7–2}

Теперь можно оценить степень согласованности мнений экспертов по отношению к шести факторам. Для каждого из факторов наблюдается отклонение суммы рангов, указанных экспертами, от среднего значения такой суммы.

Поскольку сумма этих отклонений всегда равна нулю, для их усреднения разумно использовать квадраты значений. В нашем случае сумма таких квадратов составит S= 64, а в общем случае эта сумма будет наибольшей только при полном совпадении мнений всех экспертов по отношению ко всем факторам:

Smax m2 • (n3– n) / 12 {7 –3} что в нашем примере дает 280.

М. Кэндаллом предложен показатель согласованности или коэффициент конкордации, определяемый как

W = S / Smax {7–4} принимающий, в отличие от обычных (парных) коэффициентов ранговой корреляции, значения от 1 (при наибольшей согласованности) до 0.

В нашем примере значение коэффициента конкордации составляет около 0.23 и явно недостаточно для принятия гипотезы о согласованности мнений экспертов.

Существуют специальные таблицы, позволяющие отыскивать значения сумм S, настолько близких к **S**max , что вероятность ошибки при принятии гипотезы о полной согласованности мнений экспертов не превосходит 5%. Вот одна из таких таблиц с критическими (достаточными) значениями сумм квадратов отклонений рангов S для n=3…7 факторов при m= 3…15 экспертов.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| m \ n | 3 | 4 | 5 | 6 | 7 |
| 3 | – | – | 64 | 104 | 157 |
| 4 | – | 50 | 88 | 143 | 217 |
| 5 | – | 63 | 112 | 182 | 276 |
| 6 | – | 76 | 136 | 221 | 335 |
| 8 | 48 | 102 | 184 | 299 | 453 |
| 10 | 60 | 128 | 231 | 377 | 571 |
| 15 | 90 | 193 | 350 | 571 | 865 |

Для нашего примера указанная вероятность соответствует сумме квадратов отклонений S= 143, что намного больше наблюдаемой суммы 64. Поэтому гипотезу о согласованности мнений экспертов придется отбросить.

# Материал семинарских занятий

## Введение в комбинаторику

При изучении курса математической статистики приходится использовать методы одного из разделов математики, который хотя формально и не относится к высшей, вузовской математике, но, к сожалению, не изучается в средней школе.

Этот раздел – комбинаторика, “наука о способах подсчета вариантов”. Эта наука имеет тот же, примерно 300 летний возраст, что и сама статистика. Комбинаторика – сверстница теории вероятностей, теоретического фундамента прикладной статистики. Как и в древней, в современной статистике невозможно обойтись без навыков просчитывать в уме или, по крайней мере, быстро, по простым формулам, варианты событий, размещений предметов, значений величин и т.п.

Замечание о расчетах в уме сделано не случайно. Знание основ комбинаторики позволит хотя бы оценивать числа вариантов и соотношения между ними также “профессионально” как и делаете это вы, оценивая возраст встреченного человека.

В этом плане комбинаторику можно называть “логикой вариантов” и это будет вполне резонно ­– в этой науке больше чистой логики, чем математики.

Для демонстрации необходимости знаний комбинаторики и в качестве первой практической задачи рассмотрим несколько простых, практических вопросов.

• Вам, очевидно, известно, что внутренний, “машинный” язык компьютера люди построили по образу и подобия человеческого языка: буквы, слова, предложения.

Обстоятельства надежности записи и чтения на этом языке привели к решению сделать компьютерный язык предельно бедным. В нем всего две буквы (“0” и “1”, “+ " и “–”, “да” и “нет”, ­– в зависимости от физического процесса записи), всегда 8 букв в слове, отсутствует пробел между словами (это была бы третья буква).

И вот возникает вопрос ­– а сколько вариантов у машинного слова, т.е. у одного байта? Еще проще – если одним байтом записывать числа, то сколько положительных целых чисел можно охватить 1 байтом? В поисках ответа можно терпеливо выписывать все возможные варианты слов из 8 нулей и единиц: 00000000, 00000001, 00000010 и т.д. до 11111111. Но ведь это долго и надо быть уверенным, что ничего не пропустили!

Так вот – законы комбинаторики позволяют мгновенно ­решить эту задачу и получить ответ – вариантов записи байта ровно 256.

Это чисто практический вопрос – ведь компьютер с возможностью считать в целых числах от –128 до 127 никто не купит.

Ну, если целые числа хранить в 2-х машинных словах, в 2-х байтах или в 16 “разрядах”.? Уж это новое число вариантов никто не согласится вычислять простым перебором! А ответ комбинаторики все тот же прост – в этом случае есть возможность работать с целыми числами от –32768 до 32767.

Оказывается, что эти числа не надо запоминать, поскольку алгоритм их расчетов очень прост и посилен человеку, осилившему только арифметику.

• Рассмотрим второй пример решения практического вопроса с использованием правил комбинаторики. Пусть решается вопрос об установлении проводной связи между 25 предприятиями фирмы по следующему принципу – каждое предприятие должно иметь отдельный канал связи со всеми остальными. Сколько таких каналов придется установить в фирме?

Для решения вопроса можно нарисовать выпуклый 25–угольник и провести в нем все диагонали, пересчитав в конце их число и не забыв добавить число сторон. Человек, знающий комбинаторику, во-первых, не сделает ошибки –25•24=600 каналов. Во-вторых, он мгновенно укажет верный ответ – всего требуется 300 каналов. Комментарии излишни…

Для освоения наиболее популярных применений комбинаторики нам потребуется использовать, по крайней мере, два ее основных понятия ­– перестановки и сочетания.

Перестановками называют операции над упорядоченным рядом из n различных объектов, в процессе которых “списочный состав” ряда не изменяется, но “места” объектов в этом ряду изменяются от варианта к варианту. Не будем тратить время на обоснование расчетной формулы для произвольного n, а попробуем найти число перестановок в ряду из 1, 2 и 3 предметов.

Воспользуемся для этого простенькой схемой:

n=1 A 1 вариант.

n=2 AB BA 1•2= 2 варианта.

n=3 ABC ACB BCA BAC CAB CBA 1•2•3= 6 вариантов.

Можно доказать строго, что в общем случае число перестановок в ряду из n элементов составит

 {8–1}

Сочетаниями называют операции над множеством из n различных объектов, в процессе которых образуют подмножества из k элементов, взятых из исходного множества, так, чтобы варианты подмножеств отличались друг от друга хотя бы одним элементом.

Опустим доказательство формулы для расчета числа сочетаний из n по k в общем виде и приведем лишь примеры для числа сочетаний из 3 по 2 и из 5 по 3.

• Элементы исходного множества A, B, C.

Варианты подмножеств: AB, AC, BC – всего три.

• Элементы исходного множества A, B, C, D, E.

Варианты подмножеств: ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE, CDE – всего десять.

В общем случае число вариантов сочетаний или просто – число сочетаний из n по k определяется по формуле

= {8–2}

Существует еще один способ вычисления числа сочетаний из n по k – с использованием коэффициентов в развернутой форме бинома (p+q)n. В самом деле, например, при n=3 коэффициенты при степенях разложения составляют 1, 3, 3, 1 – а это и есть сочетания из 3 по 0, 1, 2, 3 и 4 элементов.

Известна также схема простого расчета биномиальных коэффициентов, которая носит названия треугольника Паскаля:

Для n

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  | **1** |  | **1** |  |  |  |  |  |  |  |  | **1** |  |  |
|  |  |  |  |  |  | **1** |  | **2** |  | **1** |  |  |  |  |  |  |  | **2** |  |  |
|  |  |  |  |  | **1** |  | **3** |  | **3** |  | **1** |  |  |  |  |  |  | **3** |  |  |
|  |  |  |  | **1** |  | **4** |  | **6** |  | **4** |  | **1** |  |  |  |  |  | **4** |  |  |
|  |  |  | **1** |  | **5** |  | **10** |  | **10** |  | **5** |  | **1** |  |  |  |  | **5** |  |  |
|  |  | **1** |  | **6** |  | **15** |  | **20** |  | **15** |  | **6** |  | **1** |  |  |  | **6** |  |  |
|  | **1** |  | **7** |  | **21** |  | **35** |  | **35** |  | **21** |  | **7** |  | **1** |  |  | **7** |  |  |

Первый элемент любого основания равен 1, второй – номеру основания, а все последующие – сумме двух "вышестоящих".

## Методы вычисления моментов распределений

При вычислении моментов распределения случайных величин полезно использовать некоторые удобные (как для прямого расчета, так и для составления компьютерных программ) выражения.

• Пусть требуется просуммировать ряд чисел T1, T2, ……Tk, …Tm и мы замечаем, что они отличаются друг от друга на одну и ту же величину d, т.е. образуют арифметическую прогрессию. В этом случае полезна замена –

 {8–3}

Таким образом, среднее значение для ряда таких чисел составит:

. {8–4}

• Для вычисления суммы чисел натурального ряда или суммы квадратов этих чисел удобны формулы:

;  . {8–5}

• Если некоторая случайная величина Y может быть выражена через другую в виде

Y= a•X+b, то справедливы соотношения:

M(Y) = a•M(X)+b; D(Y) = a2 • D(X). {8–6}

• Если некоторая случайная величина X имеет математическое ожидание M(X) и среднеквадратичное отклонение S(X) , то "нормированная" случайная величина:

 {8–7} имеет нулевое математическое ожидание и единичную дисперсию.

## Алгоритмы простейших статистических расчетов

Несмотря на относительную простоту, статистические расчеты требуют значительных затрат времени, повышенного внимания и, связанного с этим риска ошибок. Кроме того, в большинстве случаев практики после расчетов выборочных значений и выдвижения гипотез почти всегда приходится обращаться к статистическим таблицам, т.е. к данным классических распределений.

Большую часть этих трудностей можно преодолеть ­– путем использования специальных статистических программ (или целого набора – пакета прикладных программ).

На сегодня программное обеспечение статистических расчетов выполнено, как правило, на уровне глобальных задач прикладной статистики, системного анализа и т.п. Надежных, простых в употреблении компьютерных программ практически нет – считается, что писать и распространять такие программы не престижно! С другой стороны, потребители таких программ – профессиональные статистики не испытывают затруднений в самостоятельном написании удобных (для себя) программ и даже пакетов. То, что есть – не хорошо и не плохо, просто это традиция и нарушать ее нет желания ни у фирм, производящих программы, ни у потенциальных пользователей.

Поэтому имеет смысл затратить некоторое время на анализ определенных трудностей, которые наверняка будут проявляться при программировании типовых статистических расчетов.

Оказывается, что здесь программиста поджидают "подводные камни", тупики и прочие неприятности, связанные не только с реальными возможностями компьютера, но и с самими формулами статистики, особенностями этой науки.

### Вычисление моментов выборочных распределений

Пусть у нас имеется массив выборочных значений случайной величины и соответствующие частости (числа наблюдений) этих значений, то есть матрица из двух столбцов и m строк.

Обозначим такой массив W и рассмотрим вопрос о вводе исходных данных. Конечно же, мы быстро сообразим, что ввод надо организовать для пар значений Xi, ni – только в этом варианте можно снизить вероятность ошибок.

Вопрос об общем количестве наблюдений можно не ставить в начале диалога – освободить пользователя от необходимости вычислять **N** = n1 + n2 + … + nm. Организовать сигнал конца ввода не представляет проблем – скажем, ввести отрицательное число наблюдений на очередном шаге.

Как организовать подготовку данных для расчета выборочных моментов – например, выборочного среднего Mx и выборочной дисперсии Dx?

Среди многих вариантов наилучшим будет, пожалуй, следующий.

Приготовить три контрольных величины M1, M2 и NN, предварительно присвоив им нулевые значения до начала ввода, что на языке Pascal будет выглядеть так –

Var NN, I, X, Y: Integer;

W: Array [1…2,1…m] of Integer;

M1, M2, D, S, V: Real;

M1:=0; M2:=0; NN:=0; I:= 0;

Теперь можно организовать суммирование поступающих с клавиатуры (или прямо из уже готового массива, записанного где–то на диске) выборочных данных Xi и ni.

Пусть у нас такой массив уже есть, тогдас каждой очередной парой чисел следует поступить так

Repeat

I:=I + 1; X:=W[I,1]; Y:=W[I,2];

NN:=NN+Y;

M1:=(M1+X•Y); M2:=M2+Sqr(X)•Y

Until I < m;

Операцию надо повторять до тех пор, пока мы не достигнем конца массива (при вводе с клавиатуры ­– пока не будет введено отрицательное значение очередного ni).

Если ввод окончен, то далее выборочные среднее, дисперсия и коэффициент вариации

N:=NN; M1:=M1/N;

D:=M2/N – Sqr(M1); S:=Sqrt(D); If M1#0 Then V:=S/M1;

### Проблема переполнения

В предыдущем примере программирования процедуры вычисления моментов была не отмечена опасность "переполнения" – суммы M1 и M2 могут выйти за "разрядную сетку" компьютера.

Если такая угроза очевидна, то простейший выход из положения – вычислить предварительно общее число наблюдений N и потом выполнять описанный выше алгоритм суммирования с использованием не частостей, а частот.

Более надежным, однако, является другой подход к этой проблеме. Достаточно на каждом шаге суммирования преобразовывать "старые" значения сумм M1 и M2 в "новые".

Var N, NN, I, X, Y: Integer;

W: Array [1… 2,1… m] of Integer;

А, B, M1, M2, D, S, V: Real;

M1:=0; M2:=0; N:=0; I :=0;

Repeat

I:=I + 1;

X:=W[I,1]; NN:=N+W[I,2]:

A:=N/NN; B:=W[I,2]/NN;

M1:=M1•A+X•B;

M2:=M2•A+Sqr(X)•B; N:=NN

Until I< m;

D:=M2 – Sqr(M1); V:= Sqrt(D);

If M1#0 Then V:=S/M1;

Более остро стоит проблема переполнения при вычислении факториалов, входящих в формулы вероятностей многих классических законов дискретных случайных величин.

Продемонстрируем метод решения подобной проблемы при вычислении биномиальных коэффициентов.

Если нам необходимо найти k–й коэффициент бинома n–й степени, то вполне надежным будет следующий алгоритм.

A:=N; B:=K; C:=1;

Repeat

C:=C•A/B; A:=A-1; B:=B-1

Until B>0;

Полезно также знать, что при достаточно больших N вычисление факториала можно производить по формуле Стирлинга , однако приведенный алгоритм намного проще алгоритма использования этой формулы.

### Моделирование законов распределения

Практика прикладной статистики невозможна без использования данных о классических, стандартных законах распределения. Чтобы избежать непосредственного использования статистических таблиц при выполнения расчетов ­– особенно в части проверки гипотез, можно поступить двояко.

• Ввести содержание таблиц в память компьютера (непосредственно в рабочую программу или в виде отдельного файла ­– приложения к этой программе). Но этого мало. Надо научить компьютер "водить пальцем по таблице", т.е. запрограммировать иногда не совсем элементарный алгоритм пользования таблицей. Работа эта хоть и занудная, но зато не требующая никаких знаний, кроме умения программировать решение корректно поставленных задач – описания пользования таблицами составлены четко и алгоритмично.

• Можно поступить более рационально. Поскольку речь идет о классических распределениях дискретных или непрерывных случайных величин, то в нашем распоряжении всегда имеются формулы вычисления вероятности (или интеграла вероятности). Бытует мнение, что программирование расчетов по формулам является чуть ли не самым низким уровнем искусства программирования. На самом же деле это не совсем так, а при программировании законов распределения вероятностей ­– совсем не так!

Без понимания природы процесса, который порождает данную случайную величину, без знания основ теории вероятностей и математической статистики нечего и пытаться строить такие программы. Но если всё это есть, то можно строить компьютерные программы с такими возможностями статистического анализа, о которых не могли и мечтать отцы–основатели прикладной статистики. Покажем это на нескольких простых примерах.

Нам уже известно, что выдвижение в качестве нулевой гипотезы о некотором стандартном законе распределения связано только с одним обстоятельством – мы можем предсказывать итоги наблюдения в условиях её справедливости. Но это предсказание невозможно без использования конкретных значений параметра (или нескольких параметров) закона. Во всех "до–компьютерных" руководствах по прикладной статистике рано или поздно приходится читать – "а теперь возьмем таблицу … и найдем для наших условий …". Хочется проверить ту же гипотезу при другом значении параметра? Нет проблем! Повтори все расчеты при этом новом значении и снова работай с таблицей.

Иными словами, в "до–компьютерную" эпоху вопрос ­– а что вообще можно получить из данного наблюдения (или серии наблюдений), какова максимальная информация о случайной величине заключена в этих наблюдениях, – не ставился.

Причина этого очевидна ­– сложность и большие затраты времени на расчеты. Но дело еще и в том, что неопределенность статистических выводов приводила к тупиковой ситуации, когда затраты на проведение сложных, требующих особого внимания и безупречной логики расчетов, могли оказаться куда больше возможного экономического выигрыша при внедрении результатов.

Поэтому сегодня, отдав должное изобретательности творцов прикладной статистики, следует ориентировать практику статистических расчетов исключительно на применение компьютерных программ.

Это могут быть, условно говоря, "параметрические" программы, ориентированные на тот или иной тип распределения. Их назначение – найти по данным имеющихся наблюдений статистическую значимость гипотез о параметрах таких распределений или, наоборот, по заданным пользователем параметрам рассчитать вероятности всех (!) заданных им ситуаций.

Вполне реально создание и использование "непараметрических" программ – способных анализировать входные данные наблюдений и проверять гипотезы о принадлежности случайной величины к любому из "известных этой программе" закону распределения.

Наконец, использование компьютерной техники современного уровня позволяет решать за вполне приемлемое время и небольшую цену еще один вид задач – статистического моделирования. Сущность этого термина раскрывается в специальной области кибернетики – системном анализе, но кратко может быть раскрыта следующим образом.

Пусть некоторая случайная величина Z является, по нашим представлениям, функцией двух других случайных величин – X и Y. При этом оказывается, что X зависит от двух также случайных величин A и B, а Y зависит от трех случайных событий C, D и E.

Так вот, в этом "простом" случае мы знаем или предполагаем, что знаем вероятности всех событий и законы распределения всех случайных величин, кроме "выходной" величины **Z.**

Для простоты будем считать функциональные зависимости также известными (например, – вытекающими из некоторых законов природы):

**Z** = X – ; X = A + ;

A = 1, 2 , … 16 и распределена по биномиальному закону с параметром p= 0.42;

B – распределена по нормальному закону с μ=12 и σ =2;

Y = 42, если произошло событие C, а события D и E не произошли;

Y = 177, если произошли события D и E, независимо от того, произошло ли C;

Y = –15 во всех остальных случаях.

Ясно, что попытка строить для этого примера–шутки логическую схему, по которой можно было бы вычислять возможные значения **Z** и соответствующие этим значениям вероятности, обречена на провал – слишком сложными и не поддающимися аналитическому описанию окажутся наши выкладки.

Однако же, при наличии знаний хотя бы основных положений прикладной статистики и умении программировать, вполне оправданно потратить некоторое время на создание программы и ее обкатку, проверку по правилам статистики.

Далее можно будет "проигрывать" все возможные ситуации и буквально через секунды получать "распределение случайной величины Z" в любом виде (кроме, разумеется, формульного).

Итак, надо уметь программировать операции, дающие случайную величину с заранее оговоренным законом распределения. Большинство языков программирования высокого уровня имеют встроенные подпрограммы (процедуры или функции в языке Pascal), обеспечивающие генерацию случайной величины R, равномерно распределенной в диапазоне 0…1. Будем полагать, что в нашем распоряжении имеется такой "датчик случайных чисел".

Покажем, как превратить такую величину R в дискретную с биномиальным законом распределения. Пусть нам нужна случайная величина K, с целочисленными значениями от 0 до N при значении заданном значении параметра p. Один из вариантов алгоритма такой генерации мог бы выглядеть так.

Var X, P: Real;

I, K, N: Integer;

K:=0;

For I:=1 to N Do

Begin

X:= R;

If X>(1– p)

Then K:=K+1

End;

После очередного цикла генерации мы получаем случайную величину K, распределенную по биномиальному закону настолько надежно, насколько удачной является функция генерации числа R. Во избежание сомнений стоит потратить время на обкатку такого алгоритма – повторив цикл 100 или 1000 раз и проверив надежность генерации по данным "наблюдений" с помощью теоретических значений математического ожидания N•p и дисперсии N•p•(1–p).

Несколько более сложно генерировать непрерывные случайные величины, в частности для популярных распределений ­– нормального, "хи–квадрат", Стьюдента и т.п.

Дело здесь в том, что непрерывная случайная величина имеет бесконечное число допустимых значений, даже если интервал этих значений ограничен.

Но, вместе с тем, для конкретного закона распределения непрерывной случайной величины известна плотность вероятности – предел, к которому стремится вероятность попадания такой величины в заданный интервал при сужении интервала до нуля.

Покажем эти трудности и пути их преодоления на примере нормального распределения. Пусть нам требуется генерировать нормированную случайную величину Z с нормальным законом распределения.

Для такой величины μ =0, σ=1, а попадание ее значений в диапазон более 3 или менее –3 практически невероятно (около 0.0027).

Разобьем диапазон –3…+3 на 2N+1 интервалов, шириной 2d каждый. При достаточно малом d= 3 / N, вероятность попадания Z в любой из них вычисляется легко:

P(–d <Z <+d) ≅ 2d•ϕ (0) = 2P0;

P( d <Z < 3d) ≅ 2d•ϕ (2d) = P1< P0;

P(3d <Z < 5d) ≅ 2d•ϕ (4d) = P2< P1;

P(5d <Z < 7d) ≅ 2d•ϕ (6d) = P3< P2;

……………………………………………

P(2•d– d <Z < 2k•d+ d) ≅ 2d•ϕ (2k•d) = PK;

………………………………………………………

P(2N•d– d<Z < 2N•d+ d) ≅ 2d•ϕ (2N•d) = PN;

Поскольку 2•P0+2•P1+2•P2+ …+2•PN ≅1, то можно предложить следующий алгоритм генерации нормированной случайной величины Z .

Вся шкала допустимых значений генерации равномерно распределенной случайной величины X (0…1) разбивается на интервалы с шириной, соответствующей значениям PK, в порядке убывания.

Если при очередной генерации равномерно распределенной случайной величины X ее значение попадает в интервалы –

0.5 – P0<X<0.5 + P0, то считается сгенерированным Z=0;

0.5 + P0<X<0.5 + P0+P1, то считается сгенерированным Z=2d;

0.5+P0+P1<X<0.5+P0+P1+P2, то считается сгенерированным Z=4d;

и так далее вплоть до последнего правого интервала –

0.999 <X<1, когда считается сгенерированным Z = 2N•d+d =3+d.

Точно также генерируются отрицательные значения Z – в соответствии с ситуациями генерации X<0.5 .

Описанный выше алгоритм, разумеется, не является единственным. Его рассмотрение преследовало только оду цель ­– дать представление о самой возможности решать задачи программирования "источников" случайных величин с заранее заданным законом распределения.

# Литература

|  |  |  |
| --- | --- | --- |
| *Название* | *Автор* | *Год* |
| 1. Прикладная Статистика | Айвазян С.А. и др. | 1983 |
| 1. Стохастические модели социальных процессов | Бартоломью Д. | 1987 |
| 1. Прикладная комбинаторная математика | Беккенбах Э.(ред.) | 1968\* |
| 1. Математическая статистика вып.1,2 | Бикел П., Доксам М. | 1987 |
| 1. Таблицы математической статистики | Большев Л.Н., Смирнов Н.В. | 1965\* |
| 1. Комбинаторика | Виленкин Н.Я. | 1969\* |
| 1. Многомерное шкалирование | Дейвисон М. | 1988 |
| 1. Методы анализа данных | Дидэ Э. и др. | 1987 |
| 1. Теория распределений | Кэндалл М., Стьюарт А. | 1966 |
| 1. Статистические выводы и связи | Кэндалл М., Стьюарт А. | 1973\* |
| 1. Теоретическая статистика | Кокс Д., Хинкли Д. | 1978\* |
| 1. Математические методы статистики | Крамер Г. | 1975\* |
| 1. Ранговые корреляции | Кэндалл М. | 1975\* |
| 1. Математические методы в социальных науках | Лазарсфельд П., Генри Н. | 1973\* |
| 1. Проверка статистических гипотез | Леман Э. | 1985 |
| 1. Метод наименьших квадратов | Линник Ю.В. | 1962 |
| 1. Справочник по прикладной статистике т.1 | Ллойд Э., Ледерман У. | 1989\* |
| 1. Справочник по прикладной статистике т.2 | Ллойд Э., Ледерман У. | 1990\* |
| 1. Наука об управлении. Байесовский подход | Моррис У. | 1971\* |
| 1. Вероятность | Мостеллер Ф. | 1969\* |
| 1. Вычисл. алгоритмы в прикладной статистике | Мэйндоналд Дж. | 1988\* |
| 1. Стат.оценивание и проверка гипотез на ЭВМ | Петрович М.Л., Давидович М.И. | 1989\* |
| 1. Математическое открытие | Пойа Д. | 1970\* |
| 1. Теория вероятностей | Прохоров Ю.В., Розанов Ю.А. | 1973\* |
| 1. Прикладная теория статистических решений | Райфа Г., Шлейфер Р. | 1987 |
| 1. Введение в комбинаторный анализ | Риордан Дж. | 1963\* |
| 1. Справочник по непараметрической статистике | Рунион Р. | 1982\* |
| 1. Сборник задач по теории вероятностей… | Свешников А.А. | 1965\* |
| 1. Непараметрические методы статистики | Тюрин Ю.Н. | 1978 |
| 1. Статист. модели в инженерных задачах | Хан Г., Шапиро С. | 1969\* |
| 1. Статист. выводы, основанные на рангах | Хеттманспергер Т. | 1987 |
| 1. Непараметрические методы статистики | Холлендер М., Вулф Д. | 1983 |
| 1. Элементарная теория статистических решений | Чернов Г., Мозес Л. | 1962\* |
| 1. Теория вероятностей, мат. статистика… | Шторм Р. | 1970\* |

© **От автора** 

Конспект содержит расширенное содержание лекций и семинаров по курсу "Математическая статистика" для специальностей "Финансы и кредит" (набора 1995 г.) и соответствует сокращенной программе (18 часов лекций, 18 часов семинаров). Онтодидактическое назначение курса – создание логико–математической базы для изучения курса "Основы теории систем и системного анализа", а также для курса "Экономическая статистика", от которого и была, по сути дела, отделена примерно третья часть под данный курс. Назначение его, кроме отмеченного выше, – обеспечить запас фундаментальных знаний, необходимый для восприятия еще двух дисциплин цикла информационных технологий ­– "Компьютерная техника и программирование" (в части курсового проекта) и "Основы теории информационных систем".

Проф. Корнилов Г.И.

март 1997 г.Оглавление

1. Введение в курс

1.1 Основные определения

1.2 Вероятности случайных событий

2. Распределения вероятностей случайных величин

2.1 Шкалирование случайных величин

2.2 Законы распределений дискретных случайных величин.

2.3 Односторонние и двухсторонние значения вероятностей

2.4 Моменты распределений дискретных случайных величин.

2.5 Распределения непрерывных случайных величин

2.5.1 Нормальное распределение

*2.5.2 Распределения выборочных значений параметров нормального распределения*

3. Взаимосвязи случайных величин

3.1 Парная корреляция

3.2 Множественная корреляция

4. Проверка статистических гипотез

4.1 Понятие статистической гипотезы

4.2 Критерии статистических гипотез

4.3 Ошибки при проверке статистических гипотез

5. Выборочные распределения на шкалах Int и Rel

5.1 Оценка наблюдений при неизвестном законе распределения

5.2 Оценка наблюдений при известном законе распределения

5.2.1 Оценка параметров нормального распределения

5.2.2 Оценка параметров дискретных распределений

6. Выборочные распределения на шкале Nom

6.1 Случай двухзначной случайной величины, N<50

6.2 Случай двухзначной случайной величины, N>50

6.3 Случай многозначной случайной величины

7. Выборочные распределения на шкале Ord

8. Материал семинарских занятий

8.1 Введение в комбинаторику

8.2 Методы вычисления моментов распределений

8.3 Алгоритмы простейших статистических расчетов

8.3.1 Вычисление моментов выборочных распределений

8.3.2 Проблема переполнения

8.3.3 Моделирование законов распределения

9. Литература