Министерство образования и науки

Российской Федерации

Национальный исследовательский университет «МЭИ»

Горицкий Ю. А.

**ВВедение**

**В Математическую статистику**

Учебное пособие по курсу

«Теория вероятностей и математическая статистика»

для студентов, обучающихся по направлению

«прикладная математика»

Москва

Издательство МЭИ

2015

УДК 519

Г692

*Утверждено учебным управлением МЭИ*

*в качестве учебного пособия для студентов*

Подготовлено на кафедре Математического моделирования

Рецензенты: докт. техн. наук, проф. А.Б. Фролов,

докт. техн. наук, проф. В.М. Беседин,

докт. техн. наук, проф. Ю.П. Кораблин

**Горицкий, Ю. А.**

Г692 Введение в математическую статистику: учебное пособие / Ю.А. Горицкий. М.: Издательство МЭИ, 2015. — … с.

ISBN

Пособие является конспектом лекций по основам математической статистики и содержит следующие разделы: точечное оценивание, доверительные границы и интервалы, проверка статистических гипотез, различение гипотез, элементы регрессионного анализа и метод статистических испытаний.

**УДК 519**

**ISBN**  © Национальный исследовательский

университет «МЭИ», 2015

**СОДЕРЖАНИЕ**

**ВВЕДЕНИЕ**

**ГЛАВА 1. ОЦЕНИВАНИЕ НЕИЗВЕСТНЫХ ПАРАМЕТРОВ**

**§1. Основные понятия и характеристики качества оценок**……………

**§2.** **оценивание вероятностей и моментов**…………………………………..

2.1. Оценка неизвестной вероятности случайного события………………

2.2. Оценка неизвестной функции распределения. Основная теорема статистики……………………………..……………………………...

2.3. Простейшие оценки моментов…………………………………..……

2.4. Линейная оценка среднего с минимальной дисперсией при разно точных измерениях………………………………………………..

**§3. Нижняя граница для дисперсии несмещенной оценки**

3.1. Информационное неравенство Рао-Крамера …………………

3.2. Информация Фишера……………………………………………………...

3.3. эффективные оценки. Экспонентные семейства распределений…….

**§4. Достаточные статистики**…………………………………………………

4.1. Предварительные соображения и определение…………………………

4.2. Критерий факторизации…………………………………………………..

**§5. Методы построения оценок**………………………………………………..

5.1. Метод моментов…………………………………………………………...

5.2. Метод максимального правдоподобия…………………………………

5.3. Метод порядковых статистик…………………………………………….

ГЛАВА 2. ИНТЕРВАЛЬНОЕ ОЦЕНИВАНИЕ

**§6. Доверительные границы и интервалы**………………………………….

6.1. Определения……………………………………………………………….

6.2. Пример. Доверительный интервал для среднего нормальной совокупности при известной дисперсии ……………

6.3. Способ построения доверительных границ и интервалов……………

6.4. Интервалы для параметров нормального распределения………….…

6.5. Построение центральной статистики…………………………………..

**§7. Интервалы при больших выборках** ………………………………..

7.1. Использование асимптотической нормальности оценок………………

7.2. Примеры……………………………………………………………………

**ГЛАВА 3. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ**

**§8. Критерий хи-квадрат Пирсона проверки гипотез** ………

8.1. Простая гипотеза о вероятностях………………………….

8.2. Сложная гипотеза о вероятностях……………………… .

8.3. Критерий хи-квадрат при полностью определённом гипотетическом распределении

8.4. Гипотеза о типе распределения………………………………

8.5. Проверка независимости бинарных признаков (таблица 2 × 2)

8.6. Обобщение. Проверка гипотезы о независимости признаков (таблица сопряженности признаков)…………………………………………….

8.7**.** Проверка гипотезы об однородности выборок……………………….

**§9. Критерий согласия Колмогорова**…………………………………………

**§10. Различение двух простых гипотез**……………………………………..

10.1. фиксированный объем наблюдений………………

10.2. Пример. Различение гипотез о среднем нормальной совокупности…………………………………………………………………

10.3. Последовательный критерий отношения вероятностей ……………

**ГЛАВА 4. РЕГРЕССИОННЫЙ АНАЛИЗ**

**§11. Элементы линейного анализа** ……………………….

11.1. Простая линейная регрессия…………………………………………….

11.2. Множественная регрессия. Схема Гаусса-Маркова …………………..

**ГЛАВА 5. МЕТОД МОНТЕ-КАРЛО**

**§12. Метод статистических испытаний**……………………………………..

12.1. Идея метода……………………………………………………………..

12.2. Общая характеристика методов ……………………………………….

12.3. Способы получения случайных величин………………………………

12.4. Выбор числа испытаний…………………………………………………

**СПИСОК ЛИТЕРАТУРЫ**………………………………………………………

**ВВЕДЕНИЕ**

Математическая статистика предполагает, что имеются наблюдения *x ≡ (x1, x2 … xn)* случайного характера, закон распределения которых частично или полностью неизвестен. По наблюдениям нужно сделать те или иные выводы относительно неизвестного распределения. Задачи математической статистики можно условно разделить на три этапа:

— оценить неизвестные параметры (теория точечного оценивания);

— указать интервалы, в которых находятся неизвестные параметры (теория интервального оценивания);

— ответить на вопрос: можно ли считать, что неизвестное распределение обладает тем или иным свойством (теория проверки статистических гипотез).

При рассмотрении любой статистической задачи имеющиеся наблюдения *x* ≡ (*x*1, *x*2… *x*n) являются конкретными значениями многомерной случайной величины ξ ≡ (ξ1, ξ2…ξn).

**Глава 1. Оценивание неизвестных параметров**

**§1. Основные понятия и характеристики качества оценок**

В большинстве задач математической статистики предполагается, что наблюдения (ξ1, ξ2…ξn) являются *независимыми, одинаково распределенными* случайными величинами с функцией распределения *F*(*x*), которая является частично или полностью неизвестной. Вэтом случае они называются *выборкой объема n из совокупности, распределенной по F*(*x*)*.* Выборка — объект случайный. Закон распределения выборки легко определяется. Например, для непрерывного случая, если — плотность распределения одной компоненты, то плотность распределения выборки есть

.

*Любая функция наблюдений* ϕ(ξ1, ξ2…ξn) (произвольной размерности) *называется* ***статистикой****.* Примеры статистик:

 .

Пусть ξ1, ξ2…ξn — выборка из совокупности, распределённой по закону с функцией распределения *F*(*x*; *a*), зависящей от параметра *а*. Параметр *а* неизвестен, его необходимо оценить по выборке.

Функция наблюдений *â* = φ(ξ1, ξ2…ξn), с помощью которой оценивается неизвестный параметр, называется ***оценкой*** или ***оценивающей функцией***. Заметим, что результат применения оценки *â* = φ(ξ1, ξ2…ξn) является случайной величиной.

Укажем те качества, которыми характеризуется оценка (оценивающая функция).

1. Несмещённость.

Оценка *â* = φ(ξ1, ξ2…ξn) называется ***несмещённой***, если *при любом значении параметра* *а*

Мϕ = *а*,

где М — математическое ожидание. В противном случае оценка называется смещенной, и

Мφ – *а* ≡ *b(а)*

называется *смещением*.

Выпишем условие несмещённости в интегральном виде для случая непрерывного распределения:

,

где  — плотность распределения выборки.

Обозначим ошибку оценки

δn = φ(ξ1, ξ2…ξn) – *a*.

В терминах ошибки несмещённость означает, что среднее значение ошибки равно нулю, т.е. при любом значении параметра *а*

Mδn = Мφ – *а* = 0.

2. Состоятельность.

Оценка *â* = φ(ξ1, ξ2…ξn) называется ***состоятельной***, если

φ(ξ1, ξ2…ξn) → *a*

(по вероятности) при *n* → ∞ при любом значении *а.* В терминах ошибки δ*n*состоятельность означает, что δn сходится к нулю по вероятности при любом значении *а*:

δn  0,

т.е. для любого сколь угодно малого ε > 0 P{|δn| < ε} → 1. Проверять состоятельность можно, используя *признак состоятельности*. Если

Mδn2→ 0 при *n* → ∞,

то оценка состоятельна.

Действительно, для любого ε > 0

P{|δn| < ε} = 1 - P{|δn| ≥ ε} ≥1 - → 1.

Здесь использовано обобщенное неравенство Чебышева:

 P{|ξ| ≥ t} при любом *p*>0; здесь *p* = 2.

3. Оптимальность.

Критерием качества оценки ϕ примем средний квадрат ошибки *R*ϕ(*a*) = M[φ(ξ1, ξ2…ξn) – *a*]2. Если оценка несмещённая, то критерий качества *R*ϕ(*a*) — это дисперсия оценки.

Оценка ϕ\* называется ***оптимальной***, если среди всех оценок ϕ она даёт минимальное среднее значение квадрата ошибки:

.

Если оценка несмещённая, то критерием качества является дисперсия.

Поскольку критерием качества оценки является не скаляр, а функция параметра, ясно, что оптимальной оценки может не существовать.

Используется и более общий подход к понятию оптимальности — критерием качества рассматривается , где  — потери (loss), которые несет статистик, если значение оценки — ϕ, а истинное значение параметра — *а*. Например, = 0, если , и 1 иначе; нетрудно увидеть, что в этом случае оптимальной оценкой будет та, для которой вероятность события  минимальна.

**§2.** **оценивание вероятностей и моментов**

**2.1. Оценка неизвестной вероятности случайного события**

Задачи математической статистики связаны с неизвестностью распределения наблюдений. Но распределение — это совокупность вероятностей, поэтому начнем с оценки неизвестной вероятности.

Пусть A — случайное событие, *p* *=* P(A) — его неизвестная вероятность. Пусть проведено *n* испытаний этого события, ν — количество появления события А, т.е. количество успехов. Рассмотрим в качестве оценки для *p* статистику

 , (1)

где  — относительная частота появления события А. случайная величина ν распределена по биномиальному закону *Bi*(*n, p*), и потому

.

Это означает, что оценка  несмещённая. Проверим состоятельность оценки:

,

где D — дисперсия. Т.е. оценка состоятельна.

**2.2 Оценка неизвестной функции распределения. Основная теорема статистики**

Пусть имеется выборка ξ1, ξ2…ξn с неизвестной функцией распределения *F*(*x*). Построим оценку для *F*(*x*). Зафиксируем произвольное значение аргумента *х*. Значение *F*(*x*) в точке *x* есть вероятность события Ах= {ξ<*х*}, т.е. *F*(*x*) = *P*(A*x*). Несмещённой и состоятельной оценкой для вероятности *P*(A*x*) — она же и оценка  для *F*(*x*) — является относительная частота (см. пример в п. 2.1):

 = (*x*; ξ1, ξ2…ξn), (2)



где νх — количество появления события Ах в *n* испытаниях, т. е. количество тех наблюдений ξ*i* в выборке, которые меньше *х* (ξi < *x*). Случайная величина νх имеет биномиальное распределение *Bi*(*n, F*(*x*)) с параметрами *n* и *F*(*x*). Имеет место несмещённость:

.

Также справедлива состоятельность:

.

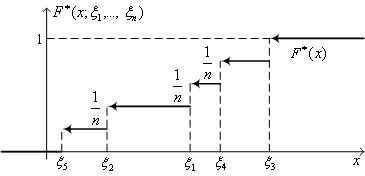
Итак, при любом *х* оценка (2) является несмещённой и состоятельной.

Функция  ≡ (*x*| ξ1, ξ2…ξn) называется ***функцией эмпирического распределения.***



**основная теорема статистики.** *Функция эмпирического распределения сходится (по вероятности) к истинной функции распределения:*

 ≡ (*x*; ξ1, ξ2…ξn). (3)



Справедливость этого утверждения показана предыдущими соотношениями. График функции эмпирического распределения показан на рис. 1:  кусочно-постоянная, делает скачок величиной 1/*n*, когда аргумент *x* переходит выборочное значение.

**Рис. 1. Функция эмпирического** Заметим, что если в *n* точках ξ1, ξ2…ξn

**распределения**  на оси *х* поместить равные вероятности 1/*n*, то получится некоторое дискретное распределение, называемое *эмпирическим,* и  — его функция распределения. Ясно, что первые два момента этого распределения таковы:

, (4)

.

В этих равенствах учтено, что при кусочно-постоянной интегрирующей функции интеграл Стильтьеса превращается в сумму. Второй центральный момент эмпирического распределения:

. (5)

Статистика  называется выборочным средним, а s2 — выборочной дисперсией.

**2.3. Простейшие оценки моментов**

Пусть имеется выборка ξ1, ξ2…ξn. Функция распределения *F*(*x*) наблюдений нам неизвестна.

*А. оценка математического ожидания.*

По определению математическое ожидание (первый момент) есть

.

Подставим в интеграл вместо *F*(*x*) несмещенную и состоятельную оценку (*x*| ξ1, ξ2…ξn). Получим (4):



.

Рассмотрим эту статистику в качестве оценки для математического ожидания *m*1. Проверим несмещённость:

.

Проверим состоятельность:

.

Таким образом,  является несмещенной и состоятельной оценкой для математического ожидания.

*Б. Оценка дисперсии.*

Дисперсия, согласно определению:

.

Вместо неизвестных *m*1 и *F*(*x*) подставим состоятельные оценки  и. Получим оценку *s*2 (5) для σ2:

*s*2 = .

Проверим несмещённость, для чего сначала преобразуем выражение для s2:

*s*2 = =

. (5а)

Здесь учтено, что . Определим математическое ожидание:

М*s*2≠σ2.

Оно не равно σ2, и потому оценка смещенная. Ясно, что ее можно исправить, умножив на константу, обратную к коэффициенту при σ2. Рассмотрим оценку

*s*12 = *s*2 = .

Эта оценка является несмещенной:

M*s*12 = M*s*2 = σ2.

Обе оценки *s*2 и *s*12 являются состоятельными, что видно из (5а), используя свойства сходимости по вероятности, аналогичные свойствам сходимости числовых последовательностей.

*В. Оценка моментов порядка k* > 2.

Для начального момента порядка *k >* 2

*mk* = Mξ*k*= ,

рассмотрим оценку, полученную заменой *F*(*x*) на (*x*| ξ1, ξ2…ξn):



.

Она является несмещенной:

.

Можно показать, что оценка состоятельна, т.е.

 .

Для центрального момента порядка *k >* 2

*μk* = M(ξ - *m1*)*k*= ,

рассмотрим оценку

.

Можно показать, что данная оценка состоятельна, т.е.

,

но несмещенной она не является.

**2.4. Линейная оценка среднего с минимальной дисперсией**

**при разноточных измерениях.**

Пусть ξ1, ξ2…ξ*n* — результаты *n* независимых наблюдений, причем среднее *m* для всех одно и то же, а дисперсии, т.е. точности измерений, различны:

Mξ*i* = *m*, Dξ*i* =  = σ2/*gi* .

Здесь зависимость дисперсии от номера *i* помещена в знаменатель. Величина σ2 может быть неизвестной. Требуется построить для *m* линейную несмещенную оценку

, M= *m*, (6)

с минимальной дисперсией:

. (7)

Условие несмещенности (6)



дает условие на коэффициенты:

. (8)

Найдем условный минимум квадратичной формы (7) при условии (8). Для определения условного экстремума запишем функцию Лагранжа с неопределенным множителем 2λ:

*L*(*c*1, с2…*cn*, λ) = .

Найдем ее экстремум:

, *k* = 1, 2…*n,*  (9)

. (10)

Из (9) выражаем *сk*:

= *hgk*,

где *h =* λ */* σ2 — неизвестное значение, которое находим из (10):

.

Получаем нужную оценку:

=. (11)

Мы видим, что чем больше *gi* (меньше дисперсия ), т.е. точнее наблюдение, тем больше вес этого наблюдения в сумме (11). Заметим, что вес обратно пропорционален *квадрату* *точности* измерения. Также заметим, что не обязательно знать дисперсии , нужно знать лишь соотношения *gi* дисперсий.

Определим дисперсию оценки:

.

**§3. Нижняя граница для дисперсии несмещенной оценки**

**3.1. Информационное неравенство Рао-Крамера**

Оказывается, никаким выбором оценочной функции невозможно сделать дисперсию ошибки меньше, чем некоторое определенное значение.

Пусть  — результаты *n* наблюдений (не обязательно независимых), являющиеся конкретными значениями многомерной случайной величины ξ ≡ (ξ1, ξ2…ξn). Закон распределения  известен с точностью до параметра *a* (будем считать плотностью распределения, если  непрерывна, и вероятностью, если ξ дискретна). Мы рассматриваем различные несмещенные оценки  для  (функция  известна, но значение параметра *a* неизвестно). Как обычно, рассматривая статистические задачи, мы отвлекаемся от конкретного значения , считая его одной из возможных реализаций случайной величины ξ ≡ (ξ1, ξ2…ξn), а оценку  рассматриваем как случайную величину.

***Теорема*.** Для любой оценки , несмещенно оценивающей , при условиях, оговариваемых ниже, справедливо неравенство:

, (1)

где обозначено

=. (2)

Т. е. дисперсия несмещенной оценки не может быть меньше величины , которая определяется правой частью неравенства и которую можно вычислить заранее, если известен закон распределения. Условия, при которых справедливо это неравенство, выпишем после проведения выкладок, из которых будет ясно, что именно нужно потребовать.

Обозначим через  множество тех значений *x*, при которых , т. е. . условие несмещенности

 (3)

продифференцируем по *a*. Учитывая, что , получим

. (4)

Здесь имеются в виду многомерные интегралы. Продифференцируем по *a* условие

. (5)

Учтем, что. Затем, умножив результат на , получим

. (6)

Вычтем из (4) равенство (6):

. (6а)

Введем обозначения

,

.

Возведя (6а) в квадрат, получим:

. (7)

Применим к левой части (7) неравенство Коши – Буняковского – Шварца:

 (8)

Получим:

. (9)

Здесь первый множитель есть , а второй , так что (9) совпадает с (1).

Условия, которые нужно наложить на  для справедливости (1):

множество  не должно зависеть от *a*, иначе нельзя дифференцирование в (3) и (5) внести под знак интеграла;

функция  должна быть дифференцируемой по *a*;

интегралы в (9) должны существовать.

Следствие. Если оценивается само значение параметра *a*, т. е. , то

. (10)

3.2. Информация Фишера

Величина  в формуле (2) называется информацией Фишера, содержащейся в выборке ξ относительно параметра *a*.

Свойства информации.

1. О вычислении.

Справедлива следующая формула:

. (11)

Действительно,

,

.

1. Аддитивность информации Фишера.

Информация, содержащаяся в n независимых наблюдениях, равна сумме информаций, содержащихся в отдельных наблюдениях.

Действительно, пусть ξ = (ξ1, ξ2 … ξ*n*), где  — независимы и распределены по закону . Тогда, если обозначено *x* = (*x*1, *x*2…*xn*), то . Вычислим информацию в наблюдениях :

, (12)

где  — информация, содержащаяся в одном наблюдении . Если все наблюдения распределены одинаково, т. е.  для всех  (это означает, что мы имеем дело с выборкой (ξ1, ξ2 … ξ*n*)), то

, (13)

где — информация в одном наблюдении.

Следствие. Если ξ = (ξ1, ξ2 … ξ*n*) — выборка, то из (10) и (13) следует:

. (14)

Это означает, что в условиях, в которых справедливо неравенство Рао – Крамера, дисперсию оценки нельзя сделать убывающей быстрее, чем .

Замечания.

*1. Об обобщениях неравенства* (1).

Неравенство Рао – Крамера обобщалось в различных направлениях — снимались ограничения, в основном, первое ограничение. Различные обобщения, в том числе на многомерный случай, можно найти в книгах [5], [6].

1. *Об использовании неравенства*.

С помощью (1) можно выносить суждения о качестве имеющихся оценок. далеко не всегда удается найти оптимальную оценку, т.к. или она не существует (неравенство не утверждает, что существует оптимальная оценка), или, если существует, то трудно реализуема в вычислительном отношении.

Пусть из каких-либо соображений построена оценка ϕ. Как оценить ее качество? Если  близка к , это означает, что оценка «хорошая». Если же нет, то с некоторой осторожностью можно считать, что оценка ϕ «плохая», поскольку известно, что при широких условиях существуют оценки с дисперсией, асимптотически (с ростом *n*) приближающейся к значению  (см. раздел 5.2).

**3.3. Эффективные оценки. Экспонентные семейства**

**распределений**

*Оценка, дисперсия которой достигает нижней границы, определенной неравенством Рао – Крамера, называетя эффективной.*

Эффективность — это самое лучшее в смысле точности, что мы можем ожидать. Возникают вопросы: при каких условиях существуют эффективные оценки, и как их искать? Ответом на второй вопрос является следующее утверждение.

Утверждение. Эффективная оценка  для , если она существует, равна

, (15)

причем зависимость правой части от параметра *а* фиктивна.

Действительно, равенство в (8) возможно тогда и только тогда, когда

, (16)

где *С* = *С*(*а*) может зависеть от *а*. Подставив это в (6а), найдем *С*(*а*):

, .

Подставив это в (16) и сократив на , получим:

, (17)

что эквивалентно (15). Заметим, что слева в (15) стоит оценка, т.е. функция наблюдений. Следовательно, зависимость правой части от параметра *а* фиктивна.

Пример 1. Проверка на эффективность оценки для вероятности случайного события.

Пусть имеется случайное событие . Его вероятность *q* неизвестна. Проведено *n* независимых испытаний, число выпадений события ξ — случайная величина, распределенная по биномиальному закону  с параметрами *n* и *q*. Роль неизвестного параметра играет *q*. Рассмотрим оценку , ее дисперсия . Эффективна ли оценка ? Нижняя граница Рао – Крамера . учитывая, что

, *x* = 0, 1…*n*,

вычислим информацию *I*(*q*):

,

откуда = . Т. е. оценка  эффективна, лучшей несмещенной оценки не существует.

Возникает вопрос: какие распределения допускают существование эффективных оценок? Ответ получим, если соотношение (15) понимать как дифференциальное уравнение относительно *p*(*x*,*a*):

.

После интегрирования по *а* получим

.

Распределения такого вида называют *экспонентными*. Здесь существенно то, что множитель, зависящий от *а*, есть экспонента, аргумент которой есть произведение функции *А(а*) от параметра на функцию ϕ(*x*) от наблюдений. Если распределение представлено в этом виде, то ϕ(*x*) есть эффективная оценка для своего математического ожидания Mϕ(ξ) = *f*(*a*). Заметим, что нормальное распределение, распределения Пуассона, биномиальное, геометрическое, гамма являются экспонентными семействами распределений.

§4 Достаточные статистики

4.1. Предварительные соображения и определение

Достаточная статистика — понятие фундаментальное. Часто возникает следующий вопрос. Имеется большая совокупность ξ = (ξ1, ξ2 … ξ*n*) наблюдений случайного характера, по которой нужно делать какие-либо выводы относительно чего-то неизвестного; обозначим это неизвестное через *a*. можно ли сжать информацию, то есть хранить меньший объем данных, не потеряв при этом информацию об *a*?

Сначала ответим на предварительный вопрос: нужны ли нам наблюдения ξ, если распределение *p*ξ(*x*) для ξ от *a* не зависит? ответ очевиден: нет, не нужны, наблюдения ξ можем не хранить.

Следующий вопрос: имеется две совокупности наблюдений ξ и τ. Известно, что распределение *p*τ(*t,a*) для τ зависит от *a*; также известно, что *условное распределение* ξ *при условии известного значения τ от a не зависит.* Нужны ли нам в этом случае наблюдения ξ? Ответ очевиден: нет, не нужны, мы можем сжать информацию, выбросив ξ и оставив только τ, поскольку распределение *p*τ(*t,a*) зависит от *a*.

Пусть ξ ≡ (ξ1, ξ2 … ξ*n*) — вектор наблюдений, принимающий значения *x* ≡ (*x*1, *x*2…*xn*) и распределенный по закону , где *а* ≡ (*а*1, *а*2…*аk*) есть *k*-мерный параметр. Пусть есть функция  (нам интересны значения *r* < *n*). Образуем случайную величину . обозначим через  распределение τ. Теперь имеем пару случайных величин ξиτ. При дальнейших рассуждениях полагаем случайную величину **ξ** дискретной (следовательно, τ - тоже).

формула умножения вероятностей дает

.

Однако ясно, что слева написано распределение ξ, т.к. из события  следует событие :

.

В результате имеем соотношение

. (1)

Т.е. распределение для всей совокупности ξ есть произведение распределения статистики  на условное распределение ξ при условии известного значения τ. Пусть второй сомножитель , т. е. условное распределение ξ, при условии известного значения τ, от *a* не зависит. Это означает, что значение ξ ничего не добавляет к знаниям о параметре *a*, полученным на основании статистики , всю информацию об *a* содержит .можно отбросить ξ, оставив только . В этом случае  называется достаточной статистикой для *а*.

Если ξ непрерывна, то рассуждения остаются справедливыми, нужно лишь в (1) вероятности заменить на плотности

. (2)

Для условного распределения имеем

. (3)

Если условное распределение ξ при известном  не зависит от , то ξ можно не хранить, оставить только .

Определение. Статистика  называется **достаточной** **для *параметра a****,* если условное распределение ξ при условии известного значения  не зависит от *a*, т.е. если

.

Практический смысл достаточной статистики состоит в том, что любые статистические выводы о неизвестном параметре *a* можно делать без ущерба для качества, основываясь не на всех исходных данных, а только на достаточной статистике. Верно очевидное утверждение, что *для любого способа  обработки всей информации* ξ *существует другой, эквивалентный способ обработки, основанный на достаточной статистике* *.* Тривиальный эквивалентный способ состоит в следующем: по исходным наблюдениям ξ вычисляем достаточную статистику , затем с помощью генератора случайных чисел генерируем с.в. ξ’ с условным законом распределения , не зависящим от параметра *а.* При этом распределения для ξ’ и ξ совпадают. Применив исходную процедуру δ к ξ’, **, получаем результат, основанный на статистике , но он эквивалентен *.*

Пример 1. Пусть для определения параметра λ некоторого однородного пуассоновского потока (например, источника радиоактивного излучения) -кратно в течение промежутков времени одинаковой продолжительности  измеряется количество поступающих частиц. На языке математической статистики это означает, что имеется  независимых наблюденийξ1*,* ξ2…ξ*n* над случайной величиной ξ, распределенной по закону Пуассона с неизвестным параметром . Возникает вопрос: можно ли не хранить все значения *х*1, *х2*…*хn*, которые приняли случайные величиныξ1, ξ2…ξ*n*, а хранить только суммарное значение ? Другими словами, является ли  достаточной статистикой?

Для ответа на этот вопрос нужно определить условную вероятность (3) получения значений *х*1, *х2*…*хn* при условии, что их суммарное значение известно и равно :



. (4)

В приведенной выкладке учтено, что сумма  независимых пуассоновских случайных величин распределена по закону Пуассона с параметром, равным сумме *na* параметров. Поскольку условная вероятность (4) от  не зависит, то  является достаточной статистикой, и значения , …,  можно не хранить. Полезно отметить, что распределение (4) является полиномиальным с равными вероятностями , при *i* = 1, 2…*n*. Действительно, на отрезок, состоящий из *n* промежутков длиной , бросим независимо общее количество равномерно распределенных точек, и определим вероятность попадания , *х*2 …  точек на интервалы ∆*Т*1, ∆*Т*2…∆*Тn* одинаковой длины . Эта полиномиальная вероятность равна правой части равенства (4). Мы получили важный результат для теории вероятностей: *для простейшего потока при известном числе точек на отрезке, положения точек независимы и равномерно распределены*.

**4.2. Критерий факторизации.**

Необходимость вычислять условное распределение  (или распределение ), чтобы потом определить условное распределение делением  на , является весьма неудобным способом определения достаточности. Простой способ дает теорема, называемая критерием факторизации.

Теорема. Статистика T(ξ) достаточна для а тогда и только тогда, когда справедлива факторизация (представление):

.

Существенным в этой записи является то, что *множитель, зависящий от* а*, от x зависит только через ,* а понимается как плотность или вероятность.

Докажем утверждение для дискретного случая.

Достаточность.

Пусть  (иначе выписанная ниже вероятность равна 0). Рассмотрим условную вероятность:

.

Полученное выражение не зависит от а.

.

Необходимость.

Пусть  не зависит от а. Тогда в силу (1)

,

то есть справедлива факторизация (2).

Замечание. Пусть  — достаточная статистика. Любая функция , по которой можно получить , является достаточной. Другими словами, если , то  достаточна. Этот факт тривиально следует из критерия факторизации:

=.

В частности, если  и  связаны взаимооднозначно, то  — достаточна.

Пример 1. Пусть , ξ2 … — *n* независимых наблюдений над случайной величиной, распределенной по закону Пуассона, другими словами, , ξ2… — выборка из совокупности, распределенной по закону Пуассона. Распределение выборки имеет вид:

.

Видим, что множитель, зависящий от а, т.е., от *x* зависит только через .По критерию факторизации  является достаточной статистикой.

Пример 2. Пусть случайное событие *А* с неизвестной вероятностью *q* испытывается независимо *n* раз. Пусть результат ξ*i* -го испытания — случайная величина (*i* = 1, 2…*n*), определенная следующим образом:



Пусть  — число наступлений события *А* в серии из *n* испытаний. Выпишем распределение вероятностей для выборки , ξ2… , принимая во внимание, что любая из случайных величин  может принимать лишь два значения (*xi* = 0 или *xi* = 1):

.

Из этого выражения следует, что  является достаточной статистикой.

Пример 3. Пусть  — случайная величина, распределенная по равномерному закону на отрезке , . Плотность распределения (для ):



Пусть ξ1, ξ2…ξn есть *n* независимых наблюдений над ξ0. Плотность распределения выборки:



Если ввести функцию  — индикатор неотрицательности аргумента *z*, —определив ее так:



то плотность распределения выборки запишется в виде:

,

откуда по критерию факторизации следует, что  является достаточной статистикой.

Пример 4. Пусть ξ1, ξ2…ξn — выборка из совокупности, распределенной по нормальному закону с параметрами *m* и  (роль параметра *а* в критерии факторизации играет здесь двумерный параметр ).

Плотность распределения выборки имеет вид:

.

Из этого выражения по критерию факторизации следует, что двумерная статистика  является достаточной для , так же как и

, связанная с τ1 взаимно-однозначно и имеющая своими компонентами несмещенные оценки  и  для математического ожидания *m* и дисперсии . Поэтому *любые статистические задачи, связанные с нормальным распределением, можно решать, опираясь на оценки  и *.

**§5. Методы построения оценок**

Рассмотрим лишь три наиболее популярные метода.

**5.1. Метод моментов.**

Пусть ξ1, ξ2…ξ*n*  выборка, т.е. *n* независимых наблюдений над случайной величиной ξ0, имеющей функцию распределения *F*(*x;a*1, *а*2…*ak*), зависящую от неизвестных параметров *a* = (*a*1, *а*2… *aR*). Требуется оценить параметры. Идея метода состоит в том, чтобы *неизвестные параметры выразить через начальные моменты, а затем вместо моментов подставить несмещенные и состоятельные оценки моментов*.

Выразим *R* моментов через *R* параметров:

, *j* = 1, 2…*R.* (1)

Из этой системы равенств можно выразить параметры через моменты:

 *j* = 1, 2…*R.* (2)

Подставив вместо моментов *m1, m2…mR* оценки моментов , получаем:

, где , *k* = 1, 2…*R*.

Мы получили некоторые оценки .

Справедливы следующие свойства (см., например, [1], [3]):

1) *если функции gj (⋅), j = 1, 2...R, непрерывны, то оценки состоятельны;*

2) *если функции gj(⋅), j = 1, 2...R, дифференцируемы, а распределение при любом a имеет 2R моментов, то оценки  асимптотически нормальны:*

** ~ *N (aj*, .

В справедливости этих свойств нетрудно убедиться. Несмещённость оценок не гарантируется.

**Замечания.**

1. В равенствах (1) вместо первых *R* моментов можно использовать любые *R* моментов; важно лишь, чтобы система была разрешима относительно параметров.

2. Моментные оценки не всегда обладают высокой точностью. Однако, обычно они достаточно просты в вычислительном отношении.

**Пример 1**. Оценим дисперсию σ2 методом моментов. Дисперсия σ2 выражается через первые два момента:

σ2 =.

Подставив оценки моментов, получим оценку *s*2 для дисперсии:

*s*2 = . (3)

Последнее равенство нетрудно проверить:

.

Оценка (3) совпадает с оценкой *s*2, которая была проанализирована в разделе 2.3.

**Пример 2**. Оценка параметров равномерного распределения.

Пусть ξ1, ξ2…ξ*n —* выборка из совокупности, распределенной по равномерному закону R[*a, b*] на отрезке [*a*, *b*]. Оценим два неизвестных параметра *a* и *b*. Первые два момента выражаются через два параметра:

*m*1 = (*a + b*) / 2,

*m*2 - *m*12 = σ2 = (*b – a*)2 / 12.

В этих уравнениях относительно *a* и *b* заменяем неизвестные моменты выборочными, при этом во втором уравнении слева, исходя из (3), имеем s2. Получаем:

 = 2,

 =2*s*.

Откуда:

 = –*s*,

 = +*s*.

**5.2. Метод максимального правдоподобия.**

Пусть ξ1, ξ2…ξ*n* — выборка, *q*(*xi*; *a*) — плотность распределения одного *i*-го наблюдения (в дискретном случае *q*(*xi*; *a*) — вероятность принятия дискретного значения *xi*), *a* = (*a*1, *а*2…*aR*) — неизвестный параметр, *p*ξ(*x*;*a*) =  — распределение выборки *x* = (*x*1, *х*2…*xn* ).

*Функция p*ξ*(x*;*a), как функция параметра а, при фиксированном х, называется* ***функцией правдоподобия****.*

*Оценкой максимального правдоподобия (мп оценкой) а\* параметра а называется такое значение, при котором функция правдоподобия p*ξ*(x;a) достигает максимума:*

*а*\*: *p*ξ(*x*;*a*\*) = . (4)

Если максимум достигается во внутренней точке области определения функции, то *а*\* удовлетворяет системе уравнений:

, *i* = 1, 2…*R*. (4а)

Использование логарифма не изменяет точки максимума, но упрощает выкладки при независимых наблюдениях. Оценка *а*\*= *а*\*(*x*) является функцией наблюдений *x*. Чтобы подчеркнуть случайность аргумента, напишем *а*\*(ξ).

**Пример 1**. МП оценка параметров нормального распределения.

Пусть ξ1, ξ2…ξ*n* — выборка из нормальной совокупности *N*(*m*, σ2), здесь *а ≡* (*m*, σ2). Параметры *m* и σ2 неизвестны. Плотность распределения выборки:

*p*ξ(*x*; *m*,σ2) = .

Логарифм функции правдоподобия:

ln *p*ξ(*x*; *m*, σ2) = .

Система уравнений для определения оценок:



Из первого уравнения находим

*m*\* = ≡ . (5)

Из второго уравнения находим

(σ2)\* = . (6)

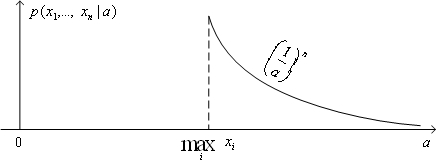
В данном случае оценки совпадают с выборочными средним и дисперсией.

**Пример 2.** МП оценка параметра равномерного распределения.

Пусть ξ1, ξ2…ξ*n* — выборка из совокупности, распределенной по равномерному закону R[0, *a*] с неизвестным правым концом *a* > 0. Плотность распределения выборки равна

*p*ξ(*x*1, *х*2…*xn* ; *a*) = 

При фиксированных *x*1, *х*2…*xn* функция правдоподобия убывает  при и равна 0 при *а* < (рис. 2). Максимум достигается при



*а\** = .

**Рис. 2. Функция правдоподобия**

Проанализируем эту оценку. Ее функция распределения:



Плотность распределения:

 иначе 0.

Математическое ожидание:

M*а\**= ,

т.е. оценка смещенная.

Ее дисперсия:

D*а\**=  – (M*а\**)2 = .

Оценку легко исправить, т.е. сделать несмещенной, умножив на , в результате чего получим оценку  = . Она уже несмещенная, М = *а* при любом *а,* и дисперсия

D = =.

Из вышесказанного видно, что дисперсия убывает быстрее, чем 1/*n*, что противоречит неравенству (14) раздела 3.2. Однако, в этом примере условия неравенства Рао-Крамера не выполняются, и дисперсия может убывать быстрее, что будет являться примером сверхэффективной оценки.

**Свойства** оценок максимального правдоподобия.

Пусть ξ ≡ (ξ1, ξ2*…*ξ*n*) *—* выборка объема *n* из совокупности, распределенной с плотностью *q*(*x;a*),и

*p(x*1*, х*2 *...xn ;a)* =  (7)

является плотностью распределения выборки.

При некоторых весьма широких условиях (см. ниже) оценки максимального правдоподобия:

— состоятельны;

— асимптотически эффективны;

— асимптотически нормальны.

Для одномерного случая:

M*а\** → *а, Dа\**→= при *n* → ∞ . (8)

условия, при которых вышеприведенные свойства верны, совпадают с условиями неравенства Рао-Крамера:

а) независимость от параметра *а* множества *X =* {*x: q(x/a) ≠* 0};

б) существование производных  и ;

в) существование интеграла .

Доказательство справедливости этих свойств можно найти, например, в [5]. Примем на веру состоятельность и покажем, как возникает асимптотическая эффективность и асимптотическая нормальность.

Рассмотрим случайную функцию от *а*

*Sn*(*a,*ξ) = . (9)

Учитывая (4а) и (7), ясно, что оценка *а*\* является корнем этой случайной функции

*Sn*(*а*\**,*ξ) = 0.

Пусть *а*0 — истинное значение параметра. Рассмотрим случайную величину *Sn*(*а,*ξ) в точке *а = а*0. Учитывая состоятельность, т.е. *а*\* → *а*0, и гладкость функции *Sn*(*a,*ξ), по теореме Лагранжа имеем:

*S*n(*а*0*,*ξ) = *S*n(*а*\**,*ξ) + (*а*0 – *а*\*) *S’*n(*,*ξ), (10)

где  — промежуточная точка между *а*0 и *а*\*, причем  → *а*0.

В силу предыдущего уравнения, справа первое слагаемое равно 0. Умножим это соотношение на :

*Sn*(*а*0*,*ξ) = (*а*0- *а*\*) *S’n*(*,*ξ), (11)

Слева имеем случайную величину ζ=*Sn*(*а*0*,*ξ), которая, учитывая суммирование случайных величин в (9), асимптотически нормальна *N*(0, *I*(*a*0)) с параметрами

Mζ =M*Sn*(*а*0*,*ξ) = .

При вычислении интеграла учтено, что

.

Дисперсия равна информации Фишера в одном наблюдении в точке *а*0:

Dζ = = *I*(*a*0).

Определим параметры случайной величины *S’n*(*,*ξ) в (11) при *n* *→* ∞ с учетом того, что  → *а*0:

M*S’n*(*,*ξ) =,

D*S’n*(*,*ξ) = .

Это означает, что *S’n*(*,*ξ) сходится к константе *I*(*a*0). Из (11) в пределе получаем

ζ= - (*а*0 - *а*\*),

что означает

*а*\* = *а*0 + .

Из этого следует, что оценка *а*\* асимптотически нормальна, а дисперсия {*nI*(*a*0)}-1. Это значение совпадает с границей Рао-Крамера.

**Замечания.**

1. *Эффективная оценка,**если она существует, является оценкой максимального правдоподобия*.

Действительно, если ϕ(*x*) — эффективная оценка для параметра *a*, то по лемме из раздела 3.3 имеем

,

откуда, приравнивая производную к нулю, получаем .

2. *Оценка максимального правдоподобия является функцией достаточной статистики, если последняя существует*.

Действительно, если *t*(*x*) — достаточная статистика, то в силу критерия факторизации в разделе 4.2 справедливо представление

*p*(*x*;*a*) = g(*t*(*x*), *a*)*h*(*x*),

и потому

*p*(*x*;*a*) = *h*(*x*)⋅**g(*t*(*x*), *a*),

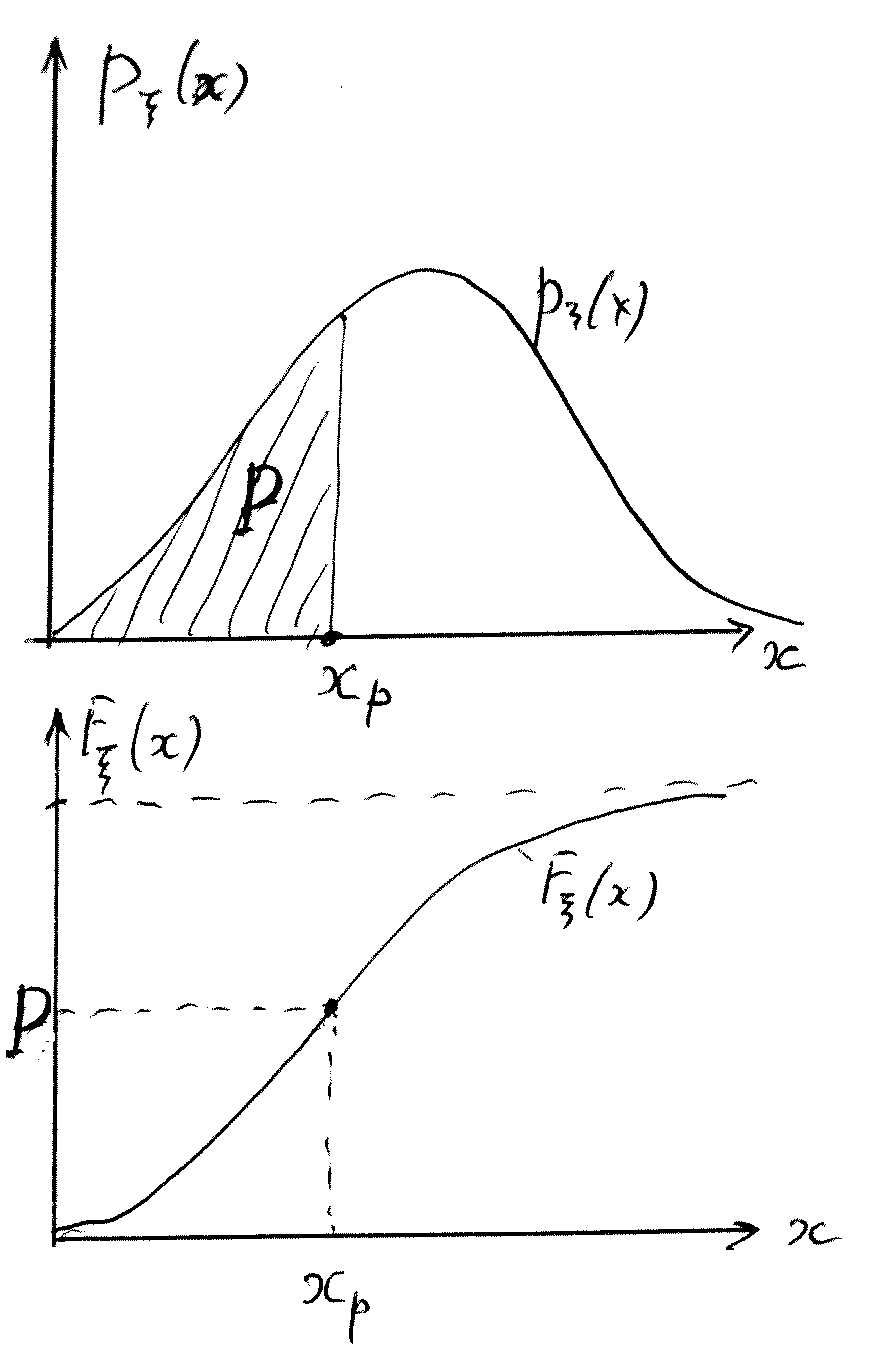
откуда экстремальная точка *a\** = *a\**[(*t*(*x*)].

**5.3.** **Метод порядковых статистик.**

В статистикешироко используется система числовых характеристик, называемых *квантилями*.

*Значение xp случайной величины ξ называется* ***p-квантилью****, если* P{ξ< *xp*} = *p*,

где *xp* — это корень уравнения *F*ξ(*xp*) = *p* (рис. 3).



Примеры р-квантили:

*x*0,5 — медиана — характеристика среднего значения случайной величины;

*x*0,98 — максимальное, в некотором смысле, значение случайной величины, т.к. P{ξ < *x*0,98} = 0,98;

*x*0,02 — минимальное, в некотором смысле, значение случайной величины, т.к. P{ξ ≥ *xp*} = 1 – P{ξ < *xp*} = 1– *p* = 0,98;

*x*3/4 и x1/4 — верхняя и нижняя квартили; их разность(*x*0,75 – *x*0,25) — межквартильная широта — служит характеристикой разброса.

**Рис. 3. Графическая иллюстрация квантили *xp***

*Оценка p-квантилей.* Неизвестные *p*-квантили легко оцениваются по выборке. Действительно, пусть *x*1*, х*2*...xn*— результаты *n* независимых наблюдений над случайной величиной ξ с функцией распределения *F*(*x*). Упорядочив их по возрастанию, получаем вариационный ряд *x*(1) ≤ *x*(2) ≤ *...* ≤ *x*(*n*). Чтобы подчеркнуть случайность ряда, запишем его греческими символами ξ (1) ≤ξ (2) ≤ *...* ≤ ξ(*n*). член вариационного ряда ξ (*i*) с номером *i* (заметим, что это случайная величина) называется *i*-й *порядковой статистикой*. по вариационному ряду построим функцию  ** эмпирического распределения, и, согласно общему принципу о том, что выборочные характеристики являются состоятельными оценками характеристик распределения генеральной совокупности, рассмотрим в качестве оценкидля *p*-квантили *xp* выборочную квантиль *ζp*, т.е. корень уравнения

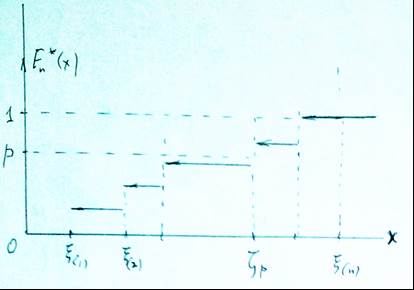
** = *p*. (8)

Поскольку  — функция кусочно-постоянная, то корнем является одна из порядковых статистик

ζ*p* = ξ ([*np*]+1), (9)

с номером [*np*]+1, т.е. целая часть числа *np* плюс 1(рис. 4).

Нетрудно показать, что ζ*p* *является состоятельной оценкой для* *xp.* Кроме того, известна **теорема Крамера**, которая гласит, что *для непрерывных распределений с плотностью q(x) оценка* ζ*p асимптотически нормальна с параметрами*:



**Рис. 4. Графическая иллюстрация выборочной квантили**

Mζ*p* = *xp*, Dζ*p* =. (10)

*Метод оценки параметров* основан на оценках ζ*p* при разных *p.* Пусть ξ1, ξ2…ξ*n* — выборка с функцией распределения *F*(*x*;*a*), зависящей от параметра *a,* значение которого требуется оценить. Выберем *р* так, чтобы квантиль *xр* зависела от параметра:

*xр = f*(*a*).

Выразим параметр *а* через квантиль *xр*:

*а* = *g*(*xр*),

и вместо *xр* подставим выборочную квантиль ζ*p* = ξ([*np*]+1), в результате чего получим *состоятельную оценку*

**= *g*(ξ ([*np*]+1)).

Таким же образом можно построить оценки и для неодномерного параметра. Основное и очень важное преимущество оценок, основанных на порядковых статистиках, — *их устойчивость к засорению наблюдений и к изменениям закона распределения*.

**Примеры** оценок параметров нормального распределения.Пусть ξ1, ξ2…ξ*n* — выборка из нормальной совокупности *N*(*m*, σ2).

*1) Оценка среднего m.* Известно или нет значение σ — безразлично. В силу симметрии нормального распределения параметр *m* является медианой, т.е. квантилью уровня ½, и потому может быть оценен выборочной медианой:

= ζ½ = ξ([*n/2*]+1).

Можно сравнить по точности эту оценку с эффективной оценкой  с дисперсией . согласно (10), D≈ , т.е. очень простая и устойчивая к засорению оценка  уступает по точности оценке в  раза, т.е. 25 %.

*2) Оценка стандартного уклонения* σ.

Легко проверить, что верхняя и нижняя квартили равны соответственно *x*3/4 = *m* + 0,675σ и  *x*1/4= *m* – 0,675σ, откуда

σ = (*x*3/4 - *x*1/4) / 1,35,

и потому оценивать σ можно следующим образом:

.

*3) Оценка стандартного уклонения* σ *по размаху*.

Пусть ξ (1) и ξ (*n*) — минимальный и максимальный член выборки, разность которых называется размахом *w*:

*w* = ξ (*n*) – ξ (1).

Ясно, что M*w* = c(*n*)σ, и потому оценкой для σ может служить

 = *w/*c(*n*) = *k*(*n*) ∙ *w*,

где *k*(*n*) берем из статистических таблиц [4]. Ниже приведены значения коэффициента *k*(*n*)и коэффициента эффективности *eff* =, где  — нижняя граница Рао-Крамера, а также потеря точности (1–) ∙ 100, измеряемая в процентах, по сравнению с нижней границей Рао-Крамера.

Табл. 1. Значение коэффициентов *k* и *n*

|  |  |  |  |
| --- | --- | --- | --- |
| *n* | 2 | 5 | 10 |
| *k(n)* | 0,866 | 0,430 | 0,325 |
| *eff* | 1,000 | 0,955 | 0,855 |
| потеря точности, (1 –)100, % | 0 | 2,5 | 7 |

Для устойчивости оценки к засорению используют подразмахи *wm* порядка *m,* где *m =* 1, 2, 3…:

*wm* = ξ (*n-m+*1) - ξ (*m*),

так что оценка имеет вид:

 = *km*(*n*) *wm*.

Значение коэффициента *km*(*n*) берется из таблиц.

4*) Распределение порядковых статистик.* При анализе оценок, получаемых рассматриваемым методом, необходимо знать распределения порядковых статистик. Если распределение одного наблюдения ξ непрерывно с плотностью *p*(*x*) = *F’*(*x*), топлотность распределения для *k*-й порядковой статистики ξ (*k*) выражается следующей формулой:

,

которая получается вычислением вероятности события

,

означающего, что при *n*-кратном испытании случайной величины ξ событие ,вероятность которого , появится (*k*-1) раз, событие , вероятность которого (), появится (*n*-*k*) раз, и событие , вероятность которого , появится 1 раз.

**глава 2. Интервальное оценивание**

**§6. Доверительные границы и интервалы**

результатом применения оценки *(x*1*, х*2*...xn)* является одно числовое значение, которое не дает представления о точности, т.е. о том, насколько близко полученное значение к истинному значению параметра. Интуитивно ясно, что такое представление может дать, например, дисперсия оценки, так что истинное значение неизвестного должно находиться где-то в пределах

± *(*2÷4*)*

Внесем уточнения.

**6.1. Определения**

Пусть ξ ≡ (ξ1*,* ξ2*...*ξ*n*) — *n* независимых наблюдений над случайной величиной с функцией распределения *F*(*x*;*a*)*,* зависящей от параметра *a*, значение которого неизвестно.

Интервал *I*(ξ) *=* (*a1*(ξ)*, a2*(ξ)) со случайными концами (случайный интервал)*,* определяемый двумя функциями наблюдений, называется *доверительным интервалом* для параметра *a* с уровнем доверия *Р*Д (обычно близким к 1), если

****P**{*I*(ξ) ∍ *a*} ≡ ****P**{*a*1(ξ)< *a* < *a*2(ξ)} = *P*Д*,* (1)

т.е. если при любом значении параметра *a* *вероятность* (зависящая от *a*) *накрыть случайным интервалом* *I*(ξ) *истинное значение a* *велика* (не менее *Р*Д), и хотя бы в одной точке равна *P*Д*.*

Функция наблюдений *â*н(ξ1*,* ξ2*...* ξ*n*) (случайная величина), называется *нижней доверительной границей* для параметра *a* с уровнем доверия *Р*Д (близким к 1), если

****P**{*â*н(ξ1*,* ξ2*...* ξ*n*)< *a*}= *P*Д. (2)

Т.е. если при любом значении параметра *a* вероятность события {*â*н(ξ) < *a*} велика (не меньше *P*Д), и хотя бы в одной точке равна *P*Д.

Функция наблюдений *â*в(ξ1*,* ξ2*...* ξ*n*) (случайная величина) называется *верхней доверительной границей* для параметра *a* с уровнем доверия *Р*Д, если

****P**{*â*в(ξ1*,* ξ2*...* ξ*n*)> *a*}= *P*Д . (3)

Т.е. если при любом значении параметра *a* вероятность события {*â*в(ξ) > *a*} велика (не меньше *P*Д), и хотя бы в одной точке равна *P*Д.

Вероятность *Р*Д называют также *доверительной вероятностью*.

**6.2. Пример.** **Доверительный интервал для среднего нормальной совокупности при известной дисперсии**

Пусть ξ = (ξ1*,* ξ2*...*ξ*n*) — выборка из нормальной *N*(*a, σ2*)совокупности*.* Достаточной оценкой для *а* является

= (ξ1*,*ξ2…ξ*n*) = ≡‾ξ,

распределенная по нормальному закону *N*(*a,* ). Пронормируем её, образовав случайную величину

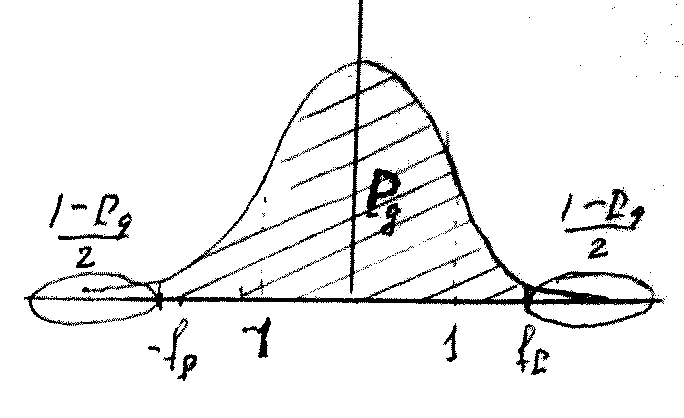
, (4)

которая распределена нормально *N*(0*,*1) при любом значении *а.*

По заданному уровню доверия *Р*Д (рис. 5) определим для ϕ симметричный интервал (-*fp,* *fp*) так, что

. (5)

Ясно, что *fp* есть квантиль порядка (1 + *Р*Д) / 2 стандартного нормального распределения *N*(0,1). Заметим, что ϕ зависит от *а*, и (5) верно при любом



**Рис. 5. Выбор интервалов при нормальном распределении**

значении *а*. Подставим в (5) выражение для ϕ из(4) и разрешим неравенство под знакомвероятности в (5) относительно *а*. получим соотношение

 (6)

верное при любом значении *а.* под знаком вероятности имеем две функции наблюдений

** и **,(7)

определяющие случайный интервал

*I*(ξ1*,* ξ2 *...*ξ*n*) = (,), (8)

который в силу (6) накрывает неизвестное значение параметра *а* с большой вероятностью *Р*Дпри любом значении *а*, и потому, по определению доверительно интервала, он является доверительным для *а* с уровнем доверия *Р*Д.

**6.3. Способ построения доверительных границ и интервалов**

Для построения доверительного интервала (или границы) необходимо знать закон распределения статистики ζ *=* ζ(ξ1*,* ξ2*...*ξ*n*)*,* по которой оценивается неизвестный параметр (такой статистикой могут быть сама оценка (ξ1*,* ξ2*...*ξ*n*), статистика, от которой зависит оценка , достаточная статистика или статистика, близкая к достаточной). Один из способов построения состоит в следующем.

1) Построим случайную величину ϕ = ϕ(ζ, *a*), зависящую от статистики ζ и неизвестного параметра *a* таким образом, что:

— закон распределения для ϕ известен и не зависит от *a*;

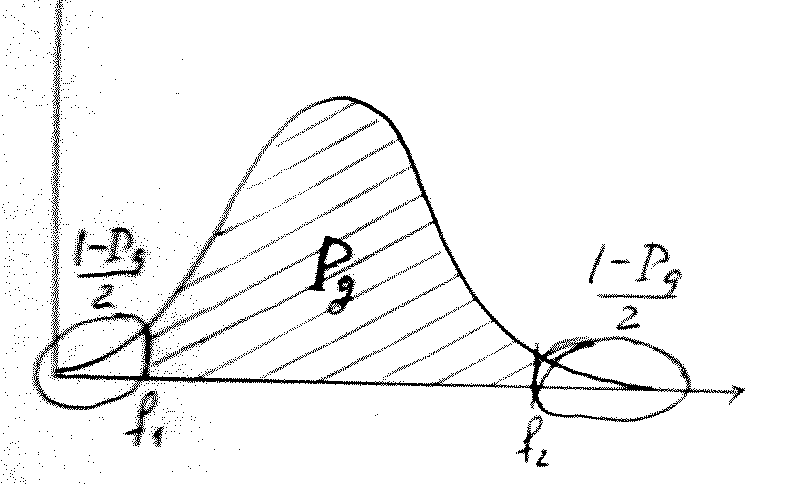
— ϕ(ζ*, a*) непрерывна и монотонна по *а*.

Такая случайная величина ϕ(ζ*, a*) называется *центральной статистикой*.

2) Выберем интервал (*f1, f2*) так, чтобы попадание в него случайной величины ϕ было практически достоверным (с вероятностью *P*Д):

*P*{*f1* <ϕ(ζ*, a*) < *f2*}= *P*Д, (9)

для чего достаточно в качестве *f1*и *f2* взять квантили распределения для ϕ уровня (1-*Р*Д)/2 и (1+*Р*Д)/2 соответственно (рис. 6).



3) Перейдем в (9) к другой записи случайного события, разрешив неравенства относительно

**Рис. 6. Выбор интервалов при произвольном распределении**

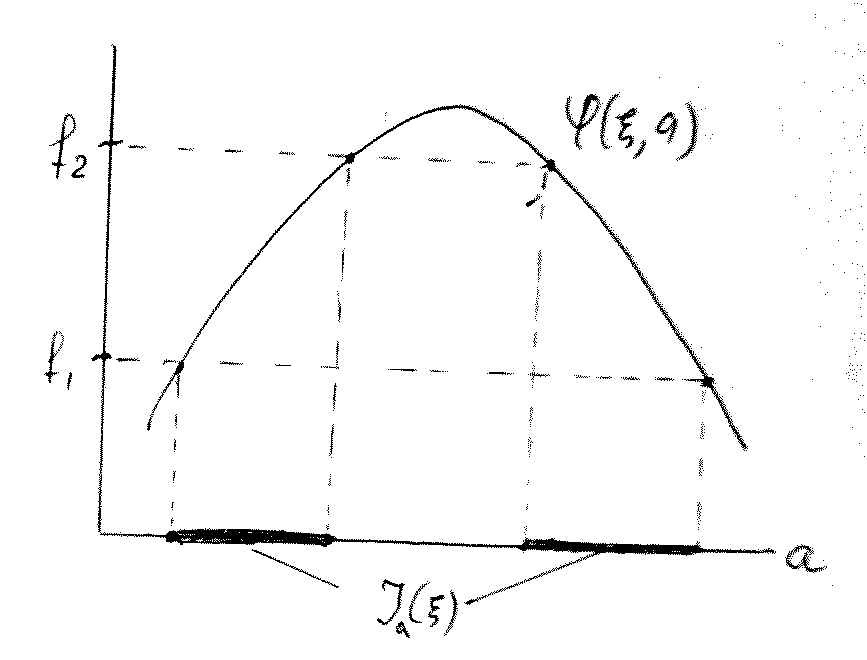
параметра *a*. Предполагая *монотонное* *возрастание* ϕ по *а* получим:

*P*{*g*(ζ*, f1*)< *a* < *g*(ζ*, f2*)} *= P*Д.

Это соотношение верно при любом значении параметра *a* (поскольку это так для (9)), и потому, согласно определению, интервал со случайными концами (*g*(ζ*, f*1)*, g*(ζ*, f*2)) является доверительным для *a* с уровнем доверия *Р*Д*.* Если имеем монотонное *убывание* ϕ по *а*, интервалом будет (*g*(ζ*, f*2), *g*(ζ*, f*1)).

**Замечания.**

1. сделанное выше предположение о монотонности ϕ(ζ*, a*) по *а*, не является существенным. В случае, если монотонности нет, результатом разрешения неравенств в (9) под знаком вероятности относительно параметра *a* является не интервал, а более сложное *доверительное множество*, например, два интервала (рис. 7).



**Рис. 7. Доверительное множество при отсутствии монотонности**

2. Если требуется построить *одностороннюю доверительную границу* (верхнюю или нижнюю), нужно использовать только одно из неравенств под знаком вероятности в (9). Используем

или **P**{ϕ(ζ*, a*) > *f*1} *= PД*, *f*1=*Q*(1 *- PД*),

или**P**{ϕ(ζ, *a*) < *f*2} *= PД ,* *f*2 = *Q*(*PД*)*,*

согласовав знак неравенства с характером монотонности ϕ по *а* (возрастание или убывание) и с характером границы (верхняя или нижняя). В неравенствах обозначено *Q(P)* — квантиль уровня *P*. После разрешения неравенства под знаком **P** получим односторонние доверительные границы для *a.*

3. О числовых значениях интервала. После применения формул для интервала, например, (8), к конкретным наблюдениям мы получаем конкретный интервал, например, для *n* = 9, при σ = 1, *Р*Д= 0,95 (соответствующее значение *f*P = 1,96) и *=* 3,87 получаем по (8) *a*1 = 3,22, *a*2 = 4,52.

Смысл замечания состоит в том, что нельзя писать аналогично соотношению (6)

,

поскольку под знаком вероятности нет случайного события. Параметр *а* не является случайной величиной, и интервал (3,22, 4,52) тоже не случайный, в отличие от формулы (6), где интервал является случайным. Коэффициент доверия *Р*Д= 0,95 — это *характеристика* *способа* определения интервала, но не конкретного интервала.

4. О точности интервального оценивания (о ширине интервала (6)).

а) Ошибка оценивания δ, в соответствии с (6),

δ=

не превосходит значения  с вероятностью *Р*Д < 1. При массовых вычислениях подобного рода точность не хуже  в числе случаев, доля которых *Р*Д. Уровень доверия *Р*Дозначает, что правило определения интервала дает верный результат свероятностью*Р*Д*,* которая обычно выбирается близкой к единице, однако, единице не равна.

б) Хотелось бы иметь высокую точность с высокой вероятностью. Однако, при увеличении *Р*Д значение *fp* тоже увеличивается, так что с высокой вероятностью можно гарантировать относительно низкую точность —чем надежнее гарантия, тем меньше она гарантирует.

в) Связь точности с σ и *n*: чем меньше σ, тем выше точность. Увеличение *n* также улучшает точность, которая изменяется как . Чтобы повысить точность в 10 раз, нужно число наблюдений увеличить в 100 раз.

**6.4. Интервалы для параметров нормального распределения**

*А. Распределение хи-квадрат c k степенями свободы.* Для рассмотрения типичных практических примеров потребуются сведения о некоторых распределениях. Многие задачи статистики связаны с распределением хи-квадрат (χ2(*k*)).

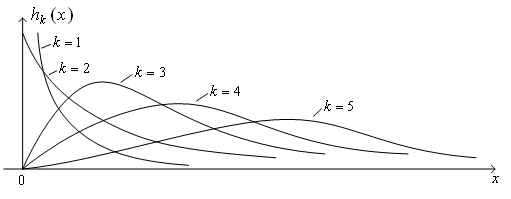
Пусть α1, α2…α*k* — независимые случайные величины, распределенные по стандартному нормальному закону *N*(0,1). Рассмотрим сумму их квадратов и обозначим соответствующую случайную величину через :

. (10)

Распределение этой случайной величины называют *распределением хи-квадрат с k степенями свободы*. Нетрудно показать (см., например, [2], §24), что плотность этого распределения выражается следующей формулой:

, *x* > 0, (11)

где *Сk* =  — нормирующий множитель,  — гамма-функция. На рис. 8 показаны графики при различных значениях *k*.



**Рис. 8. Семейство плотностей распределения χ2**

Заметим, что при *k* = 2 получаем показательное распределение:

.

Из соотношения (10) получаем первые два момента:

M = *k*, D = 2*k*,

откуда ясно, что с увеличением числа *k* степеней свободы распределение χ2(*k*) смещается вправо и расплывается, а также, что оно асимптотически нормально (в силу центральной предельной теоремы):

χ2(*k*) ~ *N*(*k*, 2*k*) при *k*→∞;

при *k* > 30 можно пользоваться таблицами нормального распределения.

Это распределение является частным случаем *гамма-распределения*, для которого плотность выражается формулой

*p*(*x*; λ*,a*) = *C*(λ*,a*)**, *x* >0, λ > 0, *a* > 0,

где *C*(λ*,a*) = - нормирующий множитель; λ – параметр формы, *a* – параметр масштаба,  - известная гамма-функция. Первые два момента *m*1 и σ2 равны соответственно

*m*1 = λ/*a*, σ2 = λ/*a*2.

Характеристическая функция *f(t)* этого распределения выражается формулой:

*f(t)* =  = = . (12)

Если λ *—* целое число, то распределение называется *распределением Эрланга*, которому подчиняется сумма λ независимых случайных величин, показательно распределенных с плотностью **.

Справедливость формулы (11) можно легко показать, определив характеристические функции для α12 и затем для  из соотношения (10). характеристическая функция для случайной величины  равна (1–2*it*)-*k*/2, откуда следует, что соответствующее распределение является гамма-распределением с параметрами λ = *k*/2, *a* = 1/2.

Б*. Совместное распределение выборочных среднего и дисперсии нормальной совокупности***.**

**Теорема.** Пусть ξ = (ξ1, ξ2…ξ*n*) — выборка из совокупности, распределенной по нормальному закону *N*(*m*, σ2), ‾ξ=— выборочное среднее, *s*2 =  — выборочная дисперсия.

Утверждения:

1) статистики‾ξ и *s*2 независимы;

2) случайная величина  подчиняется стандартному нормальному закону *N*(0, 1);

3) случайная величина *n*s2/σ2 подчиняется распределению χ2(*n*-1) — хи-квадрат с числом степеней свободы (*n*-1).

Доказательство. Перейдем к новым случайным величинам η*i* = (ξ*i*–*m*)/σ, *i* = 1, 2…*n*, которые образуют выборку η =(η1, η2…η*n*) из совокупности, распределенной по *N*(0, 1). Тогда

= (‾ξ–*m*)/σ,

== =; (13)

здесь предпоследняя сумма есть умноженная на *n* дисперсия выборочного распределения.

Преобразуем вектор η с помощью ортогонального преобразования с матрицей С: ζ = Сη, где первая строка матрицы С состоит из одинаковых элементов, равных 1 /. Дисперсионная матрица ζ, с учетом того, что M(ηηT) = I и С T= С -1, равна

Dζ= MСη(Сη)T = СM(ηηT)С T = I,

где I — единичная матрица, и потому ζ1, ζ2…ζ*n* — независимые случайные величины, распределенные по *N*(0, 1). Если учесть, что ортогональное преобразование не меняет расстояния, т. е.  =, а для первого элемента справедливо соотношение

ζ1 = = ,

то выражение (13) примет вид

== .

Последняя сумма *ns*2/σ2 распределена по закону хи-квадрат с (*n*–1) степенями свободы и не зависит от ζ1== , т.е. от‾ξ. Именно это утверждает данная теорема.

*В. Доверительный интервал для дисперсии нормальной совокупности***.** Пусть ξ = (ξ1, ξ2…ξ*n*) — выборка из совокупности, распределенной по нормальному закону *N*(*m*, σ2). Задан коэффициент доверия *P*Д. Параметр *а* может быть известен или не известен, поэтому рассматриваем два случая одновременно. в качестве несмещенных оценок для σ2 используем статистики

*s*2 = 

полученные методом максимального правдоподобия (с исправлением смещения во втором случае). Рассмотрим случайную величину

ϕ(ξ*, m,* σ) = , где *k* =

Очевидно, что в обоих случаях случайная величина ϕ подчиняется закону распределения хи-квадрат с *k* степенями свободы. Определим интервал (*f*1, *f*2) так, чтобы

**P**{*f*1 < ϕ < *f*2} = *P*Д.

В качестве *f*1 и *f*2 возьмем квантили уровней соответственно (1 – *Р*Д) / 2 и (1 + *Р*Д) / 2 распределения хи-квадрат с *k* степенями свободы:

(1-*Р*Д)/2, (1+*Р*Д)/2.

Разрешая под знаком вероятности два неравенства относительно σ

*f*1 < ϕ =< *f*2,

получим соотношение

= *P*Д

верное при любых значениях *m* иσ, откуда следует, что интервал  является доверительным для σ с доверительной вероятностью *P*Д.

**Пример.** Пусть среднее *m* неизвестно, *n* = 2, *P*Д = 0,95. Тогда

*s* = ,

и доверительный интервал весьма широк — (0,5*s,* 30*s*). При *n* = 10 доверительный интервал составляет (0,7*s,* 1,8*s*), при *n* = 20 — (0,87*s,* 1,17*s*).

*Г. Распределение Стьюдента***.** Многие задачи статистики приводят к рассмотрению следующей случайной величины. Пусть имеются две независимые случайные величины: α, распределенная по стандартному нормальному закону *N*(0,1), и , распределенная по закону хи-квадрат с *k* степенями свободы. Образуем новую случайную величину *Tk* следующим образом:

*Tk* =. (14)

Распределение этой случайной величины называется *распределением Стьюдента* (псевдоним английского статистика В. Госсета) с *k* степенями свободы и обозначается *S*(*k*). Плотность *sk*(*x*) распределения выражается формулой:

*sk*(*x*) = , - ∞ < *x* < ∞,

где Ck =  — нормирующий множитель. При *k* = 1 получаем распределение Коши с плотностью . При увеличении *k* знаменатель в (14) сходится к 1, поскольку математическое ожидание , а дисперсия , и потому распределение *S*(*k*) сходится к стандартному нормальному. При *k* > 30 для вероятностей p > 0,01 можно пользоваться таблицами нормального распределения.

*Д. Доверительный интервал для среднего нормальной совокупности при неизвестной дисперсии.*

Пусть ξ = (ξ1, ξ2…ξn) — выборка из совокупности, распределенной по нормальному закону *N*(*m*, σ2). Построим доверительный интервал с коэффициентом доверия *P*Д. параметр σ неизвестен, но именно он определяет точность оценки, т.е. ширину интервала, поэтому его тоже нужно оценить.

Пусть‾ξ=— выборочное среднее, *s*2 =  — исправленная выборочная дисперсия. В силу теоремы раздела 6.4, Б, нормированная погрешность ** распределена по стандартному нормальному закону, а (*n–*1)*s*2/σ2 — по закону хи-квадрат с (*n*–1) степенями свободы, и они независимы. Построим статистику *Tn*-1 делением

*Tn*-1 =  = .

Неизвестное значение σ сократилось. в силу (14) и теоремы раздела 6.4, Б эта статистика подчиняется закону Стьюдента с (*n–*1) степенями свободы. По заданному коэффициенту доверия *P*Д определяем симметричный интервал (- *tP*, *tP*) такой, что

**P**{- *tP <Tn*-1 < *tP* } = *P*Д.

Очевидно, что *tP* есть квантиль уровня (1+*P*Д) / 2. Разрешая под знаком вероятности два неравенства относительно параметра *m*

- *tP <*< *tP*,

получаем

**P**{‾ξ - *tP* *< а* <‾ξ+*tP*} = *P*Д.

Последнее соотношение верно при любых значениях параметров *m* и σ, и потому случайный интервал {‾ξ - *tP*,‾ξ + *tP*} является доверительным с вероятностью *P*Д.

**Замечание.** Сравним полученный интервал с интервалом (, ) из (8), построенным при известной дисперсии. Видно, что в полученном интервале вместо известного значения σ фигурирует оценка *s* для σ, и вместо квантили *fP* нормального распределения *N*(0,1) появилась квантиль *tP* распределения *S*(*n-*1) Стьюдента. Отметим, что при равных доверительных вероятностях *tP* > *fP*. В табл. 2 для примера приведены некоторые значения.

**Табл. 2. Сравнительные значения *fP* и *tP***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *P*Д | *fP* | *tP* | | | |
| *n* = 5 | *n* = 10 | *n* = 20 | *n* = 50 |
| 0.95 | 1.96 | 2.57 | 2.23 | 2.09 | 2.00 |
| 0.99 | 2.58 | 4.03 | 3.17 | 2.85 | 2.66 |

**6.5. Построение центральной статистики.**

В общем случаецентральную статистику ϕ(ξ,*а*) в разделе 6.3 можно построить, основываясь на следующей лемме.

**Лемма.** Пусть ζ — непрерывная случайная величина с функцией распределения *F*(*x*). Тогда случайная величина η = *F*(*ξ*) *равномерно* распределена на отрезке [0,1]. Действительно, определив функцию распределения *Fη*(*y*) случайной величины η:

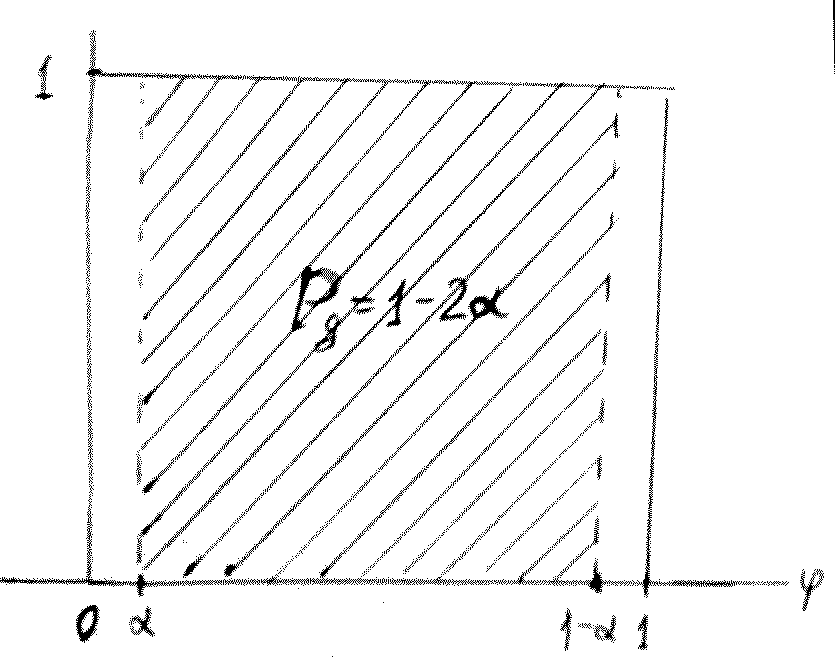
*Fη*(*y*) = *P*{η ≡ *F*(ξ) <*y*} = 

убеждаемся, что это так.

Пусть при построении доверительного интервала мы используем некоторую статистику ζ = ζ(ξ1, ξ2…ξn). Определим функцию распределения *F*(*z*,*a*)статистики ζ (*F* зависит от *а*). случайная величина ϕ = *F*(ζ,*a*), если *а* есть *истинное значение* параметра, в силу леммы, распределена *равномерно на отрезке* [0, 1] при любом истинном значении *а*, и потому мы можем ее принять в качестве центральной статистики ϕ(ξ,*а*) = *F*(ζ,*a*). в качестве (9) имеем соотношение (рис.9):

*P* {α<ϕ = *F*(ζ,*a*) <1-α} *=* 1 - 2α≡ *PД*,

справедливое при любом значении параметра *а*. Разрешив два неравенства под знаком вероятности относительно *а*, получим доверительный интервал.



Такой подход с некоторыми изменениями можно применить и для дискретных распределений.

**Рис. 9. Выбор интервала для *F*(ζ,*a*)**

Можно рассуждать иначе. При любом фиксированном значении истинного параметра *а* определим интервал (*z*1(*a*), *z*2(*a*)) так, чтобы

P{*z1*(*a*)< ζ< *z2*(*a*)} = *РД*. (15)

Ясно, что в качестве *z1* и *z2* можно взять квантили, т.е. определить *z1* и *z2* из условий

*F*(*z*1,*a*) = (1 *- РД* ) / 2*, F*(*z*2,*a*) = (1 *+ РД* ) / 2*.*

Если *z1*(*a*) и *z2*(*a*) монотонно *возрастают* по *а,* то, разрешив два неравенства под знаком Р в (15) и учитывая, что *z*1(*a*) < *z*2(*a*),получим

*P*{ *z*2-1(ζ) < *a* < z1-1(ζ) } = *РД*,

верное при любом *а. Я*сно, что интервал (*z2-1*(ζ) *, z1-1*(ζ)), определяемый двумя функциями от ζ, является доверительным с уровнем доверия *Р*Д. Если *z1*(*a*) и *z2*(*a*) монотонно *убывают* по *а*, то доверительный интервал получим таким же образом.

При построении *односторонних границ* нужно вместо двух неравенств использовать лишь одно, согласовав знак неравенства с типом границы (верхняя или нижняя) и с характером монотонности (убывание или возрастание).

**Пример 1.** Пусть ξ1, ξ2…ξ*n* — независимые наблюдения над нормальной *N*(*m*, σ2) случайной величиной. Пусть дисперсия σ2 известна. Построим доверительный интервал для среднего *m* с уровнем доверия *Р*Д = 1 – 2α. Оценкой для *m* является статистика‾ξ = , имеющая функцию распределения , где Ф(*x*) - функция стандартного нормального распределения. Согласно лемме, случайная величина  распределена равномерно на отрезке [0,1], следовательно,

= 1 – 2α = *P*Д

при любом истинном значении *m*. После разрешения двух неравенств под знаком вероятности получим

= *P*Д,

где учтено, что = - . Обозначив  = *fP*, получаем интервал (, ), совпадающий с (8).

**Пример 2.** Пусть ξ1, ξ2…ξ*n* — независимые наблюдения над случайной величиной, распределенной равномерно на отрезке [0, *a*], где *a* — неизвестный параметр, для которого нужно построить верхнюю доверительную границу с доверительной вероятностью *P*Д = 1 – α. Оценивающей статистикой является ζ =  — достаточная для *а* статистика. определим функцию распределения случайной величины ζ:

*F*ζ(*z*, *a*) = 

Случайная величина  распределена равномерно на [0,1], и потому при любом *а*

,

что означает, что статистика  является верхней доверительной границей для *а* с коэффициентом доверия *P*Д*.*

Верно также при любом*а*

,

т.е. статистика  является нижней доверительной границей.

*Общая логика* построения доверительного множества заключается в следующем. Для каждого значения параметра *а* построим множество Z(*a*) значений z случайной величины ζ вероятности *РД*:

P{ζZ(*a*)}= *Р*Д  *a*.

Далее для любого *z* построим множество A(z) значений *а,* включив в него те значения *а*, для которых Z(*a*) содержит *z*:

A(z) = {*a*: zZ(*a*)}, т.е. *a*A(z)  zZ(*a*),

и потому случайное множество A(ζ) содержит истинное значение *а* с вероятностью

P{A(ζ) ∍ *a*} = P{ζ  Z(*a*)} = *Р*Д  *a.*

**§7. Интервалы при больших выборках**

**7.1. Использование асимптотической нормальности оценок.**

Пусть по выборке ξ = (ξ1, ξ2…ξ*n*) оценивается неизвестный параметр *а*, и пусть оценка = (ξ) асимптотически нормальна со средним *а* и дисперсией , зависящей от неизвестного параметр *а*. Рассмотрим нормированную погрешность

.

Эта случайная величина распределена приближенно по нормальному закону *N*(0,1) при любом значении параметра *а*. По заданному значению доверительной вероятности *PД*определяем симметричный интервал(-*f*P,*f*P), который несет в себе вероятность *PД* нормального *N*(0,1) распределения, и потому при любом значении параметра *а* верно приближенное соотношение:

*P*{⎜ϕ(ξ*, a*)⎜ < *f*P}≈ *PД* .

Полагая монотонность ϕ по *a* и разрешая под знаком вероятности неравенства

, (16)

относительно *а,* получим соотношение

,

верное при любом значении параметра *а.* Это означает, что  является доверительным интервалом для параметра *а* с коэффициентом доверия, приближенно равным *PД*.

**7.2. Примеры**

*А*.*доверительный интервал для вероятности*. Пусть *Р*(А) = *а* — неизвестная вероятность некоторого события А, и ξ — число появлений А в серии *n* независимых испытаний. Несмещенной оценкой для *а* является , которая по теореме Муавра-Лапласа при больших значениях числа испытаний *n* является асимптотически нормальной с параметрами

, ,

и потому нормированная погрешность



распределена приближенно по нормальному закону *N*(0,1) при любом значении параметра *а*. Разрешая неравенство (16)

< *f*P2

относительно *а*, получаем доверительный интервал , где *а*1, *а*2 — корни соответствующего квадратного уравнения

*а*1,2 = .

При больших *n* формула упрощается:

*а*1,2 ≈ .

При малых значениях *n*, когда нельзя пользоваться приближенной нормальностью, пользуются статистическими таблицами, в которых для заданных *n* и *P*Д и полученному значению оценки указаны левый и правый концы интервала.

*Б.**Доверительный интервал для коэффициента корреляции.*Пусть имеется пара случайных величин (ξ, η), для которых по имеющейся выборке (ξ1, η1), (ξ2, η2)…(ξ*n*, η*n*) нужно определить доверительный интервал для коэффициента корреляции

.

Методом моментов получаем оценку для *r*:

*,*

где обозначено

, .

Если распределение случайных величин (ξ, η) является нормальным, оценка  распределена при больших *n* приближенно нормально [4], причем

М = *r*, D= .

Этого достаточно, чтобы определить приближенный доверительный интервал. Однако, более удобным является другой способ, основанный на преобразовании z Фишера:

*z* = *z*() = .

Эта статистика распределена приближенно (при *n*  20) по нормальному закону [4] со средним

*mz*(*r*)≈ 

и дисперсией σ2 ≈ 1 / (*n* – 3), а статистика

 ~ *N*(0,1)

приближенно нормальна. Доверительный интервал для *r* получаем из неравенства

,

где *f*P = Q((1 – *P*Д) / 2) — квантиль уровня (1 – *P*Д) / 2 нормального распределения. В статистических таблицах (например, [4]) даны интервалы для заданных *n*, *P*Д и . Для примера укажем, что при *P*Д = 0,95 и = 0,8 интервалы составляют:

(0,32; 0,95) при *n* = 10 и (0,53; 0,92) при *n* = 20.

**Глава 3. Проверка статистических гипотез**

Задачи проверки статистических гипотез возникают в ситуациях следующего общего вида. Есть предположение (гипотеза *Н*) о чем-то неизвестном, что непосредственно недоступно для наблюдения. Имеются данные наблюдений случайного характера, в законе распределения которых предположение отражается в виде некоторого свойства. Проверить гипотезу *Н* означает ответить на вопрос, обладает ли закон распределения наблюдений этим свойством. В первом приближении предполагается два возможных ответа: «да» или «нет».

Примеры возможных вопросов.

1. Есть ли связь между двумя признаками человека: ξ — цветом глаз и η — характером? Каждый признак имеет несколько уровней: *xi*, *i* = 1, 2…*m*, *yj*, *j* = 1, 2…*k*; каждый человек представлен парой значений *xi*, *yj*. Имеются данные по совокупности из *n* человек. По этим данным требуется проверить гипотезу Н (связи нет), что означает ответить на вопрос: равна ли вероятность *Р*(*xi , yj*) встретить любое сочетание (*xi , yj*) признаков произведению вероятностей *Р*(*xi*)*Р*(*yj*)?

2. Имеются данные о числе отказов ν0 устройства до модификации и ν*m* - после модификации, причем в последнем случае отказов меньше (ν*m* < ν0). Можно ли считать, что надежность увеличилась после усовершенствования, или наблюдаемое уменьшение числа отказов вызвано чисто случайными факторами? Этот вопрос сводится к проверке гипотезы о том, равны ли вероятности отказа в обоих случаях: *pm* = *p*0?

Один простой пример рассмотрим подробно.

3*. Гипотеза о симметричности монеты*. Проведено *n* бросаний монеты, при этом выпало *m* гербов. Можно ли считать, что монета симметрична в следующих трех случаях?

а) *n* = 10 бросаний, *m* = 6 гербов;

б) *n* = 100 бросаний, *m* = 60 гербов;

в) *n* = 1000 бросаний, *m* = 600 гербов.

В первом случае нет оснований подозревать не симметрию, поскольку имеем *типичный результат* при симметричной монете. В третьем случае, как подсказывает жизненный опыт, результат слишком сильно отличается от того, который был бы при симметричной монете, т.е. в окрестности значения 500. *Отклонение 100 от среднего 500 слишком велико*, оно противоречит предположению о симметрии. Второй случай промежуточный, и наш опыт не позволяет сделать уверенный вывод. Проведем рассуждения и расчеты.

Ясно, что если отклонение от среднего значения слишком велико, то следует признать, что монета несимметрична. Но что значит «отклонение слишком велико»? Это значит, что слишком мала вероятность такого отклонения, поскольку чем больше отклонение, тем оно менее вероятно. Определим вероятность получения на симметричной монете отклонения не меньшего наблюдаемого, и если она слишком мала, т.е. меньше, чем выбранное α, то гипотезу о симметрии отклоним.

Эту же мысль можно выразить иначе. Определим при симметричной монете типичный диапазон возможных значений (т.е. диапазон, в котором практически достоверно, с вероятностью (1 – α), должны находиться наблюдения). Если наблюдения не попали в этот диапазон, гипотезу о симметрии отклоним.

Пусть ξ — случайная величина, число выпадений герба. В наших трех случаях вероятности получения на симметричной монете отклонений(гипотеза *Н*), не меньших наблюденных 1, 10, 100, таковы:

а) P{⎢ξ - 5⎟ ≥ 1⎢Н} ≈ 0,75;

б) P{⎢ξ - 50⎟ ≥ 10⎢Н} ≈ 0,12;

в) P{⎢ξ - 500⎟ ≥ 100⎢Н} ≈ 10-8.

В случае (б) нет оснований считать монету несимметричной, поскольку вероятность 0,12 получить отклонение 10 или более не так уж мала; диапазон от 41 до 59 маловат для того, чтобы считать его практически достоверным, поскольку с немалой вероятностью 0,12 можно не попасть в этот диапазон. В случае (в) вероятность 10-8 слишком мала, чтобы верить в осуществление события ⎢ξ - 500⎟ ≥ 100.

Итак, действуем по следующей схеме. Предполагаем, что гипотеза Н истинна (монета симметрична). В этом предположении определяем вероятность отклонения от «нормы» (в данном случае, среднего значения) на величину наблюденного значения иди большего. Если вероятность слишком мала, меньше, чем некоторое α, гипотезу следует признать неверной (т.е. следует признать, что *наблюдения противоречат гипотезе*).

Величина α — вопрос выбора. Чтобы выбрать α, нужно ответить на вопрос: можно ли пренебречь возможностью ошибиться, если вероятность ошибки оценивается величиной α? Ответ зависит от тяжести последствий возможной ошибки. Например, если событие «опоздание на электричку» имеет вероятность 0,1, то мы можем пренебречь этим событием, т. к. потери в случае его осуществления невелики: ожидание не более десятков минут. Но если событие «опоздание на самолет в Австралию» имеет вероятность 0,1, то мы не будем пренебрегать этим событием, т. к. потери в случае его осуществления окажутся весьма большими.

**§8. Критерий хи-квадрат Пирсона проверки гипотез**

Критерий хи-квадрат является весьма общим методом построения тестов (процедур) для проверки различных гипотез. Чтобы воспользоваться этим критерием, выборочные данные предварительно группируют, т.е. переходят к частотному представлению данных. Рассмотрим исходную схему.

**8.1. Простая гипотеза о вероятностях**

Пусть результатом одного наблюдения могут быть *A*1*, А*2*...Am* — *m* возможных исходов. Обозначим:

*— p*1*, р*2*...pm* — соответствующие истинные (неизвестные) вероятности, , *n* — число независимых повторений опыта;

— *v*1*,v*2 *... vm* — число появлений соответствующих исходов в *n* опытах, ;

*— p, p...p* — *теоретические* (*гипотетические*) значения вероятностей, *p*>0*,* .

Требуется по наблюдениям *v*1*,v*2 *... vm* проверить гипотезу *Н* о том, что истинные вероятности *p*1*, ..., pm* имеют значения *p, p...p*, т.е.

*Н: pi = p* *, i=*1*, 2...m.*

Оценим по наблюдениям *v*1*,v*2 *... vm* вероятности *p*1*, p*2*...pm*. Пусть = ν1 */ n* *...*= ν*m / n —* оценки вероятностей. Мерой расхождения между теоретическими (гипотетическими) *p, p...p*и эмпирическими *,**...*вероятностями принимается величина

**,

которая с точностью до множителя *n* есть усредненное (при истинности *Н*) значение квадрата относительного отклонения оценок  от теоретических значений *p**.* Статистика *X2* называется статистикой хи-квадрат Пирсона. Для ее вычисления используются две эквивалентные формулы:

*.* (1)

Условно статистику *X*2 можно записать так:

,

где Н — наблюдаемые частоты *ν*i, Т — теоретические (ожидаемые) частоты *np*.

Поскольку по закону больших чисел *→ pi*при *n* → ∞, то

.

Последняя величина равна 0, если верна *Н* (*pi= p*); если же *Н* неверна, то она равна некоторому ε > 0, и тогда *X2* → *n*ε → ∞ при увеличении *n*. И потому процедура проверки гипотезы состоит в том, что если величина *X2* принимает «слишком большое (критическое)» значение *h*, т.е.

если *X2  ≥ h*, то гипотеза *Н* отклоняется. (2)

Если это не так, будем говорить, что «*наблюдения не противоречат гипотезе*». На вопрос, что означает «слишком большое» значение, отвечает теорема Пирсона.

**Теорема К. Пирсона.** Если гипотеза *Н* верна и 0 < *pi*0 *<*1, *i =* 1, 2...*m*, то при *n →* ∞распределение статистики *Х*2 асимптотически подчиняется распределению хи-квадрат с (*m* – 1) степенями свободы, т.е.

*Р*{*X2* < *x*/ *H*}→ *Fm-*1(*x*)≡ *P* {χ*2m-*1 < *x*}*.*

покажем, как возникает распределение хи-квадрат. Рассмотрим частный случай *m* = 2. Действительно, так как ν2 = *n -* ν1, *p*= 1 *- p* имеем

*,*

и статистика *X2* принимает вид

**. (3)

Если *H* верна, то ν1 подчиняется биномиальному распределению *Bi*(*n*, *p*) с параметрами *n* и *p*, а отношение , в силу теоремы Муавра-Лапласа, асимптотически нормально *N*(0,1). В правой части (3) имеем квадрат этого отношения, что означает сходимость соответствующего распределения к распределению хи-квадрат с одной степенью свободы. Для произвольного *m* теорема доказывается методом полной математической индукции.

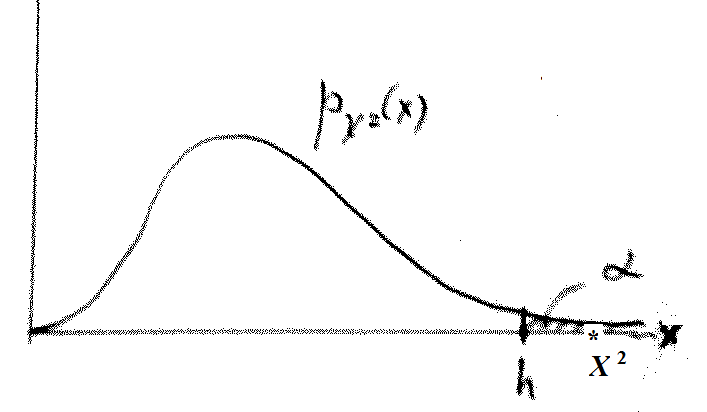
Теорема означает, что если гипотеза *Н* верна, то при достаточно большом *n* можно считать, что распределение статистики *X2* подчиняется хи-квадрат распределению. Порог (критическое значение) *h* в (2) выберем из условия, что *вероятность ошибки первого рода*, т.е. вероятность отклонения гипотезы, когда она верна, должна быть достаточно малой, т.е. равной выбираемому значению α — уровню значимости (рис.10):

*P*{отклонить *H*⎟*H* верна}= *P*{*X2*≥*h*⎟*H*} ≅ *P*{χ*2m-*1≥*h*} = α,

откуда

*h = Q*(1 *-* α*, m -* 1) (3a)

квантиль уровня (1 - α) распределения хи-квадрат с (*m* – 1) степенями свободы.



Решения (2) и (3а) процедуры проверки *Н* могут быть записана иначе в эквивалентном виде: гипотеза *Н* отклоняется, если

*P*{χ*2m-*1≥ *X2*}≤α, (4)

т.е. если мала вероятность

**Рис. 10. Выбор критического значения**

получения расхождения (при справедливости *Н*) не меньшего, чем в опыте. Вероятность слева в (4) называется минимальным уровнем значимости. при любом значении α, большем *P*{*X2m-*1 ≥ *X2*}, гипотеза отклоняется.

**Замечание**. Теорему Пирсона можно применять, если *n* > 50, все наблюдаемые частоты

νi ≥5, *i=*1*,* 2 *... m* (5а)

и теоретические частоты

*np* ≥ 10, *i=*1*,* 2 *… m.* (5б)

Если (5) не выполняется, необходимо объединять некоторые исходы из множества *A*1*, А*2*... Am*.

Пример. Имеется механизм, который предназначен генерировать случайную величину, принимающую с равными вероятностями  значения 0, 1...9. В табл. 3 приведены количества цифр, появившихся в результате  независимых наблюдений.

Табл. 3. Результаты наблюдений

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| цифры | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|  | 35 | 16 | 15 | 17 | 17 | 19 | 11 | 16 | 30 | 24 |
| - *np* | - 15 | - 4 | - 5 | - 3 | - 3 | - 1 | - 9 | - 4 | 10 | 4 |

Необходимо проверить гипотезу о том, что каждая цифра появляется с равной вероятностью. В этом примере , значение . Для уровня значимости  порог *h* равен 16,9, т.е.ь  = 0,05. Поскольку , гипотезу о равных вероятностях следует отклонить. Судя по табличным данным вероятности цифр 0 и 8 превосходят 1/10, а вероятность цифры 6 меньше .

**Свойство состоятельности критерия.** Решающее правило, описанное формулами (2) и (3а), обладает важным свойством *состоятельности:* *если гипотеза Н неверна, то с ростом числа наблюдений оно отклоняет гипотезу с вероятностью, стремящейся к* 1. Это свойство выражается следующим соотношением для мощности *W(p)* критерия:

*W(p)* ≡ .

Важную характеристику, мощность *W(p)*, можно определить приближенно, опираясь на следующую теорему.

**Теорема**. Если гипотеза неверна, то при *n* → ∞ распределение статистики *X*2 сходится к распределению  — нецентральному хи-квадрат с числом степеней свободы (*m* – 1) и параметром нецентральности *a*, причем

.

Эта формула получается из (1) заменой наблюдаемых частот νi на истинные *npi*.

**Справка** **о нецентральном распределении хи-квадрат** .

Пусть α1, α2…α*k* — независимы и нормально распределены по *N*(0,1). Составим случайную величину

,

где *a*1, *a*2…*ak* — произвольные константы. Нетрудно увидеть, что распределение этой случайной величины зависит не от *k* параметров *a*1, *a*2…*ak*, а только от суммы их квадратов:

.

Это означает, что составленную случайную величину можно представить в виде:

.

Первые два момента:

, .

При увеличении *k*, согласно центральной предельной теореме, распределение асимптотически нормально *N*(*k + a*, 2*k* + 4*a*). Если воспользоваться этим приближением, то для функции мощности приближенно будем иметь

*W(p)* = ≈.

Существует более точное приближение для  — приближение Патнайка. Оно основано на приближении  величиной , где *с* и *m* подбираются из условий равенства первых двух моментов:

, ;

результат подбора:

, .

Тогда функция распределения приближенно:

.

**8.2. Сложная гипотеза о вероятностях***.*

Пусть *A*1*, А2...Am*— *m* исходов некоторого опыта, *n* — число независимых повторений опыта, ν1*,* ν2*...*ν*m*— числа появлений исходов. Проверяемая гипотеза *Н* предполагает, что вероятности исходов *P(Ai)* являются известными функциями  параметра *a* (произвольной размерности *k*: *а =* (*a*1*, а*2*...ak*)), т.е.

*Н: Р(Аi) =* *, i =* 1*,* 2 *... m*,

но значение *а* неизвестно.

Пусть * —* оценка для *а,* минимизирующая значение статистики *X2.* Определим статистику

 . (6)

Оказывается справедливой теорема Фишера.

**Теорема Фишера.** Если *Н* верна, то при *n* → ∞ распределение статистики асимптотически подчиняется распределению хи-квадрат с числом степеней свободы *f = m* – 1 – *k,* и потому ***отклоняем*** *Н,* если

 ≥ *h,* (7)

где *h = Q*((1 –α)*, f*) — квантиль уровня (1 - α) распределения хи-квадрат с числом степеней свободы *f*. Такой порог обеспечивает выбранный уровень α вероятности *P* (отклонить *Н / Н*) ошибки первого рода. Если (7) не выполняется, делаем вывод, что ***наблюдения не противоречат гипотезе***.

Распределению хи-квадрат с *f = m* – 1 – *k* степенями свободы асимптотически подчиняется также статистика (6), если вместо оценок  по минимуму *X*2 используется  — оценка максимального правдоподобия для *а.*

Процедура (7) может быть записана иначе. Определяем по таблицам *P{*χ*f2* ≥}, и если

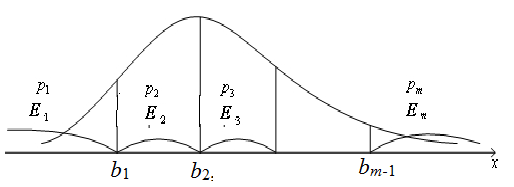
*P{*χ*f2* ≥*} ≤*  α , (8)

то гипотеза *Н* отклоняется.

**8.3. Критерий хи-квадрат при полностью определенном гипотетическом распределении.**

Пусть по выборке ξ1, ξ2…ξ*n* требуется проверить гипотезу *H* о том, что функция распределения случайной величины ξ равна *F*(*x*). Разобьем диапазон значений случайной величины ξ на *m* промежутков *Е*1, *Е*2…*Еm* без общих точек (рис. 11). Такое разбиение осуществляется числами *b*1, *b*2… *bm*-1. При этом правый конец каждого промежутка исключается из соответствующего промежутка, а левый — включается.

Обозначим через *Ai* событие «попадание наблюдения в *Еi*»:



*Ai =* {ξ ∈ *Еi*}, *i* = 1, 2…*m.*

Обозначим через *p* = *P*{*Ai*} *= P*{ξ∈*Еi*}, *i* = 1, 2…*m*. Пусть ν*i* — число членов выборки, попавших в *Еi*. Если верна *Н*, то *P*{*Ai*} = *p*. Таким образом, требуется проверить гипотезу о вероятностях *p, p... p* по наблюдениям ν1*,* ν2 *...* ν*m*. Задача сведена к задаче раздела 8.1. Обычно промежутки выбирают так, чтобы

**Рис. 11. Разбиение множества значений СВ**

вероятности были одинаковы и равны *p*=, причем, чтобы , то есть , однако, слишком малые значения *m* нежелательны, так как статистика не будет реагировать на изменение распределения внутри интервала.

**8.4. Гипотеза о типе распределения**

Пусть требуется проверить гипотезу о том, что выборка ξ1, ξ2…ξ*n* извлечена из совокупности, распределенной по некоторому закону, известному с точностью до параметров *а =* (*а*1*, а*2*...аk*), значения которых неизвестно, с функцией распределения *F*(*x*; *а*) (например, требуется проверить гипотезу о нормальности или о пуассоновости). Если разбить, аналогично разделу 8.3, диапазон значений случайной величины ξ на *m* промежутков *Е*1, *Е*2…*Еm* без общих точек, то в этом случае вероятности попадания наблюдения в *Еi*, т.е. *P*{*Ai*}= *p*(*а*), являются вполне конкретными функциями параметров, и возникает сложная гипотеза о вероятностях (см. раздел 8.2). Итак, наши действия таковы: получим значение оценки  методом максимального правдоподобия (или по минимуму *X*2), разобьем весь диапазон наблюдений на *m* интервалов, определим значения *νi* (число наблюдений в *i*-м интервале), определим вероятности или попадания в *i*-й интервал, вычислим (6) и примем решение по (7).

**8.5. Проверка независимости бинарных признаков (таблица 2×2)**

Пусть имеется *n* объектов, выбранных случайно из большой совокупности. Каждый объект характеризуется двумя признаками, которые могут или присутствовать, или отсутствовать в каждом из объектов. Возникает вопрос, имеется ли связь между этими признаками. Например, влияет ли форма собственности на рентабельность производства (*А* — частная собственность, ‾*А* — общественная собственность, *В* — рентабельно, ‾*В* — нерентабельно), влияет ли курение на легочные заболевания и т.д.?

Проверяется гипотеза *Н* о независимости признаков *A* и *B*. Если обозначить , , то гипотеза о независимости сводится к следующему: *Н*: , (9)

т. е. вероятность встретить сочетание признаков *А* и *В* равна произведению вероятностей встретить *А* и встретить *В*. В результате анализа на присутствие признаков могут появиться события *АВ,* *,* , . Пусть эти комбинации появились ν1*,* ν2*,* ν3*,* ν*4* число раз соответственно. Гипотеза о независимости *A* и *B* сводится к гипотезе о том, что вероятности этих событий

 = ,  =,

= , =  соответственно. Вероятности  и

**Рис. 11. Разбиение множества значений случайной величины ξ**

— два неизвестных параметра. Образуем случайную величину :

|  |  |
| --- | --- |
| =  . |  |

Решив задачу на минимум  по  и , найдем оценки, которые получатся такими, как мы ожидаем:

|  |  |
| --- | --- |
| . |  |

Подставив  и в получим:

.

При истинности *H* статистика  асимптотически распределена по  — хи-квадрат с одной степенью свободы. Пусть задано . Из условия  находим порог *h*. Итак, если , то гипотезу *H* отклоняем.

Приведем конкретные рабочие формулы. Для этого представим данные в следующей таблице 2×2:

|  |  |  |  |
| --- | --- | --- | --- |
|  | *A* |  |  |
| *B* |  |  |  |
|  |  |  |  |
|  |  |  |  |

Если провести необходимые выкладки, получим следующую формулу:

 , (10)

где в круглых скобках в числителе стоит квадрат определителя матрицы, а в знаменателе — произведение частных сумм (точка в обозначениях — суммирование по соответствующему индексу).

**Пример.** Медики испытывали очередную противогриппозную сыворотку на некоторой группе из 25 человек, 5 из которых не приняли сыворотку, 20 приняли. Из 5 человек, не принявших сыворотку, заболели 4, т.е. почти все, из 20 принявших заболели только 6 человек. Медики сделали вывод о том, что получили прекрасное подтверждение действенности сыворотки. Однако, проверим гипотезу Н о бесполезности сыворотки, т.е. о независимости признаков Б (заболеваемость) и С (принятие сыворотки).

Исходные данные записаны в таблице.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Б |  | Σ |
| С | 6 | 14 | 20 |
|  | 4 | 1 | 5 |
| Σ | 10 | 15 | 25 |

Вычисляем значение статистики по формуле (10)**:**

= .

Какую вероятность α ошибки мы можем допустить, если:

α = P{отклонить Н|Н верна} = P{сыворотка хорошая | сыворотка бесполезна }?

Если такая ошибка произойдёт, то сотни тысяч людей останутся без медицинcкой помощи. Значение α = 0,05 слишком велико в этой ситуации. Возможно, подходит α = 0,003. Соответствующий порог *h = Q*((1 – α)*,* 1) *= Q*(0,97*,* 1) = 32 = 9, и мы имеем < *h*, так что приходится сделать вывод, что *наблюдения не противоречат гипотезе о бесполезности сыворотки*.

**8.6. Обобщение. Проверка гипотезы о независимости признаков (таблица сопряженности признаков)**

Предположим, имеется большая совокупность объектов, каждый из которых обладает двумя признаками *А* и *В*. Признак *А* имеет *m* уровней: *A*1, *А*2...*Am*, а признак *В* — *k* уровней: *B*1, *В*2...*Bk* . Пусть уровень *Аi* встречается с вероятностью *P(Ai),* а уровень *Bj* — c вероятностью *P*(*Bj*). независимостьпризнаков *А* и *В* означает, что

P(*Ai Bj*) *=* P(*Ai*)*⋅*P(*Bj*)*, i =* 1*,* 2*...m, j =* 1*,* 2...*k*,

т.е. вероятность встретить комбинацию *AiBj* равна произведению вероятностей. Пусть признаки определены на *n* объектах, случайно извлеченных из совокупности; *νij* — число объектов, имеющих комбинацию *AiBj*, =*n*.

По совокупности наблюдений {ν*ij*} (таблица *m* × *k*) требуется проверить гипотезу *Н* о независимости признаков *А* и *В*. Задача сводится к случаю с неизвестными параметрами, которыми являются вероятности

*P(Ai), i =* 1*,* 2...*m* и *P(Bj), j =* 1*,* 2...*k,*

всего (*m –* 1) *+* (*k –* 1). Оценки этих вероятностей

, 

(точка в обозначениях — суммирование по соответствующему индексу) и статистика (8) принимает вид:

. (11)

Если гипотеза *Н* верна, то по теореме Фишера статистика  асимптотически распределена по закону хи-квадрат с числом степеней свободы

*f = mk –* 1 *–* (*m –* 1) *–* (*k –* 1) *=* (*m –* 1)(*k –* 1)*,*

и потому, если

**, (12)

то гипотезу о независимости признаков следует отклонить.

Ясно, что по (11), (12) можно проверять независимость двух случайных величин, разбив диапазоны их значений на *m* и *k* частей.

**8.7. Проверка гипотезы об однородности выборок**

Пусть имеется *k* выборок объемами *ni*, *i* = 1, 2…*k*, извлеченных из различных совокупностей. Измеряемая величина в каждой из выборок может иметь *m* уровней *Bj*, *j* = 1, 2…*m*. Пусть *pij* — истинная (неизвестная) вероятность получить *Bj* в *i-*й совокупности. Требуется проверить гипотезу *Н* о том, что исходные совокупности распределены одинаково, т.е. . Обозначим *νij* — число наблюдений в *i*-й выборке, имеющих уровень *Bj*, причем , . Имеем таблицу *k* × *m* наблюдений аналогично предыдущему разделу 8.5. Если *Н* верна, то имеем неизвестные параметры *q*1, *q*2…*qm*. Если бы они были известны, можно было бы составить статистику

*X*2(*q*1,…, *qm*) = **, (13)

причем  при истинности *Н* приближено подчинялась бы распределению хи-квадрат с (*m* – 1) степенями свободы, а вся сумма в (13) — распределению хи-квадрат с *k*(*m* – 1) степенями свободы в силу независимости выборок. После подстановки оценок  для теоретических вероятностей  количество степеней свободы уменьшится на (*m* – 1) и станет равным *f* = (*k –* 1)(*m* – 1). процедура проверки гипотезы примет вид, аналогичный (11), (12). Решающая статистика:

.

Если , то гипотезу о независимости признаков следует отклонить.

**§9. Критерий согласия Колмогорова**

Проверяется гипотеза *H* о том, что последовательность независимых наблюдений ξ1, ξ2…ξ*n* извлечена из совокупности с непрерывной функцией распределения . Построим вариационный ряд ≤≤ …≤ и функцию эмпирического распределения . Мерой качества согласования эмпирического и гипотетического распределения примем максимальное отклонение функции эмпирического распределения  от гипотетического , то есть верхняя грань модуля разности

.

Замечателен тот факт, что распределение статистики  при истинности *H* не зависит от  [4].

Процедура проверки гипотезы *H*:

— выбирается порог  (о выборе порога, см. ниже);

— вычисляется ;

— если , то *H* отклоняется.

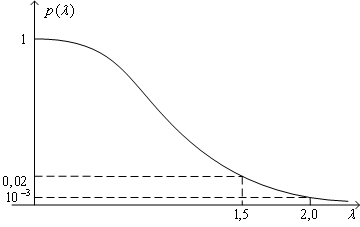
Критическое значение  выбирается заранее так, чтобы обеспечить заданную вероятность ошибки при истинности *H:*

.

В таблицах [4] имеются значения  для значений  = 1, 2, 5, 10, 20 %, . Там же имеются асимптотические формулы. Оказывается, что для любой  при  для распределения статистики  при истинности гипотезы  справедливо соотношение:

,

где  — функция распределения Колмогорова. Этот факт при *n* ≥ 20 позволяет простым способом приближенно обеспечить заданную вероятность ошибки первого рода (уровень значимости)



**Рис. 12. График функции *P*(λ)** .

График функции *P*(λ) условно показан на рис. 12. Итак,  выбирается по таблицам как функция α : , а затем вычисляется порог λ1 = λ(α) /. На рис.12 показаны две характерные точки функции P(λ): при α = 0,02  ≈ 1,5, при α = 10-3  ≈ 2,0. Для  хорошим приближением является , причем с ростом  точность улучшается.

**§ 10. Различение двух простых гипотез**

**10.1. фиксированный объем наблюдений**

Пусть имеется совокупность наблюдений *x =* ( *х*1, *х*2*...хn*)*,* являющихся реализацией случайных величин ξ ≡ (ξ1, ξ2…ξ*n*),относительно которой имеется два предположения (гипотезы):

1. *H*0*:* ξ распределена по закону *p*0(*х*)*;*
2. *H*1*:* ξ распределена по закону *p*1(*x*) (еслиξ— непрерывна, то *p*0(*х*)*, p*1(*х*) — плотности, если дискретна — вероятности).

По *х* требуется принять одно из двух решений: или«*верна Н*0» (это решение обозначим 0), или «*верна Н*1» (решение 1). Ясно, что дело сводится к определению решающей функции δ(*х*)*,* имеющей два значения 0 и 1, т.е. к определению разбиения *Г =* (*Г*0*, Г*1) пространства *Х* всех возможных значений *х*:

δ(*x*) = 

При использовании любой решающей функции δ(*х*) возможны ошибки двух типов:

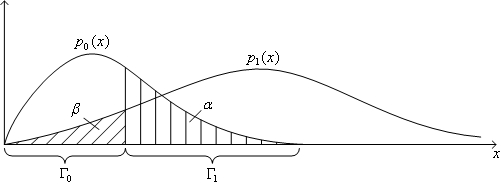
1. ошибка первого рода — принятие *Н*1 при истинности *Н*0;
2. ошибка второго рода — принятие *Н*0 при истинности *Н*1.

любая решающая функция характеризуется двумя условными вероятностями:

α *= Р*(принять *Н*1| *Н*0) *= ,* (1)

β *= Р*(принять *Н*0| *Н*1) *= ,*

которые называются вероятностями ошибок первого и второго рода соответственно (рис. 13). Хотелось бы иметь α и β близкими к нулю, но из (1) ясно, что если одна из них уменьшается, например, α (за счет уменьшения *Г*1*)*, то другая, β, увеличивается (за счет увеличения Г0; Г0 ∪ Г1 = *Х,* Г0 \ Г1 = ∅. Существуют различные подходы к определению оптимального правила.



**Рис. 13. Вероятности ошибок α и β**

*Байесовский подход*

Будем считать, что многократно сталкиваемся с проблемой выбора между *Н*0 и *Н*1. В этом случае можно говорить о частоте, с которой истинна *Н*0 (или *Н*1*)*, т.е. о том, что истинность *Н*0 (или *Н*1) — событие случайное, причем вероятность события, когда верна *Н*0, а когда верна *Н*1, известны:

*Р*(*Н*0) = *q*0*, Р*(*Н*1) = *q*1*, q*0 *+ q*1 *=* 1*.*

Кроме того, будем считать, что при каждой ошибке первого рода несем потери *W*0 (вероятность этого события *Р*(*Н*0) *Р*(принять *Н*1| *Н*0)), а при ошибке второго рода — потери *W*1 (с вероятностью *Р*(*Н*1) *Р*(принять *Н*0| *Н*1)). Если пользуемся правилом δ (с разбиением *Г*), то средний штраф от однократного использования:

*R*(Г) *= q*0⋅α(Г)⋅*W*0*+ q*1⋅β(Г)⋅*W*1*.* (1а)

Назовем правило δ (соответственно, разбиение Г ≡ (Г0, Г1)) в байесовском смысле оптимальным, если

*R*(Г) *= .*

Оказывается справедливой следующая теорема.

**теорема.** *оптимальным является правило, для которого область* Г1 *принятия гипотезы Н*1 *определяется соотношением*:

Г1 =  . (2)

Действительно, пусть T = (T0, T1) — произвольное разбиение; для него средние потери

*R*(*T*) *= q*0⋅α(*T*)⋅*W*0*+ q*1⋅β(*T*)⋅*W*1 = *q*0*W*0**+ *q*1*W*1*–*

– *q*1*W*1**+ *q*1*W*1**.

Здесь во второй строке добавлено и вычтено одно и то же слагаемое. После объединения интегралов получаем

*R*(*T*) *= q*0*W*0**+ *q*1*W*1.

Очевидно, для того чтобы получить минимальное значение средних потерь, в область интегрирования *T*1 нужно включить те точки, в которых подынтегральная функция в квадратных скобках отрицательна, т.е.

*,*

что дает в эквивалентной записи оптимальную область Г1  в (2).

**Замечание.** В частном случае, если *W*0 *= W*1 *=* 1*,* то *R*(Г) в (1а) имеет смысл безусловной вероятности ошибки, а соответствующее оптимальное правило называется правилом *“идеального наблюдателя”* или правилом Зигерта- Котельникова.

*Подход Неймана-Пирсона*

Оптимальным (в смысле Неймана-Пирсона) назовем такое правило, которое имеет заданную вероятность α ошибки первого рода, а вероятность β ошибки второго рода при этом минимальна, т.е. правило δ (соответственно разбиение Г) оптимально, если

β(Г) *= ,* (3a)

при условии α(Г*’*) ≤α0*.* (3б)

Оказывается справедливой следующая теорема.

**теорема.** *оптимальным является правило, для которого область* Г1 *определяется соотношением*:

Г1 *=* , (3в)

где *h* определяется из условия

α(*h*) *=* α0. (4a)

Действительно, во-первых, будем предполагать, что можно подобрать такое *h*, при котором в (4а) имеем равенство. В пользу этого говорит то, что α(*h*) является невозрастающей по *h* функцией: если *h*2 > *h*1, то Г1(*h*2) ⊆ Г1(*h*2), и потому α(*h*2)≤ α(*h*1), причем α(0) = 1 и α(∞) = 0 (при Г1(0) = X, Г1(*∞*) = ∅).

Далее, пусть T = (T0, T1) — произвольное разбиениеX, причем

α(Т) *= Р*(принять *Н*1|*Н*0, T) *= * ≤ α0. (4б)

Покажем, что β(T) > β(Г). Оценим разность, обозначив *u* = T0 ∩ Г0 — общую часть T0 и Г0, которую можно убрать из T0 и Г0 при оценке разности:

β(T) – β(Г) = *Р*(принять *Н*0| *Н*1, T) – *Р*(принять *Н*0|*Н*1, Г) =

=* - * =* -.*

Учтем, что из T0 убрали точки, входящие в Г0, так что A ≡ (T0 \ U) ⊆ Г1, из Г0 убрали некоторые точки, но B ≡ (Г0 \ U) ⊆ Г0, и потому в силу (3в) для точек из A справедливо , а для точек из B — . Учитывая эти неравенства, добавив  и учитывая (4), получаем

β(T) – β(Г) > **= * =*

= *h*{*Р*(принять *Н*0|*Н*0,T) – *Р*(принять *Н*0|*Н*0,Г)} =

= *h*{[1-α(Т)] – [1-α(Г)]} = *h*{α0- α(Т) ≥ 0.

**Замечание.** Приведенный результат есть частный случай фундаментальной леммы Неймана-Пирсона, справедливый при условии, что существует корень *h* уравнения (4а). Можно, однако, привести примеры, когда α(*h*)имеет скачки, и тогда (3в) требует некоторого простого уточнения.

**10.2. Пример. Различение гипотез о среднем**

**нормальной совокупности**

На вход канала связи подается сигнал *S*, который может принимать два значения:

*S =* 0 (сигнала нет), *S = а* ≠0 (сигнал есть).

В канале действует аддитивная случайная ошибка ε, нормально распределенная со средним *М*ε *=* 0 и дисперсией *D*ε *=* σ*2*; результатом является*х′= S +* ε. Измерения повторяются *n* раз, так что на выходе имеются наблюдения (*х*1*, х*2*...хn*)≡ *х*, по которым нужно решить, есть ли сигнал (*H*1*: S = a*) или нет (*H*0*: S =* 0)*.* Требуется построить решающее правило δ, имеющее заданную вероятность α0 ошибки первого рода (вероятность ложной тревоги)

α≡ *Р*(принять *Н*1⏐*Н*0) = α0

при минимальном значении вероятности β ошибки второго рода (вероятности пропуска).

считая ошибки независимыми, с учетом того, есть ли сигнал (*Н*1) или его нет (*Н*0), имеем

*р*1(*х*) *= *,  *р*0(*х*) *= .*

В соответствии с (3в), решение о наличии сигнала нужно принять, т.е. принять *Н*1, если *х* попадает в Г1, где

Г1===,

где обозначено .

Итак, если

, (5)

то принимается *Н*1, в противном случае принимается *Н*0*.* Порог *h*2 определяется из (4а):

α(*h*2) *= P*{принять *Н*1/*Н*0} = = α0.

если верна *Н*0, то  распределена нормально со средним 0 и дисперсией *n*σ*2*, и потому последнее условие принимает вид:

α(*h2*) *=* 1– Ф*=* α0*,*

откуда

*h2 =* σ*Q*(1 *–* *α*0)*,* (6)

где Ф(*х*) — функция нормального *N*(0, 1) распределения; *Q*(1 *–* α0) *—* квантиль порядка (1 – *α*0) этого распределения.

Определим вероятность β ошибки второго рода для процедуры (5) с порогом (6). Если верна *Н*1, то  распределена нормально со средним *na* и дисперсией *n*σ*2,* и потому

β = *P*(принять*.Н*0 */H*1)*= P* {**< *h2 /H*1} *=* Ф= Ф(*Q -* ).

Положим, *а* = 0,2, σ = 1,0 (т.е. ошибка σ в 5 раз больше сигнала *а*), *n* = 500, α = 10-2 ; при этом

*h2* = 1 ⋅2,33 = 52, β = Ф(2,33 – 0,2 ⋅ 22,4) = Ф(–2,14) = 1,6⋅10-2.

Как видим, вероятности ошибок порядка 10-2.

**10.3. Последовательное различение двух простых гипотез**

**(последовательный критерий отношения вероятностей Вальда**)

Многие практические задачи для достаточно надежного принятия решения требуют уменьшения времени набора данных, например: испытания надежности, военные задачи, оценка эффективности экономических и других сложных систем. Возникает задача разработать *ускоренные статистические процедуры*, имеющие заданные характеристики качества. Общая идея построения таких процедур состоит в том, чтобы не фиксировать объем наблюдений, а накапливать наблюдения до тех пор, пока не будет достаточно информации для получения выводов требуемой надежности. Соответствующие процедуры разрабатываются в разделе математической статистики, который называется *последовательным анализом*. Рассмотрим задачу, с которой началось это направление: последовательное различение двух простых гипотез.

Мы хотели бы пользоваться таким правилом различения, которое имело бы заданные уровни вероятностей ошибок α и β, и при этом требовало бы минимальное в среднем число наблюдений. Ниже такое правило описано.

Пусть *х*1*, х*2*...хn* — последовательность независимых, одинаково распределенных случайных величин. Относительно распределения имеется два предположения (гипотезы):

*Н*0: наблюдения распределены с плотностью *р*0 (*х*);

*Н*1: наблюдения распределены с плотностью *р*1(*х*); (если наблюдения дискретны, то *р*0(*х*)*, р*1(*х*) — вероятности, а аргумент дискретный).

После каждого наблюдения статистику предоставляется выбор из трех возможных решений:

— принять *Н*0и закончить наблюдения;

— принять *Н*1 и закончить наблюдения;

— не принимать ни одну из гипотез и продолжить наблюдения.

***А. Формулировка решающего правила и его оптимальность****.* Рассмотрим следующую процедуру δ\*, называемую *последовательный критерий отношения вероятностей* (ПКОВ). Зафиксируем два порога — верхний *А* и нижний *В*, при этом0 *< В <* 1 *< А*. Пусть уже получено *n* наблюдений (*n* = 1, 2...); запишем отношение правдоподобия:

*Ln*(*x*1*, ..., xn*) *= * ≡ *,* (7)

где обозначено ** Процедура δ\* на очередном шаге *n* такова:

1) если *Ln*(*x*1*, х*2*... xn*)≤ *В,* то принимается *Н*0 и наблюдения заканчиваюся;

2)если *Ln*(*x*1*, х*2*...xn*)≥ *A*, то принимается *Н*1и наблюдения заканчиваются; (8)

3)если *В* < *Ln(x*1*, х*2*...xn) < А*, то наблюдения продолжаются (т.е. делается еще одно наблюдение).

Очевидно, эта процедура характеризуется некоторыми вероятностями ошибок и средними числами наблюдений:

α *=* α(*А, В*) *= Р*{принять *Н*1 */Н*0}*,* β *=* β(*А, В*) *= Р*{принять *Н*0 */H*1}*,*

*n*0 *= n*0(*А, В*) *= М*(ν */H*0)*, n*1 *= n*1(*А, В*) *= М*(ν */H*1)*,*

где ν — число наблюдений (случайная величина) до принятия окончательного решения. Если α0 и β0 заданы, то можно найти пороги *А* и *В*, т.е. правило δ\*. Оказывается, такое правило обладает свойством оптимальности.

**Теорема** (Вальд и Вольфовиц, 1948 г.). Среди всех решающих правил δ′, обладающих свойством

α(δ*′*) ≤α0, β(δ*′*) ≤ β0, (9)

последовательный критерий отношения вероятностей δ\* имеет минимальные средние числа наблюдений:

*n*0(δ\*) ≤ *n*0(δ′)*, n*1(δ\*) ≤ *n*1(δ′). (10)

Заметим, что минимальность достигается сразу по двум характеристикам (см. [8]).

***Б. Связь порогов с вероятностями ошибок****.* Легко показать справедливость неравенств, связывающих пороги с вероятностями ошибок:

*А* ≤ *, В* ≥ *.* (11)

Действительно, обозначим через *Rn* множество тех последовательностей *x*1*, х*2*...xn* длины *n*, для которых проверочная процедура заканчивается на *n*-м шаге принятием *H*1:

*Rn* = .

Оценим α:

α *=* α(*А, В*) *= Р*{принять *Н*1 */Н*0} *= Р*{принять *Н*1 на *n*-м шаге | *Н*0} =

= *Р*{принять *Н*1 на *n*-м шаге | *Н*1} =

*Р*{принять *Н*1 | *Н*1} ,

что дает первое из неравенств в (11). Аналогично показывается справедливость второго неравенства введением S*n* множества тех последовательностей *x*1*, х*2*... xn* длины *n*, для которых проверочная процедура заканчивается на *n*-м шаге принятием *H*0:

*Sn* = .

***В. Выбор порогов****.* Вместо неизвестных значений *А* и *В* возьмем их приближенные значения *А′* и *В′*:

*А ≈ А′* =*, В ≈ В′ = .* (12)

при таком выборе порогов вероятности ошибок α*′* и β*′* не равны α и β. Для них справедливы неравенства (11):

**≥ *А’*=**, *≤ В’= *. (13)

Из них получаем связь между α*′*, β*′* и α, β:

*≤ ≤, ≤ ≤.*

Поскольку знаменатели справа незначительно меньше 1, значения α*′* и β*′*, если и превышают заданные значения α и β, то весьма незначительно. Более того, если неравенства (13), записанные в виде

**,

*≤ *

сложить, то получим *.* На самом деле, как показывает более детальный анализ [8], α*′* и β*′* меньше заданных вероятностей α и β.

***Г. Тождество Вальда.*** Пусть ζ1, ζ2 … ζ*k*-1 — последовательность независимых, одинаково распределенных случайных величин, и пусть ν — момент остановки наблюдений, причем случайное событие {ν≥ *k*} (быть или не быть следующему *k*-му наблюдению) определяется значениями ζ1, ζ2…ζ*k*-1, т.е.

{ν ≥ *k*} = *f*(ζ1, ζ2…ζ*k*-1).

Справедливо следующее утверждение:

M= MνMζ. (14)

Доказательство.

Введем случайные величины

η*k* =  =g(ζ1, ζ2, …, ζ*k*-1).

Тогда

= и

M=== = Mν⋅Mζ.

***Д. Среднее число наблюдений****.* Процедуру δ\*, определяемую соотношениями (8) и (7), можно представить так (рис.14):

если

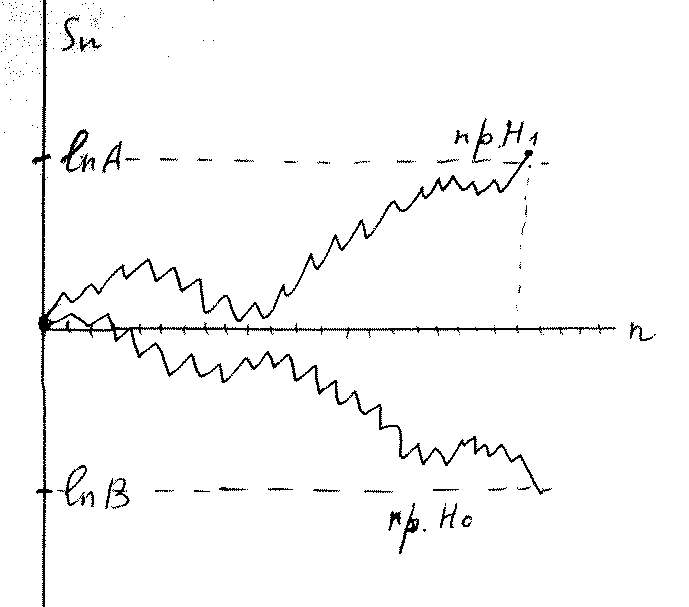
ln*В* < *Sn* ≡ ln*Ln* =*= <* ln*А*,

то наблюдения продолжаются; здесь обозначено ζ*k*=;

если *Sn* = ln*Ln* =**≤ln*В,* то принимается *Н*0, т.е. ν = *n*;

если *Sn* = ln*Ln*=**≥ln*A*, то принимается *Н*1, т.е. ν = *n*.

Слагаемые ζ1, ζ2…ζ*n* в сумме *Sn* = **являются независимыми, одинаково распределенными случайными величинами, при этом в момент остановки имеем *Sν* = **, и событие {ν ≥ *n+*1} определяется значениями ζ1, ζ2…ζ*n*. Тождество Вальда дает



**Рис. 14. Реализация процедуры Вальда**

M*Sν* = Mν⋅Mζ. (15)

С другой стороны, M*Sν* можно определить из условия, что в момент остановки *Sν* приближенно (с точностью до величины «перескока») принимает или ln*A* илиln*B*.

Если верна *Н*1, то

*Sν*  ≈ 

M(*Sν* ⎢*Н*1) ≈ .

С учетом (15) имеем

*n*1 ≡ *М*(ν /*H*1)*≈ *. (16)

Если верна *Н*0, то

*Sν*  ≈ 

M(*Sν* ⎢*Н*0) ≈ .

*n*0 ≡ *М*(ν /*H*0) *≈ *, (17)

где М(ζ/*Hi*) *=* , *i =* 0*,* 1.

**Замечание о выигрыше** по числу наблюдений. По заданным вероятностям α и β можно построить две процедуры: одна с фиксированным количеством наблюдений, другая — последовательная. Выигрыш второй по сравнению с первой по числу наблюдений зависит от соотношения α и β. В большинстве практических задач последовательная процедура выигрывает не менее чем в 2 раза (подробнее см. в [8]).

**ГЛАВА 4. РЕГРЕССИОННЫЙ АНАЛИЗ**

**§11. Элементы линейного анализа**

Регрессионный анализ возник в результате потребности прогнозировать одни величины по значениям других: метеоролог прогнозирует погоду по атмосферным измерениям, экономист прогнозирует прибыль по экономическим показателям, технолог — твердость стали по химическим добавкам. В эту область статистики входит широкий круг задач, связанных с построением (восстановлением) зависимостей между группами числовых переменных

*x ≡* (*x*1*, х*2 *... xp*)и *y=* (*y*1*, у*2*...ym*)*.*

Предполагается, что *x* — независимые переменные (факторы) влияют на значения *y —* зависимых переменных (откликов). По имеющимся эмпирическим данным (*xi , yi*), *i* = 1, 2, ..., *n* требуется построить функцию *f*(*x*), которая приближенно описывала бы изменение *y* при изменении *x*:

*y ≈ f*(*x*).

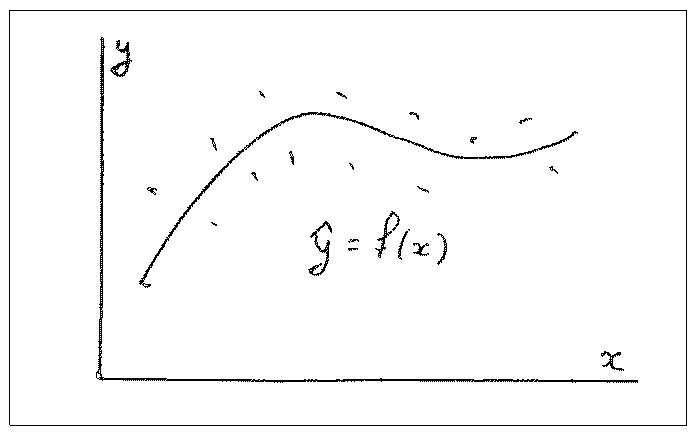
Эта функция нам нужна для того, чтобы в дальнейшем можно было строить прогноз  по значению *x.* При построении функции *f*(*x*) будем считать, что

*y = f*(*x*) + δ,  (1)

где первое слагаемое — закономерное изменение *y* от *x*; δ — случайная составляющая с нулевым средним. Т.е. *f*(*x*) является условным математическим ожиданием *y* при условии известного *x*:

M(*y* ⎢*x*)= *f*(*x*)

и называется *регрессией y по x* (рис. 15). Случайная составляющая выражает или внутреннюю изменчивость, присущую отклику *y*, или влияние факторов, неучтенных в *x,* или и то, и другое вместе.



Различают два случая:

1) значения факторов *x* случайны (например, атмосферные измерения);

2) значения *x* — неслучайны. **Рис. 15. Подбор функции *f* по наблюдениям**

В последнем случае рассматриваются два варианта:

а) значения *x* можно задавать;

б) нет возможности задавать значения *х*, нужно использовать имеющиеся данные. Здесь рассматриваются неслучайные значения факторов, однако, при случайных значениях формулы получаются такие же.

множество допустимых функций, из которого подбирается *f*(*x*), считается параметрическим:

*f*(*x*) = *f*(*x,* θ),

где θ — неизвестный параметр (многомерный).

*Если функция f(x,* θ*) линейна по* θ:

*f*(*x,* θ) = *h*(*x*)⋅θ,

*то регрессионный анализ называется* *линейным*.

**11.1. Простая линейная регрессия**

Пусть *x* и *y —* одномерные величины, а функция *f*(*x,* θ) имеет вид *f*(*x*, θ) = *а + bx*, где θ = (*а, b*). Относительно имеющихся наблюдений (*xi , yi*), *i* = 1, 2, ..., *n*, полагаем, что

*yi = a'+ bxi +*δ*i ,* (2)

где δ1*,* δ2*...*δ*n*— некоррелированные (и ненаблюдаемые) одинаково распределенные случайные величины с нулевыми средними, Mδ*i* = 0, и одинаковыми дисперсиями, Dδ*i* = σ2. Можно различными методами подбирать «лучшую» прямую линию.

Широко используется *метод наименьших квадратов.* Построим оценку параметра θ = (*а, b*) так, чтобы величины

*ei = yi −*= *yi − f* (*xi,* θ) = *yi − а' − bxi ,*

называемые остатками, были как можно меньше, а именно, чтобы сумма их квадратов была минимальной:

= min по (*а', b*). (3)

Чтобы упростить формулы, положим в (2) *xi = xi − ;* получим:

*yi = a + b* (*xi − *) + δi *, i* = 1, 2...*n*, (3)

где  = , *a =a' + b.* Сумму  минимизируем по (*a, b*), приравнивая нулю производные по *a* и *b*. Получим систему линейных уравнений относительно *a* и *b*. Ее решение () легко находится:

, где , (4)

** . (5)

**Свойства оценок**. Нетрудно показать, что если *M*δ*i =* 0*, D*δ*i =* σ2, то:

1) *M= а, М= b*, т.е. оценки несмещенные;

2) *D=* σ2 */ n*, D**= σ2 / ;

3) *cov* () = 0;

4) оценки ** и ** нормально распределены и независимы;

5) остаточная сумма квадратов

*Q*2 =  (6)

независима от (**,**), а *Q*2*/* σ2 распределена по закону хи-квадрат  с (*n –* 2) степенями свободы.

Последние два свойства оценок имеют место, если дополнительно предположить нормальность распределения δ*i*.

**Оценка для σ2  и доверительные интервалы.** Свойство 5, т.е. равенство М*Q*2*/*σ2 = *n* – 2, дает возможность несмещенно оценивать неизвестный параметр σ2 величиной

*s*2 *= Q*2/(*n –* 2). (7)

Поскольку s2 независима от ** и **, отношения

 и , где ,

имеют распределение Стьюдента с (*n* – 2) степенями свободы, и потому доверительные интервалы для *a* и *b* таковы:

 , , (8)

где *tp* — квантиль уровня (1 + *PД*) / 2 распределения Cтьюдента с (*n* – 2) степенями свободы; *PД*  — коэффициент доверия.

**Проверка гипотезы о нулевом коэффициенте наклона.** Обычно возникает вопрос: может быть, *y* не зависит от *х*, т.е. *b* = 0, и изменчивость *y* обусловлена только случайными составляющими δ*i* ? Проверим гипотезу *Н: b* = 0. Если 0 не входит в доверительный интервал (8) для *b*, т.е.

, (9)

то гипотезу *Н* следует отклонить; уровень значимости при этом α *=* 1 *− PД.*

Другой способ (в данном случае эквивалентный (9)) проверки гипотезы *Н* состоит в вычислении статистики

*F* = , (10)

распределенной, если *Н* верна, по закону *F*(1, (*n* − 2)) Фишера с числом степеней свободы 1 и (*n* − 2). Если

*F > F*1*−*α *,*  (11)

где *F*1*−*α *—* квантиль уровня (1 −α*)* распределения *F*(1, *n* – 2), то гипотеза *Н* отклоняется с уровнем значимости α*.*

**11.2. Множественная регрессия. Схема Гаусса-Маркова**

Обобщением линейной регрессионной модели с двумя переменными является модель множественной регрессии.

*А. Анализируемая модель.* Будем предполагать, чтофункция *f*(*x*) в (1) линейна по *x,* ии результатом наблюдения являетсяη— скалярная случайная величина**:**

η = *f*(*x*)+ δ = *b*1*x*1 *+b*2*x*2 + *... +*  *bk xk+*δ, (1)

где *b*1, *b*2 … *bk* — неизвестные коэффициенты регрессии. Пусть *n* раз измерены значения факторов *x*1 *, x*2 *... xk*  и соответствующие значения отклика η*.* Результаты *n* измерений:

η1 *= x*11*b*1 *+ ... +*  *xk*1*bk +* δ1,

…….. (1а)

ηn  *= x*1n*b*1 *+ ... +*  *xknbk +*δn

(первый индекс фактора *х* относится к номеру фактора, а второй — к номеру наблюдения). Предполагается также, что δ1, δ2 … δ*n* — некоррелированные случайные величины с одинаковыми дисперсиями

M(δ*i*δ*j*) *=* 0,  *i ≠ j,* и Mδ*i =* 0, Mδi2= σ2. (1б)

Соотношения (1а) удобно записывать в матричной форме:

η *=* X*Tb +* δ*,* (1в)

где η = (η1*,* η2 *...* η*n*)*T* — вектор-столбец значений зависимой переменной; *Т* — символ транспонирования; *b =* (*b*1, *b*2 *... bk*)*T*— вектор-столбец (размерности *k*) неизвестных коэффициентов регрессии;δ = (δ1, δ2 ... δ*n*)*T* — вектор-столбец случайных отклонений;

**

— матрица *k* × *n* значений факторов; в *i*-м столбце находятся значения факторов при *i*-м наблюдении.

*Б.* *Оценка коэффициентов регрессии***.** Пусть β является оценкой для *b*, тогда *=* X*T*β есть вектор прогнозов для имеющихся значений откликов η. Построим оценку β для вектора *b* так, чтобы вектор прогнозов  = ХTβ минимально (в смысле квадрата нормы разности) отличался от вектора ηзаданных значений (метод наименьших квадратов):

*S*(β) ≡  по β. (2)

Используя правило векторного дифференцирования

,

где *a, b —* векторы-столбцы, зависящие от *x*, получим необходимое условие экстремума:

**= 0. (3)

Если определитель detXXT ≠ 0, то решением является

β = (XXT)-1Xη = Z-1Xη, (4)

где введено обозначение Z = XXT. Убедимся в том, что найденное значение β доставляет минимум функции *S*(β).

Пусть β1 = β + ε.Покажем, что *S*(β1) ≥ *S*(β). Действительно,

*S*(β1) ≡ [(η – XTβ) – XTε]T[(η – XTβ) – XTε] =

= *S*(β) – εTX(η – XTβ) – (η – XTβ)TXTε + εTXXTε = *S*(β) + εTZε ≥ *S*(β); (5)

здесь дважды было учтено равенство (3).

Отметим свойства полученной оценки (4):

1. Оценка линейна по набюдениям η*.*

2. Оценка несмещенно оценивает *b.* Действительно,

Mβ = Z-1XMη = Z-1XXT*b* = *b*. (6)

3. Определим дисперсионную матрицу Dβ оценки. Учитывая (4), (6) и (1в), имеем

(β – *b*) = Z-1Xη**–** Z-1XXT*b* =Z-1X(η – XT*b*)= Z-1Xδ.

Тогда

Dβ = M(β – *b*)(β – *b*)T = M(Z-1XδδTXTZ-1) = σ2Z-1XXTZ-1 = σ2Z-1. (7)

Как видно, точность оценок зависит от значения факторов. Задачей выбора этих значений занимается *теория планирования эксперимента*.

4. Свойство оптимальности оценки — **теорема Гаусса-Маркова.** *В классе линейных несмещенных оценок в условиях* (1б) *оценки* βi, *i* = 1, 2 … *k* (4) *являются оценками с минимальной дисперсией.*

*В. Оценка дисперсии* σ*2 наблюдений.*Рассмотрим значениефункции *S*(*b*) (2), в точке β = *b* с учетом (1в):

*S*(*b*) = (η – XT*b*)T(η – XT*b*) = δTδ = .

Математическое ожидание

M*S*(*b*) = *n*σ2. (8)

ОпределимM*S*(*b*) иначе. Запишем *S*(*b*) в другом виде с учетом (5):

*S*(*b*) = *S*(β+(*b –* β)) = *S*(β) +(*b –* β)TZ(*b –* β). (9)

Обозначим элементы симметричной матрицы Z через z*ij,* а элементы матрицы Z-1 — через z*ij*. Математическое ожидание второго слагаемого с учетом (7) будет иметь вид:

M(*b –* β)TZ(*b* – β) =  = = σ2tr(ZZ-1) = σ2tr(Ek) = = *k*σ2,

обозначено Ek- единичная матрица размерности *k*.

Приравнивая математическое ожидание выражения (9) значению *n*σ2 из (8), имеем

M*S*(β) + *k*σ2 = *n*σ2,

откуда получаем

M*S*(β) / (*n – k*) = σ2.

Это равенство означает, что несмещенной оценкой для σ2 является

= =. (10)

Вектор

*e* ≡=

называется остаточным вектором или вектором невязок.

*Г. Доверительные интервалы для коэффициентов регрессии.* Еслипредположить, что случайные составляющие δ*j*, *j* = 1, 2…*n*, распределены по нормальному закону *N*(0, σ2), тогда оценки β*i*, *i* = 1, 2…*k*, распределены по *N*(*bi*, σ2z*ii*). Известна теорема, обобщающая теорему раздела 6.4., Б, согласно которой отношение  распределено по закону хи-квадрат  с (*n – k*) степенями свободы, и β и *S*(β)независимы. Ясно, что отношение

= ≡*Tn-k*  (11)

подчиняется закону Стьюдента с (*n –* *k*) степенями свободы. Соответственно, доверительный интервал определяется соотношением

, (12)

где *TP* =Q((1+*P*д) / 2, (*n – k*)) — квантиль уровня (1 + *P*д) / 2 распределения Стьюдента с (*n –* *k*) степенями свободы; *P*д — коэффициент доверия.

Основываясь на статистике (11), проверяется гипотеза *bi* = 0 о нулевом значении коэффициента регрессии — если

 *TP*, (13)

то гипотеза о нулевом значении отклоняется.

*Д. Нелинейная функция регрессии f(x)*. Связь между факторами *x* иоткликомη может быть нелинейной, например, в виде полинома:

η *= P*(*x*) *+* δ,

где *P*(*x*) = *b*1 *+ b*2 *x + ...+ bk xk-*1*,* (*k –* 1)— степень полинома; *b*1, *b*2…*bk* — неизвестные коэффициенты; δ — случайная составляющая;Мδ *=* 0*, D*δ *=* σ2 .

Пусть *n* раз измерены значения фактора *x* и соответствующие значения отклика η. Результаты *n* измерений:

η*i* = *b*1 *+ b*2 *xj + ...+ bk xjk-*1*+*δ*j , j =* 1, 2...*n,*  (14)

или, как и (1в), в матричной форме:

η *=* XT*b +* δ *,* (15)

где .

Имеем задачу (1в), и потому формулы (2) — (13) оказываются справедливыми. Слово «линейный» в названии «линейный регрессионный анализ» означает линейность относительно параметров *bi*, но не относительно факторов *xi*.

Кроме полиномиальной широко используются следующие модели:

1) логарифмическая — если зависимость η *= a*1*,* то после логарифмирования получаем

lnη *=* ln *a*1*+ a*2ln *x = b*1 *+ b*2ln *x;*

2) гиперболическая (при обратной зависимости, т.е. при увеличении *х,* признак η уменьшается)

η *= b*1*+ ;*

3) тригонометрическая

*y = b*1*+ b*2 sinω*x + b*3cosω*x*

и другие.

*Е. Нелинейная функция регрессии (обобщение).* Пусть связь между *p* факторами (*х*1*, х*2*...хр*) ≡ *x* и откликом η выражается следующим образом:

η = *f*(*x*,*b*) + δ *= * ,

где ϕ*i*(*x*) (*i* = 1, 2...*k*) — система некоторых функций. Имеется *n* наблюдений при различных значениях *х* ≡ (*х*1*, х*2*...хр*): *x*1 , *x*2... *xn*:

η*j  = *, *j* = 1, 2...*n*,

или в матричной форме:

η *=* XT*b +* δ,

где Х — матрица (*k* + 1) × *n*, в *j*-м столбце которой находятся элементы ϕ1(*xj*), ϕ2 (*xj*) ... ϕ*k*(*xj*); обозначения η*, b*, δ аналогичны(1в). Получили задачу (1в), и потому все формулы (4) — (13) оказываются справедливыми.

Если, например, отклик η нужно приблизить многочленом второй степени от двух переменных *x* и *y,* т.е.

*f*(*x*,*y*,*b*) = *b*0 *+ b*1*x + b*2*y* *+ b*3*x*2 *+ b*4 *xy* *+ b*5*y*2,

то в матрице X в *j-*м столбце будут находиться следующие 6 элементов:

1, *xj*, *yj*, *xj2*, *xjyj*, *yj2*.

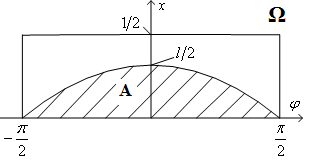
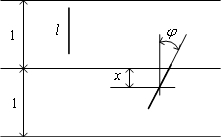
**Глава 5. Метод Монте-Карло**

***Далее нужно проверять***

§12. Метод статистических испытаний Монте-Карло

12.1. Идея метода

Этот метод относится к вычислительной математике. Идея метода стара, она восходит к французскому естествоиспытателю Бюффону (Жорж Луи Леклерк, 1707–1788). Он заметил, что если иголку бросать случайным образом на разлинованную плоскость, то, подсчитывая число пересечений ее с линиями (вернее, подсчитывая долю пересечений), можно вычислять число . Действительно, пусть на плоскости имеется семейство параллельных прямых линий, расстояние между которыми равно 1; иголка длины *l* < 1 бросается на плоскость (рис. 16). Исход этого эксперимента можно характеризовать двумя числами (*x*, ϕ), где  — расстояние от центра иглы до ближайшей прямой,  — угол наклона иголки относительно прямых линий. множеством всех исходов Ω ={(*x*, ϕ): , )} является прямоугольник (рис. 17). Событию «пересечение» благоприятствуют точки области *А* прямоугольника, *А* = {(*x*,ϕ): }.



**Рис.** **16. Игла Бюфона** **Рис**. **17. Вычисление вероятности пересечения**

Нетрудно увидеть, что вероятность *p* = *P*(A) этого события есть отношение площадей S(*A*) и S(Ω):

.

Проделав много  бросаний, среди которых *K* закончились пересечением линий, и оценив вероятность  → *p* при *N* → ∞, можно оценить π: → π.

в этом примере существенным является следующее. Была исходная математическая задача: определить некоторое число — число π. Задача сведена к определению характеристики случайного события — вероятности *P*(A) некоторого события. эта характеристика определена статистически, из опыта (конечно, приближенно).

Идею метода Монте-Карло можно пояснить следующим образом:

1) имеется задача определения некоторого математического объекта ;

2) конструируется случайная величина  (или случайный процесс), у которой математическое ожидание совпадает с искомым : ;

3) проводится много реализаций этой случайной величины ξ1, ξ2…ξ*N*, и  оценивается статистически:

.

По закону больших чисел имеем сходимость к *y* при *N* → ∞. Следует учесть, что запись эта условная:, так же как и , может быть вектором, функцией, или вектор-функцией. Значение  должно быть большим, метод реализовывают на ЭВМ, и потому, как метод вычислений, он возник с появлением ЭВМ в 40-х гг. XX в. Название подчеркивает связь со случайностью.

Как конструируется  по заданной задаче? Общего рецепта нет. Разным математическим задачам соответствуют разные . Совокупность этих соответствий и есть методы Монте-Карло.

**12.2. Общая характеристика методов.** Области применения методов можно условно разделить на две части.

Первая область охватывает задачи вычислительной математики: вычисление интегралов, обращение матриц, решение систем линейных алгебраических уравнений, решение уравнений в частных производных и другие задачи. Основная идея: конструируется какой-либо случайный процесс (или случайная величина), математическое ожидание которой равна интересующей нас величине.

**Пример 1.** Вычисление интеграла

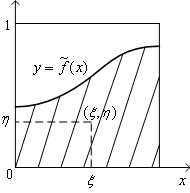
.

Первый путь решения. Пусть  — случайная величина, распределенная равномерно на отрезке . Тогда случайная величина  имеет математическое ожидание

, и .

Имея много реализаций: ξ1, ξ2…ξ*N*, можно оценить значение математического ожидания , и оценкой для  является

.



О точноcти и преимуществах такого вычисления будет сказано ниже.

Второй путь решения. Линейно деформируем отрезок  и функцию  в функцию

= 

так (рис. 18), чтобы:

1) отрезок  перешел в отрезок [0, 1] , **Рис. 18. Интеграл после**

2) . **преобразования**

Пусть (ξ, η) — случайная точка, равномерно распределенная в квадрате со стороной длины 1. Вероятность  попадания этой точки в заштрихованную область равна интегралу

*p* = P{}=.

Производим  бросаний; результаты (ξ1, η1), (ξ2, η2)… (ξ*N*, η*N*). Пусть *K* — число попаданий в заштрихованную область, т.е. когда . Очевидно, величина  является оценкой интеграла. Формально, пусть

 где *i* = 1, 2…*N.*

Тогда Mε*i* = P{} = *p* и

.

Пример 2. Интегралы вида



выделением полного квадрата можно привести к виду:

,

и тогда этот интеграл есть математическое ожидание *y* = , где ξ подчиняется нормальному закону *N*(*m*, σ2). Сгенерируем *N* раз независимые, нормально распределенные случайные величины ξ1, ξ2…ξ*N* и вычислим

.

Пример 3. Общий прием вычисления несобственных интегралов  состоит во введении в интеграл подходящей плотности *p*(*x*):

.

Если  — случайная величина, распределенная с плотностью , то , и тогда .

Многомерные интегралывычисляются аналогично**,** и именно для них сказываются преимущества метода. Объем вычислений с ростом размерности *k* по квадратурным формулам *V*кв растет экспоненциально, а для метода Монте-Карло *V*МК — линейно:

*V*кв= С1*k,*  *V*МК = С2*k*.

В таблице приведены данные из [12] о сравнении объема *V* и времени *Т* вычислений по одной и той же задаче; *k* — размерность интеграла. При малых размерностях выигрывают квадратурные формулы, а при больших — методы Монте-Карло.

**Таблица. Сравнение методов по объему вычисления**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *k* | Метод квадратур | | Метод Монте-Карло | |
| *V*кв | *T*кв | *V*МК | *T*МК, |
| 1 | 102 | 10-4 с | 104 | 10-2 с |
| 3 | 106 | 1 с | 3⋅104 | 3 10-2 с |
| 6 | 1012 | 280 ч | 6⋅104 | 6 10-2 с |

Заметим, что для одной и той же задачи можно по-разному сконструировать вероятностную модель ξ, так что объем вычислений при одной и той же точности будет разным. Например, в задаче Бюффона, если бросать скрепленные в крест две иголки, считая бросание креста за два бросания иголки, то выигрыш в числе бросаний составит 12,2 раза. Если бросать три иголки, скрепленные звездочкой, то выигрыш составит 44,3 раза. Если бросать четыре скрепленных иголки, то выигрыш составит 107,2 раза.

Эти примеры говорят о том, что сконструировать вероятностную модель ξ дело непростое, и оно требует некоторой математической изобретательности. По этому поводу Хемминг в книге «Численные методы» справедливо замечает: «Прежде чем начинать вычисление по Монте-Карло, советую получить консультацию у хорошего статистика, чтобы узнать, чем он может помочь». ***Не понимаю, что исправлять(Ю.Г.)***

Вторая область: моделирование случайных процессов и явлений с целью выяснения их статистических характеристик. Эта область характерна тем, что уже имеется математическое описание случайных объектов. вообще говоря, характеристики этих объектов можно вычислить через интегралы. Но сделать это бывает весьма трудно. В этих случаях придумывать случайные процессы для моделирования не надо, т.к. они заданы самой задачей. Примеры:

1. *Вычисление распределений случайных величин*. Заданы случайные величины ξ1, ξ2…ξ*n* с плотностями *p*1(*x*), *p*2(*x*)…*pn*(*x*) соответственно; пусть η = *f*(ξ1, ξ2…ξ*n*), где функция  — известна. Нужно определить функцию распределение  для η. Например, заданы случайные величины ξ1, ξ2,ξ3, и η = *f*(ξ1, ξ2,ξ3) = 2 ξ1 + ξ22 – . Для решения нужно много, *N* раз, сгенерировать серии (ξ1, ξ2,ξ3), *N* раз вычислить η: η1, η2…ηN, — и оценивать  функцией эмпирического распределения согласно разделу 2.

2. *Моделирование сложных вероятностных процессов* физики, химии, техники, систем управления с целью выяснения их качества, например, стрельба по многим целям из совокупности орудий, находящихся на различном расстоянии и имеющих различную вероятность поражения. Целей может быть различное число, они могут лететь на различных скоростях и высотах. Вычислить такие характеристики, как распределение времени на уничтожение целей (или среднее время уничтожения), распределение числа залпов до уничтожения, практически невозможно. Записать в формульном виде эти характеристики, хотя вероятностный процесс задан, весьма трудно, проделывать физический эксперимент — весьма накладно. Разумный способ решения состоит в проведении статистического эксперимента на цифровой модели.

3. *Моделирование систем массового обслуживания* с целью выяснения их характеристик и влияния различных факторов на эти характеристики.

4. *Моделирование алгоритмов обработки* статистической информации с целью выяснения их качества (например, алгоритма распознавания образов, или алгоритма, по которому управляется ракета, догоняющая цель, или алгоритма, по которому на приемной стороне системы связи производится обработка сигнала, и т.д.).

12.3. Способы получения случайных величин

основой моделирования случайных объектов являются способы получения случайных величин. Будем предполагать, что мы располагаем генератором случайной величины ξ, равномерно распределенной на отрезке [0,1]: ξ ~ R[0,1].

1. *Моделирование случайного события*, имеющего вероятность *p*. Событие , где *с* — константа, является случайным. Надо лишь подобрать *с* так, чтобы вероятность *Р*(*А*) события была равна заданной *p*. Если , то ясно, что *с = р*. Итак,  — случайное событие, имеющее заданную вероятность *p*.
2. *Биномиальная случайная величина* η с параметрами  и  может быть получена как число успехов в серии из *n* независимых испытаний с вероятностью успеха *p*. Генерируем *n* независимых, распределенных по , случайных величин ξ1, ξ2…ξ*n* и вычисляем

, где 

1. *Нормальная случайная величина.* Наиболее простой способ ее получения состоит в использовании центральной предельной теоремы. Случайная величина

,

где  независимы и распределены равномерно на [0,1], подчиняется приближенно нормальному закону с параметрами 0 и 1. выбор значения  зависит от того, насколько хорошо мы должны выдержать нормальными «хвосты» распределения . Для большинства практических применений достаточно брать *n* = 6; получим

.

Это соответствует совпадению функций распределения с точностью 0,002 в любой точке. Если нужна нормальная случайная величина  с математическим ожиданием *a* и дисперсией , мы ее получим из :

.

*4.* *Коррелированные случайные величины.* Требуется получить случайные величины  с коэффициентом корреляции *r*1,2 = *r* и *M*η*i* = 0, *D*η*i* = 1, *i* = 1, 2. Генерируем независимые случайные величины (ε1, ε2), *M*ε*i* = 0, *D*ε*i* = 1, *i* = 1, 2 и полагаем

.

Константы *с*1 и *с*2 подбираются из условий *D*η2 = 1 и *r*1,2 = *r*:



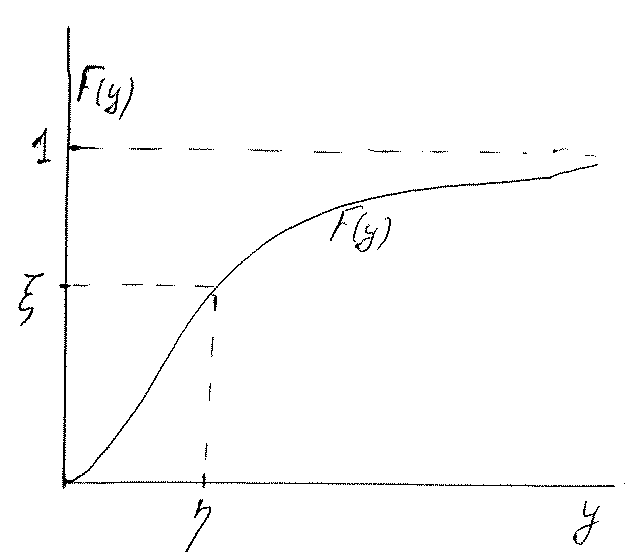
Получаем:

.

Если же требуются случайные величины (ζ1, ζ2) с коэффициентом корреляции *r*1,2 = *r* и *M*η*i* = *mi*, *D*η*i* = δ*i*2, *i* = 1, 2, то вычисляем

ζ*i* = *mi* + δ*i*η*i*, *i* = 1, 2.

*5. Случайная величина с произвольным непрерывным законом распределения.* Пусть нам нужна случайная величина  с функцией распределения . Она может быть получена как корень уравнения (рис. 19):



, где **ξ** ~ R[0, 1]. (1)

Действительно, случайная величина  имеет нужное распределение; ее функция распределения  имеет вид

**Рис. 19. Иллюстрация процедуры**

**генерации при непрерывном распределении**



в силу того, что , .

Пример. Получение случайной величины  с показательным распределением, с плотностью . Ее функция распределения

.

Функция, обратная к :

.

Пусть  равномерно распределена на [0, 1]. Тогда

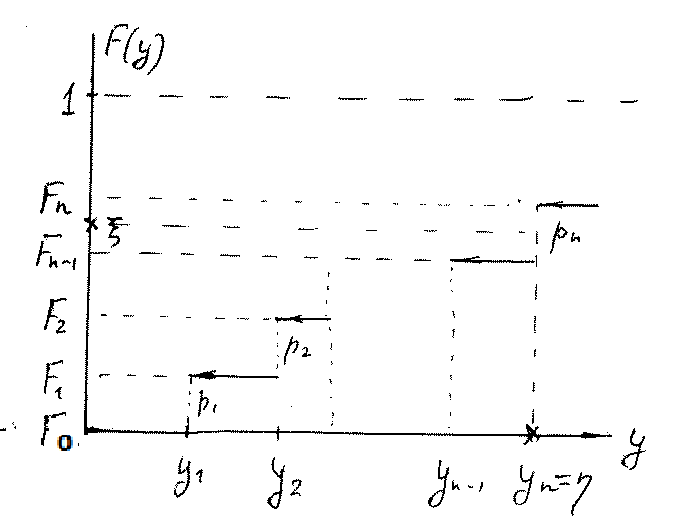


имеет распределение с плотностью . Поскольку () — тоже равномерно распределена на [0, 1], то () имеет тоже плотность .

1. *Дискретные случайные величины*. Пусть нам нужна дискретная случайная величина η, принимающая значения *у*1, *у*2 … *у*n с вероятностями , ...,  (их может быть конечное число). Идея остается той же, что и в предыдущем случае — корень уравнения , где **ξ** ~ R[0, 1], хотя вычислительные процедуры обычно отличаются (Рис. 20); здесь мы имеем дело с табличным заданием ступенчатой функции. Пусть

, *n* = 1, 2,..., F0 = 0.

Разобьем отрезок [0,1] точками . Если выпало такое значение  , что , то полагаем η = *yn*. Ясно, что . процесс вычислений строится следующим образом: значение , начиная с 1, увеличиваем до момента, когда впервые реализуется событие ;



**Рис. 20. Иллюстрация процедуры** это значение  берется для η = *yn*. **генерации при дискретном распределении**

**Пример.** *Случайная величина, распределенная по закону Пуассона.* Пусть  равномерно распределена на  и

 =, *n*≥0.

Первое целое значение  при возрастании , начиная с 0, для которого , и есть искомое значение случайной величины. Итак, первое , *n* ≥ 0, для которого , является пуассоновской случайной величиной с параметром .

12.4. Выбор числа испытаний

При реализации метода Монте-Карло возникает вопрос о выборе числа  испытаний. Пустьξ1, ξ2…ξ*n* — независимые наблюдения, причем ,  *i* = 1, 2…*N*. Рассматривается оценка  для . Требуется выбрать *N* таким образом, чтобы погрешность  была по абсолютному значению не более допустимой величины  с вероятностью не меньшей заданной  (близкой к 1), то есть должно выполняться неравенство:

.

Оценим вероятность в левой части неравенства, как функцию , используя центральную предельную теорему:

,

или

,

что дает необходимое число испытаний

.

Однако, дисперсия *D*ξ обычно неизвестна (ведь даже Мξ неизвестно). Оценим ее сверху: . Получим

 (2)

и будем использовать *N*1 испытаний (с увеличением  вероятность события  лишь увеличивается). В качестве *D*max можно использовать аналитическую оценку или статистическую верхнюю доверительную границу.

Пример. Пусть методом Монте-Карло оценивается вероятность  случайного события . Пусть  — случайная величина, равная 1, если в *i*-м испытании событие  появляется, и 0, если не появляется. Тогда  и . Будем полагать, что  достаточно велико для того, чтобы считать  нормальной случайной величиной. Для  получаем . Поскольку = Dmax; получим

.

Если δ0 = 0,01, то *N≈* 22·103.

В заключение отметим основные особенности метода Монте-Карло:

1. приспособленность метода к многомерным задачам;
2. помехоустойчивость к сбоям машины;
3. погрешность с ростом  уменьшается как , откуда следует, что, например, для уменьшения погрешности в 10 раз число испытаний нужно увеличить в 100 раз.

**ЛИТЕРАТУРА**

***В списке только дваучебника и одно учебное пособие (Ю.Г.)***

1. Крамер Г. Математические методы статистики. М., «Мир», 1975.
2. Гнеденко Б. В. Курс теории вероятностей. М., «Наука», 1965. *(учебник)*
3. Ивченко Г.И., Медведев Ю.И. Математическая статистика. М., Высшая школа, 1992. *(учебное пособие)*
4. **Большев Л. Н., Смирнов Н. В.** Таблицы математической статистики. М., «Наука», 1965.
5. Рао С. P. Линейные статистические методы и их применение. М., «Наука», 1968.
6. Кендалл М., Стьюарт А. Статистические выводы и связи. М., «Наука», 1973.
7. **Гроот М.** Оптимальные статистические решения. М., «Мир», 1974.
8. **Леман Э.** Проверка статистических гипотез. М., «Наука», 1964
9. **Тюрин Ю.Н., Макаров А.А.** Анализ данных на компьютере. М.: Финансы и статистика, 1995..
10. **Ермаков С. М.** Метод Монте-Карло и смежные вопросы. М., «Наука», 1971.
11. **Соболь Н. М.** Метод Монте-Карло. М., «Наука», 1968.
12. **Кузин Л. Т.** Основы кибернетики. М., «Энергия», 1975. *(учебник)*

Учебное издание

Юрий Александрович Горицкий

**Введение в математическую статистику**

Учебное пособие по курсу “Теория вероятностей и математическая статистика” для студентов, обучающихся по направлению

“Прикладная математика”

Редактор издательства

ЛР №……. от …….

Темплан издания МЭИ 2014 г.(II), учебн. Подписано к печати

Формат бумаги 60x84/16 Печать офсетная Физ. печ. л.

Тираж 200 Изд. №

Издательство МЭИ, Россия, 111250, Москва, Красноказарменная, 14