Математическая обработка наблюдений для астрономов

Ольга Сергеевна Сажина

Государственный астрономический институт им. П.К. Штернберга,

Московского государственного университета им. М.В. Ломоносова,

Университетский пр. 13

cosmologia@yandex.ru

2019 г.

Содержание

Введение

Учебно-методическое пособие «Математическая обработка наблюдений для астрономов» основано на курсе лекций (осенний семестр) для студентов астрономического отделения первого курса Физического факультета МГУ им. Ломоносова, читаемого автором с 2015-го года. Пособие может быть использовано студентами астрономических специальностей механико-математических и физико-математических факультетов университетов.

Пособие содержит сведения из математической статистики, необходимые для первичной обработки наблюдательных и экспериментальных данных различной природы, в том числе, приведены способы представления данных для их последующей обработки, методы вычисления основных характеристик данных с указанием погрешностей этих вычислений. Особое внимание уделяется линейной и полиномиальной регрессии для аппроксимации данных непрерывными функциями. Обсуждается вопрос проверки данных на соответствие определенному типу распределения.

Пособие снабжено большим количеством примеров, преимущественно из астрономии, а также дополнено главами из смежных областей: теории вероятностей, комбинаторики, линейной алгебры, которые делают пособие самодостаточным для решения широкого круга прикладных статистических задач без обращения к дополнительной литературе.

Литература, использованная при написании пособия, указана в библиографии. Для более углубленного изучения изложенных в пособии вопросов, рекомендуется следующая дополнительная литература. По теории вероятностей и математической статистике: В.С. Пугачёв "Теория вероятностей и математическая статистика", Москва, Н., 1979 г.; "Введение в теорию вероятностей", Москва, Н., 1968 г.; Б.В. Гнеденко "Курс теории вероятностей", Москва, Н., 1988 г.; Ю.В. Линник "Метод наименьших квадратов и основы теории обработки наблюдений", Москва, Физматгиз, 1962 г.; А.Н. Колмогоров "Основные понятия теории вероятностей", Москва, Н., 1974; Е.С. Вентцель "Теория вероятностей", Москва, Н., 1969 г.; Г. Крамер "Математические методы статистики", Москва, Н., 1975г г.; В. Феллер "Ведение в теорию вероятностей и ее приложений. в 2-х т." Москва, Мир, 1984 г.; М. Лоэв "Теория вероятностей", Москва, Изд-во иностранной литературы, 1962 г. По линейной алгебре: А.И. Кострикин "Введение в алгебру. Часть II. Линейная алгебра", Москва, Н., 2000 г.; И.М. Гельфанд "Лекции по линейной алгебре", Москва, Н., 1971 г.; Ф.Р. Гантмахер "Теория матриц", Москва, Физматлит, 2010 г. По численным методам: Н.С. Бахвалов "Численные методы", Москва, Н., 1973 г.

Сажина О.С.
2019 г.

1 Погрешности и ошибки измерений

В разделе обсуждаются особенности методов и проблем математической обработки данных и наблюдений, в том числе дается понятие погрешности (прямая и обратная задача, точная ошибка приближенного числа, предельная абсолютная погрешность, предельная относительная погрешность, погрешности простейших элементарных функций). Дается описание действий с приближенными числами (сложение, вычитание близких чисел, умножение, деление, оценка ошибки функции приближенных аргументов). Вводится понятие ошибки измерений.

1.1 О принципах, проблемах и особенностях сбора и математической обработки данных

Окружающий мир полон информации всевозможного рода. Качество сбора и обработки информации не только в точности приборов, не только в надежности экспериментальных установок, но и в понимании того, что вся информация хранит в себе элементы случайности. Невозможно провести несколько раз абсолютно одинаковые эксперименты или осуществить абсолютно одинаковые наблюдения и получить абсолютно одинаковый результат. Случайность – это неотъемлемое свойство природы и избавиться от нее невозможно, а потому надо уметь обнаруживать ее и контролировать – как качественно, так и количественно.

При наблюдениях, измерениях, экспериментах различают несколько видов возможных ошибок.

- Систематические (или инструментальные) ошибки, ошибки каталогов. Систематические ошибки являются следствием влияющих на измерение эффектов, действие которых не распознано и не устранено (или не учтено). Например, вследствие рефракции измеряемая высота светила над горизонтом оказывается больше истинной высоты. Если рефракцию не учитывать, то в измерения высоты светила над горизонтом вносится систематическая ошибка. На практике полностью исключить систематические ошибки нельзя.

- Личные ошибки наблюдателя и экспериментатора, в том числе грубые ошибки и опечатки.

- Ошибки, связанные с физическими особенностями исследуемого процесса. Например, согласно квантово-механическому принципу неопределенности, невозможно одновременно измерить точно импульс и координату частицы.

- Случайные ошибки, которые могут быть как свойствами прибора, так и свойствами самого исследуемого процесса. Эти ошибки исследуются статистическими методами, и они будут дальше обсуждаться.

1.2 Погрешности вычислений и действия с приближенными неслучайными числами

Из-за ограниченной точности измерительных приборов результаты измерений всегда приближенные. Кроме того, результаты измерений содержат и случайную составляющую, от которой нельзя избавиться никаким повышением точности. Рассмотрим сначала, как оперировать с результатами, лишенными случайной составляющей. Предположим, что существует точное числовое значение измеряемой величины (как независящая

от прибора объективная реальность). Измерение же дает какое-то другое значение. Таким образом, определяется *конечная ошибка измерения.*

Предположим, проделан ряд измерений и каждое измерение содержит свою ошибку. Далее с этими измерениями исследователь хочет производить, к примеру, простейшие арифметические действия: складывать, вычитать, умножать, делить. Определение ошибок результатов, полученных при обработке приближенных чисел с известными ошибками (с известными интервалами изменения) – *прямая задача* обработки приближенных чисел. Если же точность конечного результата задается и требуется определить, с какой точностью должны быть измерены исходные величины, то имеет место *обратная задача.*

1.2.1 Точная ошибка приближенного числа

Пусть A – точное неизвестное значение измеряемой величины, a – измеренное значение, тогда

$$\Delta_a = a - A$$

есть *точная ошибка приближенного числа.*

1.2.2 Предельная абсолютная погрешность

Наименьшее положительное число ϵ_a, такое, что

$$a - \epsilon_a \leq A \leq a + \epsilon_a$$

называется *предельной абсолютной погрешностью.*

ПРИМЕР Приведем пример вычисления предельной абсолютной погрешности. Расстояние S от Земли до планеты Глизе 581c[1] равно $a = 6.2$ парсека (1 пк $= 3 \cdot 10^{13}$ км). Округлим до целого числа парсеков, $S \approx 6.0$ пк. Предельная абсолютная погрешность (ϵ_a) округленного приближенного значения равна половине единицы последнего знака округления, т.е. $\epsilon_a = 1$ пк $/2 = 0.5$ пк.

1.2.3 Предельная относительная погрешность

Мала или велика предельная абсолютная погрешность в предыдущем примере? Важна не только малость предельной абсолютной погрешности сама по себе, но и ее малость в сравнении с измеренной величиной (так для радиуса планеты Глизе 581c $\epsilon_a = 1$ км это очень хорошо, но та же величина для измерения длины Керченского моста представляет собой очень грубую погрешность). Вводится понятие *предельной относительной погрешности*:

$$\delta_a = \frac{\epsilon_a}{|a|} \tag{1}$$

или

$$a(1 - \delta_a) \leq A \leq a(1 + \delta_a), a > 0$$

$$a(1 + \delta_a) \leq A \leq a(1 - \delta_a), a < 0.$$

ПРИМЕР Приведем пример вычисления предельной относительной погрешности. Для приведенного выше примера про планету Глизе 581c формула (1) дает $\delta_a = 0.5$ пк $/6.2$ пк $= 0.08$ или, как обычно представляется предельная относительная погрешность, 8%.

[1]Планета Глизе 581c – экзопланета в планетной системе звезды Глизе 581, которая была обнаружена в апреле 2007 г. обсерваторией Европейского астрономического сообщества в Чили.

1.2.4 Сложение приближенных чисел

Пусть $a = a_1 + a_2 + a_3 + \cdots + a_n$, где a_i – приближенные числа. Пусть также известны ϵ_i – предельные абсолютные погрешности каждой из a_i. Ставится задача (прямая задача) определить предельную абсолютную погрешность для величины a. Она есть сумма предельных абсолютных погрешностей каждого слагаемого:

$$\epsilon_a = \sum_{i=1}^{n} \epsilon_i.$$

Из этой простой формулы следует важный вывод, что не нужно стремиться получать приближенные слагаемые с разным количеством знаков после запятой.

1.2.5 Вычитание приближенных чисел

Вычитание – это алгебраическое сложение, поэтому для двух приближенных чисел, a_1 и a_2 с заданными предельными абсолютными погрешностями ϵ_1 и ϵ_2 их разность $a = a_1 - a_2$ обладает предельной абсолютной погрешностью $\epsilon_a = \epsilon_1 + \epsilon_2$, а предельная относительная погрешность, соответственно,

$$\delta_a = \frac{\epsilon_a}{|a|} = \frac{\epsilon_1 + \epsilon_2}{|a_1 - a_2|}. \tag{2}$$

Поскольку в знаменателе стоит разность двух величин, то возникает проблема роста предельной относительной погрешности, когда a_1 и a_2 мало отличаются друг от друга. Проблема может быть устранена либо увеличением числа значащих цифр после запятой либо, если первое невозможно, сведением разности $a_1 - a_2$ к разности $\alpha_1 - \alpha_2$, где $a_1 = m + \alpha_1$, $a_2 = m + \alpha_2$.

ПРИМЕР Приведем пример вычисления предельной относительной погрешности для разности двух близких чисел. Пусть необходимо вычислить предельную относительную погрешность левой части формулы, [1]:

$$\left(r_1 + r_2 + s\right)^{\frac{3}{2}} - \left(r_1 + r_2 - s\right)^{\frac{3}{2}} = u. \tag{3}$$

Это формула Эйлера, выражающая связь между двумя радиусами-векторами параболы, длиной хорды и временем. Обычно s есть малая величина по сравнению с $r_1 + r_2$ и поэтому прямое вычисление по формуле (2) приводит к потере точности. От разности близких чисел можно избавиться, умножив и разделив левую часть уравнения (3) на сумму

$$\left(r_1 + r_2 + s\right)^{\frac{3}{2}} + \left(r_1 + r_2 - s\right)^{\frac{3}{2}}.$$

Тогда получается

$$u = \frac{2\left(r_1 + r_2\right)^{\frac{3}{2}}(3\sigma + \sigma^3)}{\left(1 + \sigma\right)^{\frac{3}{2}} + \left(1 - \sigma\right)^{\frac{3}{2}}},$$

где

$$\sigma = \frac{s}{r_1 + r_2} << 1.$$

Можно воспользоваться более общим приемом, разложить преобразованное выражение (3) в ряд по σ.

ПРИМЕР Приведем пример вычисления площадей малых пересекающихся областей на небесной сфере. Ставится задача вычислить площадь некоторой малой области на небесной сфере, заданной четырьмя парами координат своих углов, а также площадь пересечения двух таких областей.

Пусть поверхность S определяется уравнением $z = f(x, y)$ и предполагается гладкой во всех точках, что означает существование в кажой точке вектора, перпендикулярного этой поверхности. Пусть D – область определения функции z на плоскости Oxy (область D есть проекция поверхности S на плоскость Oxy). Площадь поверхности S, ограниченной областью D, вычисляется по формуле:

$$S = \iint\limits_{(D)} \sqrt{1 + \left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2}\, dxdy.$$

Угол γ между перпендикуляром и осью Oz есть

$$\cos\gamma = \pm \frac{1}{\sqrt{1 + \left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2}}.$$

Построим проекцию единичной области $\Delta\sigma_{ij}$ на координатную плоскость Oxy:

$$\Delta\sigma_{ij} = \frac{\Delta x_i \cdot \Delta y_i}{\cos\gamma_{ij}},$$

где γ_{ij} вычисляется в точке c_{ij}. Полная площадь S есть предел суммы:

$$\sum_{i,j} \Delta\sigma_{ij} = \sum_{i,j} \sqrt{1 + \left(\frac{\partial f}{\partial x_i}\right)^2 + \left(\frac{\partial f}{\partial y_j}\right)^2}\, \Delta x_i \Delta y_j.$$

Окончательно,

$$S = \iint\limits_{(D)} \sqrt{1 + \left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2}\, dxdy.$$

Пусть теперь z есть неявная функция переменных x и y: $F(x, y, z) = 0$. В этом случае

$$\frac{\partial F}{\partial x} + \frac{\partial F}{\partial z} \cdot \frac{\partial z}{\partial x} = 0,$$

$$\frac{\partial F}{\partial y} + \frac{\partial F}{\partial z} \cdot \frac{\partial z}{\partial y} = 0.$$

И

$$\frac{\partial z}{\partial x} = -\frac{\frac{\partial F}{\partial x}}{\frac{\partial F}{\partial z}},$$

$$\frac{\partial z}{\partial y} = -\frac{\frac{\partial F}{\partial y}}{\frac{\partial F}{\partial z}}.$$

Окончательно,

$$S = \iint\limits_{(D)} \frac{\sqrt{\left(\frac{\partial F}{\partial x}\right)^2 + \left(\frac{\partial F}{\partial y}\right)^2 + \left(\frac{\partial F}{\partial z}\right)^2}}{\left|\frac{\partial F}{\partial z}\right|} dxdy.$$

Каждое поле определяется на сфере $x^2 + y^2 + z^2 = 1$. Для неявно заданной поверхности

$$\frac{\partial F}{\partial x} = 2x, \frac{\partial F}{\partial y} = 2y, \frac{\partial F}{\partial z} = 2z,$$

и

$$S = \iint\limits_{(D)} \frac{\sqrt{x^2 + y^2 + z^2}}{|z|} dxdy.$$

Вводя полярные координаты на сфере единичного радиуса

$$x = \cos\alpha, y = \sin\alpha,$$

мы переходим от декартовых координат к полярным с якобианом перехода r:

$$S = \int\limits_{r_1}^{r_2} \int\limits_{\alpha_1}^{\alpha_2} \frac{1}{\sqrt{1 - r^2}} r d\alpha dr.$$

Выберем в качестве координатной плоскости Oxy плоскость небесного экватора. В этом случае угол $\alpha \in (0, 2\pi)$ есть прямое восхождение, а $\delta \in (0, \pi/2)$ – склонение. Ось Oz направлена на северный полюс.

Связь между полярной координатой r и углом наклона δ есть

$$r_1 = \cos\delta_2, r_2 = \cos\delta_1$$

$$(\delta_1 < \delta_2).$$

Для каждого поля $(\alpha_1, \delta_1), (\alpha_2, \delta_2), (\alpha_3, \delta_3), (\alpha_4, \delta_4)$ будем приближенно считать $\alpha_1 = \alpha_3, \alpha_2 = \alpha_4$ и $\delta_1 = \delta_2, \delta_3 = \delta_4$. Например:

$$\alpha_1 = 170.56404, \delta_1 = 18.3216;$$

$$\alpha_2 = 169.51245, \delta_2 = 18.32824;$$

$$\alpha_3 = 170.5539, \delta_3 = 17.32329;$$

$$\alpha_4 = 169.50818, \delta_4 = 17.3299.$$

Площадь

$$S = \int\limits_{\cos\delta_2}^{\cos\delta_1} \int\limits_{\alpha_1}^{\alpha_2} \frac{1}{\sqrt{1 - r^2}} r d\alpha dr = (\alpha_2 - \alpha_1) \cdot (\sin\delta_2 - \sin\delta_1) \approx (\alpha_2 - \alpha_1) \cdot (\delta_2 - \delta_1) \cdot \cos\frac{\delta_1 + \delta_2}{2}.$$

Здесь и далее мы используем тригонометрическую формулу для устранения разности синусов близких углов. Поскольку мы хотим рассматривать два пересекающихся поля, то для подсчета полной площади нужно исключить площадь их пересечения:

$$S_{1,2} = S_1 + S_2 - S_U$$

Это легко сделать в терминах проекции: необходимо упорядочить координаты углов двух пересекающихся полей $\alpha_1^1, \alpha_2^1, \alpha_1^2, \delta_2^2$ и $\delta_1^1, \delta_2^1, \delta_1^2, \delta_2^2$. Получаем упорядоченный (вариационный) ряд $\tilde{\alpha_1} < \tilde{\alpha_2} < \tilde{\alpha_3} < \tilde{\alpha_4}$ и $\tilde{\delta_1} < \tilde{\delta_2} < \tilde{\delta_3} < \tilde{\delta_4}$. Далее,

$$S_U = (\tilde{\alpha_3} - \tilde{\alpha_2}) \cdot (\sin\tilde{\delta_3} - \sin\tilde{\delta_2}) \approx (\tilde{\alpha_3} - \tilde{\alpha_2}) \cdot (\tilde{\delta_3} - \tilde{\delta_2}) \cdot \cos\frac{\tilde{\delta_3} + \tilde{\delta_2}}{2}.$$

Рассмотрим два поля с координатами, соответственно:

$$\alpha_1^1 = 170.56404, \delta_1^1 = 18.3216;$$

$$\alpha_2^1 = 169.51245, \delta_2^1 = 18.32824;$$

$$\alpha_3^1 = 170.5539, \delta_3^1 = 17.32329;$$

$$\alpha_4^1 = 169.50818, \delta_4^1 = 17.3299.$$

и

$$\alpha_1^2 = 170.37474, \delta_1^2 = 18.66864;$$

$$\alpha_2^2 = 169.32095, \delta_2^2 = 18.67431;$$

$$\alpha_3^2 = 170.36562, \delta_3^2 = 17.67029;$$

$$\alpha_4^2 = 169.31784, \delta_4^2 = 17.67593.$$

Для определенности выберем

$$\alpha_1^1 = 170.56404, \delta_1^1 = 18.3216, \alpha_2^1 = 169.50818, \delta_2^1 = 17.3299;$$

$$\alpha_2^1 = 170.37474, \delta_2^1 = 18.66864, \alpha_2^2 = 169.31784, \delta_2^2 = 17.67593.$$

Упорядочим величины углов по возрастанию:

$$(\tilde{\alpha_1}, \tilde{\alpha_2}, \tilde{\alpha_3}, \tilde{\alpha_4}) = (169.31784, 169.50818, 170.37474, 170.56404);$$

$$(\tilde{\delta_1}, \tilde{\delta_2}, \tilde{\delta_3}, \tilde{\delta_4}) = (17.3299, 17.67593, 18.3216, 18.66864).$$

Полная площадь этих двух пересекающихся полей есть:

$$S_{12} = S_1 + S_2 - S_U = 0.99683 + 0.99686 - 0.53213 = 1.46156.$$

1.2.6 Умножение приближенных чисел

Пусть $a = a_1 \cdot a_2$ и заданы ϵ_1 ϵ_2 – предельные абсолютные погрешности величин a_1 и a_2 соответственно. Определим предельные абсолютную и относительную погрешности ϵ_a и δ_a.

Произведение точных (неизвестных) величин

$$A = A_1 \cdot A_2 = (a_1 - \Delta_1) \cdot (a_2 - \Delta_2) = a_1a_2 - a_1\Delta_2 - \Delta_1a_2 + \Delta_1\Delta_2 =$$

$$= a - a_1\Delta_2 - a_2\Delta_1 + \Delta_1\Delta_2,$$

где Δ_1 и Δ_2 – точные ошибки приближенных величин a_1 и a_2 соответственно. Точная ошибка произведения есть

$$\Delta_a = a - A = a_1\Delta_2 + a_2\Delta_1 - \Delta_1\Delta_2.$$

Последнее слагаемое есть малая величина второго порядка, поэтому можно ей пренебречь. Окончательно получаем

$$\Delta_a = a_1\Delta_2 + a_2\Delta_1$$

или, для предельной абсолютной погрешности,

$$\epsilon_a = |a_1|\epsilon_2 + |a_2|\epsilon_1.$$

Предельная относительная погрешность (1) произведения двух приближенных величин есть

$$\delta_a = \frac{\epsilon_a}{|a|} = \frac{|a_1|\epsilon_2 + |a_2|\epsilon_1}{|a_1 a_2|} = \frac{\epsilon_2}{|a_2|} + \frac{\epsilon_1}{|a_1|} = \delta_1 + \delta_2.$$

В общем случае для n сомножителей a_k

$$\delta_a = \sum_{k=1}^{n} \delta_k.$$

Таким образом, предельная относительная погрешность произведения равна сумме предельных относительных погрешностей сомножителей.

1.2.7 Деление приближенных чисел

Пусть $a = a_1/a_2$ и так же как и в предыдущем случае имеет место прямая задача, т.е. ϵ_1 и ϵ_2 заданы и требуется определить предельные абсолютную и относительную погрешности ϵ_a и δ_a. Аналогично выводу формул для умножения, для точного значения

$$A = \frac{A_1}{A_2} = \frac{a_1 - \Delta_1}{a_2 - \Delta_2}.$$

Точная ошибка приближенной величины a:

$$\Delta_a = \frac{a_1}{a_2} - \frac{a_1 - \Delta_1}{a_2 - \Delta_2} = \frac{a_2\Delta_1 - a_1\Delta_2}{a_2^2 - a_2\Delta_2}.$$

Введем величину

$$\tilde{\Delta}_a = \frac{a_2\Delta_1 - a_1\Delta_2}{a_2^2}$$

и найдем разность этой введенной величины и точной ошибки приближенной величины a:

$$\tilde{\Delta}_a - \Delta_a = -\frac{(a_2\Delta_1 - a_1\Delta_2)\Delta_2}{a_2^2(a_2 - \Delta_2)}.$$

Поскольку $a_2 \neq 0$, то $(\tilde{\Delta}_a - \Delta_a)$ есть малая второго порядка, а значит с точностью до малой второго порядка $\tilde{\Delta}_a = \Delta_a = (a_2\Delta_1 - a_1\Delta_2)/a_2^2$. Тогда предельная абсолютная погрешность отношения двух приближенных величин есть

$$\epsilon_a = \frac{|a_1|\epsilon_2 + |a_2|\epsilon_1}{a_2^2}.$$

Предельная относительная погрешность (1) отношения двух величин есть

$$\delta_a = \left(\frac{|a_1|\epsilon_2 + |a_2|\epsilon_1}{a_2^2}\right) \cdot \left|\frac{a_2}{a_1}\right| = \frac{\epsilon_1}{|a_1|} + \frac{\epsilon_2}{|a_2|} = \delta_1 + \delta_2,$$

что в точности совпадает с результатом вычисления предельной абсолютной погрешности для произведения приближенных величин.

1.2.8 Оценка ошибки функции приближенных аргументов

Пусть задана непрерывно-дифференцируемая функция f и $U = f(A)$. Если вместо точного значения аргумента A подставить его приближенное значение a, то полученное значение функции $u = f(a)$ также будет приближенным.

Пусть задана предельная абсолютная погрешность ϵ_a. Решим прямую задачу и найдем предельную абсолютную погрешность результата, ϵ_u.

$$U = f(A) = f(a - \Delta_a) = f(a) - \Delta_a f'(\xi),$$

где ξ – любое число, такое что $a - \Delta_a \le \xi \le a$ (использована теорема Лагранжа о конечном приращении). Кроме того, $\Delta_u = u - U$. Тогда

$$\Delta_u = f'(\xi)\Delta_a.$$

Если в последней формуле заменить ξ на a, то ошибка такой замены будет более высокого порядка, чем Δ_a: $\Delta_u = f'(a)\Delta_a$, или, т.к. $\epsilon_a \ge |\Delta_a|$, то предельная абсолютная погрешность искомой функции

$$\epsilon_u = |f'(a)|\epsilon_a,$$

а предельная относительная погрешность искомой функции

$$\delta_u = \left|\frac{f'(a)}{f(a)}\right| |a| \delta_a.$$

Предельная абсолютная погрешность функции нескольких аргументов имеет аналогичную структуру:

$$\epsilon_u = \left|\frac{\partial f}{\partial x}\right| \epsilon_a + \left|\frac{\partial f}{\partial y}\right| \epsilon_b, \tag{4}$$

где частные производные функции $f(x, y)$ берутся в точке $x = a, y = b$.

ПРИМЕР Приведем пример оценки ошибки функции приближенных аргументов. Ускорение силы тяжести определяется с помощью оборотного маятника следующим образом, [1]:

$$g = \frac{4\pi^2 l}{P^2},$$

где l – приведенная длина маятника, P – период колебания.

Наблюдения дали значения величин и их предельные погрешности: $l = 50.02$ см, $P = 1.4196$ сек, $\epsilon_l = 10^{-2}$ см, $\epsilon_P = 10^{-4}$ сек.

Вычислим ускорение силы тяжести g и его предельную погрешность (значение $\pi = 3.1416$, т.е. $\epsilon_\pi = 5 \cdot 10^{-5}$).

Применяя формулу (4), получим выражение для предельной абсолютной погрешности

$$\epsilon_g = \frac{8\pi l}{P^2}\epsilon_\pi + \frac{4\pi^2}{P^2}\epsilon_l + \frac{8\pi^2 l}{P^3}\epsilon_P = 0.37 \text{ см/сек}.$$

После выполнения вычислений получим $g = 979.88$ см/сек^2.

2 Основы теории вероятности и комбинаторики

В разделе рассматриваются элементы теории вероятностей. Даются основные понятия и определения: опыт, событие, вероятность, независимые события. Приводятся основные формулы комбинаторики, сумма вероятностей, умножение вероятностей, геометрическая вероятность, условная вероятность, полная вероятность, формула Байеса, повторы испытаний.

2.1 Опыт, событие и вероятность

Основными объектами, с которыми оперирует теория вероятностей, являются *опыт* (или *испытание*) и *результат опыта* (или *исход, событие*). Так, бросание игрального кубика и выстрелы по мишени – это примеры опыта, а выпадение определенного количества точек на игральном кубике и попадание (или непопадание) в мишень – это примеры соответствующих данным опытам событий.

В процессе какого-нибудь опыта событие появляется с определенной частотой. Частота служит для определения основного понятия теории вероятности – собственно *вероятности события*. Пусть проведено n одинаковых опытов («одинаковость» означает по возможности одинаковые условия всех опытов). Пусть результатами этих опытов служит, к примеру, появление некоторого события A. Тогда частота появления события A есть отношение числа появлений этого события m_A к общему числу испытаний n. Если число испытаний велико, то такая частота и называется вероятностью события[2] A:

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

Граничные значения для p, очевидно, 0 и 1 (0 – вероятность невозможного события, которое обозначается $\varnothing$, а 1 – вероятность достоверного события, которое обозначается Ω). Любому случайному событию можно поставить в соответствие его вероятность, число от 0 до 1.

Математические операции над вероятностями вводятся аналогично операциям над случайными событиями. Другими словами, используется аппарат теории множеств, [2]. Как комбинаторика, так и теория вероятностей представляют собой обширные самостоятельные дисциплины, поэтому в рамках данного пособия введем только основные понятия и обсудим важнейшие приемы и правила, необходимые для решения практических задач.

2.2 Геометрическая вероятность

Часто бывает так, что множество исходов какого-либо опыта бесконечно – например, попадание точки на заданный отрезок. В подобный случаях (для равномерного распределения) вероятность события A есть отношение соответствующих геометрических мер (длин, площадей, объемов):

$$P(A) = \frac{mes(g)}{mes(G)},$$

где mes обозначает меру соответствующей размерности, g – область допустимых исходов и G – область всех возможных исходов.

ПРИМЕР Приведем пример на вычисление геометрической вероятности. В любые моменты промежутка времени T равновозможны поступления в приемник двух

[2] Строгое определение вероятности можно найти в книге [2]

сигналов, [6]. Приемник будет забит, если промежуток времени между моментами поступления сигналов меньше τ. Определить вероятность того, что приемник забит.

Пусть x и y – моменты поступления сигналов в приемник. Отметим область их допустимых значений на декартовой плоскости. В ходе опыта величины x и y могут принимать любые значения от 0 до T, значит, область их допустимых значений представляет собой квадрат со стороной T. Область, соответствующая тому, что приемник забит (т.е., промежуток времени между моментами поступления сигналов окажется меньше τ), определяется неравенством:

$$|x - y| < \tau.$$

Мера области допустимых значений G есть площадь этой области

$$S(G) = T^2,$$

а мера искомой области g есть площадь фигуры на плоскости, лежащей в первой четверти и ограниченной осями координат и прямыми: $y - x = \tau$, $x - y = \tau$, $y = T$, $x = T$:

$$S(g) = T^2 - (T - \tau)^2$$

Вероятность того, что приемник забит:

$$p = \frac{S(g)}{S(G)} = 1 - \left(1 - \frac{\tau}{T}\right)^2.$$

2.3 Условная вероятность

Условная вероятность для двух событий A и B (вероятность того, что событие A появилось при условии появления события B) есть отношение числа опытов, в которых события A и B появились вместе (m_{AB}), к числу опытов, в которых появилось только событие B (m_B):

$$P(A|B) = \frac{m_{AB}}{m_B} = \frac{m_{AB}/n}{m_B/n} = \frac{P(A \cdot B)}{P(B)}.$$

2.3.1 Независимые события

Два события A и B называются *независимыми*, если одновременно A не зависит от B

$$P(A|B) = P(A)$$

и B не зависит от A

$$P(B|A) = P(B).$$

Из этого определения следует важное свойство независимых событий, часто принимаемое за само определение независимости

$$P(A|B) = \frac{P(A \cdot B)}{P(B)} = P(A)$$

или

$$P(A \cdot B) = P(A)P(B).$$

2.3.2 Умножение вероятностей

Используя понятие условной вероятности, вводится операция *умножения вероятностей*:

$$P(A \cdot B) = P(A) \cdot P(B|A) = P(B) \cdot P(A|B).$$

2.3.3 Сложение вероятностей

Если события A_i и A_j *несовместные*, т.е. $A_i \cdot A_j = \varnothing$ для $i \neq j$, то $P(A_i + A_j) = P(A_i) + P(A_j)$. Если события A_i и A_j *совместные*, то $P(A_i + A_j) = P(A_i) + P(A_j) - P(A_i \cdot A_j)$. Другими словами, совместность A_i и A_j означает, что $A_i \cdot A_j \neq \varnothing$. Важно отличать несовместность от независимости: если A_i и A_j – независимы, то $P(A_i \cdot A_j) = P(A_i) \cdot P(A_j)$.

Для независимых и совместных событий A и B легко доказать, что $1 - P(A + B) = P(\bar{A} \cdot \bar{B})$.

ПРИМЕР Приведем пример вычисления вероятности суммы событий. По многолетним наблюдениям известна вероятность того, что в районе обсерватории ночь будет ясной, [6]: в феврале эта вероятность равна 0.18, в марте 0.24 и в апреле 0.36. Наблюдатель будет иметь в своем распоряжении инструмент в ночь с 5-го на 6-е число и с 20-го на 21-е число каждого из этих месяцев. Найти вероятность того, что программа наблюдений будет выполнена, если для ее выполнения требуется:

- одна ясная ночь (p_1);
- две ясные ночи (p_2).

Для решения задача сначала сформулируем происходящие события:

- A_1 – ясная ночь с 5-го на 6-е числа февраля,
- A_2 – ясная ночь с 20-го на 21-е числа февраля,
- B_1 – ясная ночь с 5-го на 6-е числа марта,
- B_2 – ясная ночь с 20-го на 21-е числа марта,
- C_1 – ясная ночь с 5-го на 6-е числа апреля,
- C_2 – ясная ночь с 20-го на 21-е числа апреля.

Поскольку предоставленные астроному ночи для наблюдения отделены друг от друга большим периодом времени (15 дней), то можно рассматривать события (т.е., в данном случае наблюдения) независимыми. Вероятность того, что была одна ясная ночь означает вероятность того, что ясной была *хотя бы одна* ночь:

$$p_1 = P(A_1 + A_2 + B_1 + B_2 + C_1 + C_2).$$

Для вычисления вероятности суммы таких событий используем полезный прием перехода к дополнительному событию. Так, очевидно, для любой вероятности p верно: $p + \bar{p} = 1$. В нашем случае это означает, что

$$\begin{aligned} p_1 &= P(A_1 + A_2 + B_1 + B_2 + C_1 + C_2) = 1 - P(\bar{A}_1)P(\bar{A}_2)P(\bar{B}_1)P(\bar{B}_2)P(\bar{C}_1)P(\bar{C}_2) = \\ &= 1 - (1 - 0.18)(1 - 0.24)(1 - 0.36)(1 - 0.18)(1 - 0.24)(1 - 0.36) \approx 1 - 0.16 = 0.84. \end{aligned}$$

Здесь мы учли, что все события независимы, а вероятность произведения независимых событий равна произведению вероятностей.

Вероятность того, что будут две ясные ночи, вычислим с помощью аналогичного приема: $p_2 = 1 - P(\text{ни одна ночь не ясная}) - P(\text{ровно одна ночь ясная})$. Другими словами,

$$\begin{aligned} p_2 &\approx 1 - 0.16 - 2 \cdot 0.18 \cdot (1 - 0.18) \cdot (1 - 0.24)^2 \cdot (1 - 0.36)^2 - 2 \cdot (1 - 0.18)^2 \cdot 0.24 \cdot (1 - \\ &- 0.24) \cdot (1 - 0.36)^2 - 2 \cdot (1 - 0.18)^2 \cdot (1 - 0.24)^2 \cdot 0.36 \cdot (1 - 0.36) \approx 0.49. \end{aligned}$$

2.4 Полная вероятность

Следствием правил сложения и умножения вероятностей является правило *полной вероятности*. Это понятие важно для проверки статистических гипотез в курсе математической статистики. Остановимся на этом более подробно.

Пусть нужно найти вероятность события A, $P(A)$, причем известно, что событие A зависит от условий опыта. Об этих условиях перед началом решения задачи нужно сформулировать n взаимоисключающих предположений (или гипотез): $H_1, H_2, H_3, \cdots, H_n$. Интересно, что гипотезы могут быть сформулированы разными способами, важно помнить, что они должны быть взаимоисключающими (т.е. несовместными):

$$H_i \cdot H_j = \varnothing,$$

что означает

$$P(H_i \cdot H_j) = 0,$$

и в совокупности исчерпывать все возможные ситуации:

$$H_1 \cup H_2 \cup H_3 \cup \cdots \cup H_n = \Omega,$$

что, очевидно, означает

$$P(H_1 \cup H_2 \cup H_3 \cup \cdots \cup H_n) = 1.$$

Каждая гипотеза H_i – это случайное событие, вероятность которого *до* проведения опыта (т.е. *априорная вероятность*) оценивается как $P(H_i)$. По выбору H_i: $\sum_{i=1}^{n} P(H_i) = 1$.

Пусть также известны условные вероятности появления события A при каждой гипотезе H_i: $P(A|H_1), P(A|H_2), P(A|H_3), \cdots, P(A|H_n)$.

Тогда

$$P(A) = \sum_{i=1}^{n} P(H_i \cdot A),$$

потому что событие A может появится только с одной из гипотез

$$A = H_1 \cdot A + H_2 \cdot A + \cdots + H_n \cdot A.$$

Кроме того,

$$P(H_i \cdot A) = P(H_i) P(A|H_i).$$

Окончательно формула *полной вероятности*:

$$P(A) = \sum_{i=1}^{n} P(H_i) \cdot P(A|H_i). \tag{5}$$

ПРИМЕР Приведем пример на вычисление полной вероятности. Среди наблюдаемых спиральных галактик 23% принадлежат подтипу Sa, 31% – подтипу Sb и 46% – подтипу Sc, [6]. Вероятность вспышки в течение года сверхновой звезды в галактике Sa составляет 0.0020, в галактике Sb – 0.0035, в галактике Sc – 0.0055. Найти вероятность (P) вспышки сверхновой в далекой спиральной галактике, подтип которой определить не удается.

По условию, вероятности принадлежности галактики к определенному подтипу: $P(S_a) = 0.23, P(S_b) = 0.31, P(S_c) = 0.46$. Тогда, по формуле (5),

$$P = \sum_{i=a,b,c} P(S_i)P(H|S_i),$$

где H – событие вспышки сверхновой и $P(H|S_i)$ вероятность вспышки сверхновой, при условии, что она произошла в галактиках S_a, S_b и S_c соответственно. Подставляя численные величины из условия, получаем:

$$P = 0.23 \cdot 0.0020 + 0.31 \cdot 0.0035 + 0.46 \cdot 0.0055 = 0.0041.$$

2.5 Формула Байеса

Как следствие формулы полной вероятности и формулы умножения вероятностей, рассмотрим *формулу Байеса*.

Формула Байеса позволяет пересчитывать априорные вероятности $P(H_i)$ с учетом результата опыта. Другими словами, если событие A уже произошло, то можно определить наиболее значимый фактор, повлиявший на это событие. Таким образом, можно определить $P(H_k|A)$ – т.н. *апостериорную вероятность*.

Пусть, как и при выводе формулы полной вероятности, все факторы, влияющие на событие A каким-то образом были сформулированы в виде гипотез $H_1, H_2, \cdots, H_n$, таких что обязательно $H_i \cdot H_j = \varnothing$ и $\sum_i^n H_i = \Omega$. И пусть известны априорные, т.е., оцененные до опыта вероятности $P(H_1), P(H_2), \cdots, P(H_n)$.

Пусть, наконец, событие A произошло. Тогда можно заново пересчитать вероятности $P(H_1), P(H_2), \cdots, P(H_n)$ с учетом того, что событие A произошло: найдем $P(H_1|A), P(H_2|A), \cdots, P(H_n|A)$.

Известно, что $P(H_kA) = P(H_k)P(A|H_k) = P(A)P(H_k|A)$. Тогда *формула Байеса*:

$$P(H_k|A) = \frac{P(H_k)P(A|H_k)}{P(A)},$$

где

$$P(A) = \sum_{i=1}^{n} P(H_i)P(A|H_i).$$

ПРИМЕР Приведем пример на вычисления вероятностей, используя формулу Байеса. В продолжение задачи из примера на вычисление полной вероятности, [6]. Пусть определилось, что в течение часа наблюдений далекой спиральной галактики в ней была обнаружена вспышка сверхновой. Найти вероятности того, что галактики принадлежит подтипам S_a, S_b и S_c соответственно.

Вероятности того, что галактики принадлежит подтипам S_a, S_b и S_c:

$$P(S_a|H) = \frac{0.23 \cdot 0.0020}{0.0041} = 0.11,$$

$$P(S_b|H) = \frac{0.31 \cdot 0.0035}{0.0041} = 0.27,$$
$$P(S_b|H) = \frac{0.46 \cdot 0.0055}{0.0041} = 0.62.$$

2.6 Элементы комбинаторики

Напомним основные комбинаторные формулы, (Таблица 1), позволяющие вычислять количество способов выбора из группы элементов определенную подгруппу элементов, соблюдая определенные ограничения.

Типы этих ограничений рассмотрим на примерах.

Так, **размещения (или упорядоченная выборка) без повторений** возникают в задачах на составление чисел, причем каждая цифра может быть использована только один раз. Например, из цифр $\{1, 2, 3, 4, 5\}$ можно составить $A_5^4 = 120$ четырехзначных числа (если запретить повторение цифр).

В случае **размещения с повторениями**, когда повторение цифр разрешено, получаем $\bar{A}_5^4 = 5^4$. Здесь же приведем пример из статистической физики, [3]. Пусть механическая система состоит из n частиц и рассматривается в фазовом пространстве, разбитом на m ячеек. Сколькими равновозможными состояниями характеризуется данная система? Ответ зависит от того, различимы частицы или нет. Так, для статистики Максвелла-Больцмана, в которой частицы различимы, каждая из частиц может попасть в любую из m ячеек независмо от остальных частиц. Тогда число всех возможных состояний такой системы есть $\bar{A}_m^n = m^n$.

Задачу на **перестановки без повторений** можно рассматривать как частный случай задачи на размещения без повторения, когда количество размещаемых элементов равно количеству позиций, на которые их размещают (при $n = k$, $A_n^k = n!/(n-k)! = n! = P_n$). Например, количество способов расставить 4 разные книги на полке есть $P_4 = 4!$

Типичный пример на **перестановки с повторениями** – формирование разных слов из букв какого-либо заданного слова. Например, сколько различных семибуквенных слов можно составить из букв, образующих слово "авиация"(под словами подразумеваются любые, даже лишенные смысла, наборы букв)? Сначала пересчитаем количество типов букв и количество букв в каждом типе: $n_1 = 2$ (буква А), $n_2 = 2$ (буква И), $n_3 = n_4 = n_5 = 1$ (буквы В,Ц и Я). Всего букв $n = 7$. Тогда, учитывая, что перестановки одинаковых букв не дают новых слов, получаем $P(2, 2, 1, 1, 1) = 7!/(2!2!1!1!1!) = 1260$.

Сочетания (или неупорядоченная выборка) без повторений являются наиболее часто используемой в статистике комбинаторной формулой (схема испытаний Бернулли). Типичной задачей на применение этой формулы является выбор шаров из урны. Например, из урны, содержащей $n = 10$ шаров, можно наугад выбрать $k = 4$ шара количеством способов: $C_n^k = C_{10}^4 = 10!/4!/6! = 210$. Снова используя пример из статистической физики, рассмотрим статистику Ферми-Дирака (справедлива для электронов, нейтронов, протонов), при которой n частицы неразличимы, их число меньше числа ячеек ($n < m$) и каждая ячейка может содержать не более одной частицы. Тогда такая механическая система характеризуется числом равновозможных состояний C_m^n.

В заключение рассмотрим несколько примеров задач на **сочетания с повторениями**. Пусть необходимо составить набор $n = 10$ деталей, используя $m = 4$ типа деталей. Решим задачу сведением к предыдущей комбинаторной формуле: к перестановкам с повторением. Рассмотрим один типичный вариант возможного набора, записав его с помощью нулей и единиц, где единицы будут обозначать количество деталей определенного типа, а нули будут обозначать переход от одного набора к другому. Так,

Таблица 1:

Основные формулы комбинаторики

	Размещения (упорядоченная выборка)	Перестановки	Сочетания (неупорядоченная выборка)
без повторения	$A_n^k = \frac{n!}{(n-k)!}$	$P_n = n!$	$C_n^k = \frac{n!}{k!(n-k)!}$
с повторением	$\bar{A}_n^k = n^k$	$P(n_1, n_2, \cdots, n_k) = \frac{n!}{n_1! n_2! \cdots n_k!}$	$\bar{C}_n^m = \frac{(n+m-1)!}{(m-1)! n!}$

$\{1\,1\,0\,1\,1\,1\,1\,0\,1\,1\,0\,1\,1\}$ означает, что было набрано $k_1 = 2$ деталей первого типа, $k_2 = 4$ деталей второго типа, $k_3 = 2$ деталей третьего типа и $k_4 = 2$ деталей четвертого типа. Нули разделяют группы деталей и количество нулей есть $m - 1 = 3$. Различные варианты таких наборов – это перестановки с повторениями из 10-ти единиц и 3-х нулей: $\bar{C}_{10}^4 = P(10, 3) = 13!/3!/10! = 286$. Возвращаясь к примеру из статистической физики, рассмотрим статистику Бозе-Эйнштейна (справедлива для фотонов, атомных ядер, атомов с четным числом элементарных частиц), при которой все n частиц неразличимы и все их распределения по m ячейкам равновозможны. Такая механическая система характеризуется числом состояний C_{m+n-1}^n. Или, еще один пример, уравнение $x_1 + x_2 + \cdots + x_m = n$ при натуральном n имеет C_{m+n-1}^n неотрицательных целочисленных решений[3].

ПРИМЕР Приведем пример использования комбинаторных формул при вычислении вероятностей.

В изданном в 1784 г. каталоге Мессье, содержащем наблюдаемые на небе 108 ярких туманных объектов, имеется 39 галактик, 29 рассеянных скоплений, 29 шаровых скоплений, 6 диффузных туманностей и 5 планетарных туманностей, [6]. Определить вероятность того, что из двух объектов, наугад выбранных в каталоге,

- каждый окажется галактикой;
- один окажется шаровым, а другой – рассеянным скоплением.

В первом случае вероятность определяется отношением числа способов выбрать два объекта из имеющихся 39 галактик (т.н. *благоприятное событие*) к числу способов выбрать два объекта из полного каталога:

$$p_a = \frac{C_{39}^2}{C_{108}^2} \approx 0.128.$$

Во втором случае благоприятное событие есть выбор одного объекта из 29-ти шаровых скоплений и одновременного выбора одного объекта из 29-ти рассеянных скоплений.

[3] Обратите внимание, что в примерах использовались разные обозначения для комбинаторных параметров, важен их смысл при постановке конкретной задачи.

Общее возможное количество вариантов выбрать два объекта из всего каталога определяется так же, как в предыдущем пункте:

$$p_b = \frac{C_{29}^1 C_{29}^1}{C_{108}^2} \approx 0.146.$$

2.7 Повторение опытов (схема испытания Бернулли) и производящая функция

Пусть осуществляется последовательность n независимых опытов, в каждом из которых происходит или не происходит событие A. И пусть вероятность события A, $P(A)$, известна. Задача заключается в том, чтобы определить вероятность появления события A ровно m раз. Обозначим искомую вероятность $P_{m,n}$, а $P(A) = p$. Обозначим также вероятность того, что событие A не появилось через q. Очевидно, $q = 1-p$. Тогда искомая вероятность $P_{m,n} = C_n^m p^m q^{n-m}$. Если проводятся независимые опыты в изменяющихся условиях, т.е. $P(A_i) = p_i$ и $q_i = 1 - p_i$, то для вычисления вероятности $P_{m,n}$ построим функцию (т.н. *производящую функцию*)

$$\phi_n(x) = \prod_{i=1}^{n}(q_i + p_i x) = \sum_{m=0}^{n} P_{m,n} x^m$$

Вероятности $P_{m,n}$ есть коэффициенты при x^m в разложении производящей функции.

3 Распределение случайной величины

3.1 Основные понятия математической статистики

3.1.1 Случайная величина

Случайная величина – это величина, которая в результате *опыта* может принимать то или иное значение, заранее неизвестное, но принадлежащее множеству возможных значений. Любая *функция случайной величины* также есть случайная величина. Случайные величины могут быть как *непрерывного*, так и *дискретного* типов.

Обратим внимание на обозначения: прописными латинскими буквами $X, Y, Z, \cdots$ будем обозначать сами случайные величины, а строчными латинскими буквами $x_1, x_2, \cdots, x_n, y_1, y_2, \cdots, y_n, z_1, z_2, \cdots, z_n$ будем обозначать возможные значения, которые могут принимать эти случайные величины.

ПРИМЕР Приведем пример случайной величины дискретного типа. Число очков, выпавших при однократном бросании кубика. Множество возможных значений есть $\{x_1, x_2, x_3, x_4, x_5, x_6\} = \{1, 2, 3, 4, 5, 6\}$.

ПРИМЕР Приведем пример случайной величины непрерывного типа. Ошибка измерения скорости кометы Δv. Множество возможных значений есть $[\Delta v_{min}, \Delta v_{max}]$.

3.1.2 Генеральная совокупность

Генеральной совокупностью называется набор всех возможных значений случайной величины (т.н. *полный набор*). Важно отметить, что в практических задачах полный набор значений, которые может принимать случайная величина, никогда не известен.

3.1.3 Выборка

Выборка – это конечное число значений случайной величины, подмножество генеральной совокупности. Выборка – это то, что анализируется в любой задаче математической статистики.

3.1.4 Распределение случайной величины

Закон распределения случайной величины X – это функция $p(x)$, которая устанавливает соответствие между возможными значениями случайной величины и вероятностями этих значений. Так, каждому возможному значению случайной величины $\{x_1, x_2, \cdots, x_n\}$ ставится в соответствие своя вероятность $\{p(x_1), p(x_2), \cdots, p(x_n)\}$, причем $\sum_{i=1}^{n} p(x_i) = 1$ (поскольку случайная величина обязана принять одно из своих возможных значений и их набором исчерпываются все возможности для ее значения).

Существует несколько способов задания закона распределения случайной величины.

- *Ряд распределения* $\{x_i, p(x_i)\}$;
- *Функция распределения* $F(x)$ или интегральный закон распределения;
- *Плотность распределения* $f(x)$ или дифференциальный закон распределения.

Каждый из этих способов однозначно и полностью задает закон распределения случайной величины. Важно обратить внимание, что и функция распределения $F(x)$, и плотность распределения $f(x)$, и ряд распределения $\{x_i, p(x_i)\}$ – функции не случайного аргумента (т.е. сами не являются случайными). Они есть функции значений, которые может принимать случайный аргумент.

Таблица 2:

Представление закона распределения случайной величины в виде таблицы – статистического ряда распределения

x_i	x_1	x_2	x_3	$\cdots$	x_n
$p(x_i)$	$p(x_1)$	$p(x_2)$	$p(x_3)$	$\cdots$	$p(x_n)$

3.1.5 Ряд распределения случайной величины или статистический ряд

Простейшей формой задания закона распределения дискретной случайной величины X является *таблица*, которая в данном случае и называется *статистическим рядом распределения*, Таблица 2. Каждому значению x_i ставится в соответствие вероятность $p(x_i)$.

ПРИМЕР Рассмотрим пример построения ряда распределения случайной величины. Пусть производится два независимых опыта,в каждом из которых событие A появляется с вероятностью $p = 0.60$. Построить закон распределения случайной величины X – числа появлений события A.

Исходя из условия задачи, случайная величина X может принимать значения $\{x_0, x_1, x_2\} = \{0, 1, 2\}$. Найдем соответствующие вероятности $\{p(x_0), p(x_1), p(x_2)\}$: $p(x_0)$ есть вероятность того, что ни в первом, ни во втором случае событие A не появилось, $p(x_1)$ есть вероятность того, что событие A появилось ровно один раз (либо в первом опыте, либо во втором), $p(x_2)$ есть вероятность того, что событие A появилось в обоих опытах. Этим набором должны исчерпываться все возможные значения случайной величины X, поэтому контрольной проверкой вычисления вероятностей является проверка условия $p(x_0) + p(x_1) + p(x_2) = 1$. Итак,

$$p(x_0) = (1-p)\cdot(1-p) = 0.16,$$

$$p(x_1) = (1-p)\cdot p + p\cdot(1-p) = 0.48,$$

$$p(x_2) = p\cdot p = 0.36.$$

Вычислив вероятности, построим ряд распределения, (Таблица 3).

3.1.6 Функция распределения

Функцией распределения случайной величины X называется функция, определенная на всей действительной оси,

$$F(x) = P(X < x),$$

где X – случайная величина, а x – неслучайное фиксированное возможное значение случайной величины X. Таким образом, функция распределения представляет собой неслучайную функцию на множестве возможных значений случайной величины.

Из определения функции распределения следует, что вероятность попадания случайной величины на отрезок есть

$$P(X \in [\alpha, \beta)) = F(\beta) - F(\alpha).$$

Таблица 3:

Статистический ряд распределения случайной величины X, принимающей значения $\{x_0, x_1, x_2\} = \{0, 1, 2\}$ с вероятностями $\{p(x_0), p(x_1), p(x_2)\} = \{0.16, 0.48, 0.36\}$

x_i	0	1	2
$p(x_i)$	0.16	0.48	0.36

Не любая функция может быть функцией распределения. Функция распределения должна удовлетворять следующим условиям:

- вероятность невозможного события равна нулю: $F(-\infty) = P(X < -\infty) = P(\varnothing) = 0$;
- вероятность достоверного события равна единице: $F(+\infty) = P(X < +\infty) = P(\Omega) = 1$;
- функция распределения – неубывающая, т.е. для $x_2 > x_1$ $F(x_2) \geq F(x_1)$.

Дискретный аналог функции распределения (т.е. понятие функции распределения для дискретных величин) – это *кумулята*.

ПРИМЕР Построим функцию распределения для предыдущего примера:

$$F(0) = P(X < 0) = 0,$$
$$F(1) = P(X < 1) = P(X = 0) = 0.16,$$
$$F(2) = P(X < 2) = P(X = 0) + P(X = 1) = 0.16 + 0.48 = 0.64,$$
$$F(2 + \epsilon) = P(X < 2 + \epsilon) = P(X = 0) + P(X = 1) + P(X = 2) = 0.16 + 0.48 + 0.36 = 1.$$

Функция F определена на всей действительной оси.

3.1.7 Плотность вероятности

Плотность вероятности или *плотность распределения* вводится и имеет смысл только для непрерывной случайной величины:

$$f(x) = F'(x).$$

Плотность вероятности, как и функция распределения, не произвольная функция, а должна удовлетворять определенным условиям:

- $f(x) \geq 0$,
- $\int_{-\infty}^{+\infty} f(x)dx = 1$.

Из определения плотности вероятности следует, что функция распределения $F(x)$ геометрически есть площадь под графиком функции распределения $f(x)$:

$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

Дискретный аналог плотности вероятности – *гистограмма*.

3.1.8 Двумерная плотность вероятности

Аналогичным образом определяется двумерная плотность вероятности $f(x,y)$ двух случайных величин X и Y. Это такая функция, для которой

- $f(x,y) \geq 0$,
- $\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)dxdy = 1$.

Двумерная функция распределения

$$F(x,y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(t,\tau)dtd\tau.$$

Для непрерывных случайных величин при заданной совместной плотности можно определить соответствующие одномерные (маргинальные) плотности

$$f(x) = \int f(x,y)dy, \quad f(y) = \int f(x,y)dx.$$

Две случайные величины X и Y независимы тогда и только тогда, когда для любых значений x и y

$$f(x,y) = f(x)\cdot f(y).$$

ПРИМЕР Рассмотрим следующую кусочно-заданную функцию

$$f(x,y) = \begin{cases} x+y, \text{ если } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0, \text{ во всех остальных случаях} \end{cases}$$

и покажем, что она может служить плотностью вероясности совместного распределения двух случайных величин X и Y. Действительно,

$$\int_0^1\int_0^1 \big(x+y\big)dxdy = \int_0^1\Big(\int_0^1 xdx\Big)dy + \int_0^1\Big(\int_0^1 ydx\Big)dy = \int_0^1 \frac{1}{2}dy + \int_0^1 ydy = \frac{1}{2}+\frac{1}{2} = 1.$$

ПРИМЕР Рассмотрим еще один пример, [4]. Пусть плотность распределения имеет вид:

$$f(x,y) = \begin{cases} cx^2y, \text{ если } x^2 \leq y \leq 1 \\ 0, \text{ во всех остальных случаях} \end{cases}$$

Определим значение параметра c из условия того, что данная функция должна быть плотностью распределения. При вычислении интегралов обратим внимание, что для каждого фиксированного значения x нужно брать величину y, меняющуюся на отрезке $[x^2, 1]$. Таким образом,

$$1 = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)dxdy = c\int_{-1}^{1}\int_{x^2}^{1} x^2ydxdy = c\int_{-1}^{1} x^2\Big(\int_{x^2}^{1} ydy\Big)dx =$$

$$= c\int_{-1}^{1} x^2\frac{1-x^4}{2}dx = \frac{4c}{21}.$$

Тогда

$$c = \frac{21}{4}.$$

Теперь, для примера, вычислим вероятность того, что случайная величина X не меньше случайной величины Y:

$$P(X \geq Y) = \frac{21}{4} \int_0^1 \int_{x^2}^x x^2 y dx dy = \frac{21}{4} \int_0^1 x^2 \left(\int_{x^2}^x y dy \right) dx = \frac{21}{4} \int_0^1 x^2 \left(\frac{x^2 - x^4}{2} \right) dx = \frac{3}{20}.$$

ПРИМЕР Рассмотрим пример вычисления одномерной плотности по известной двумерной плотности. Пусть

$$f(x, y) = e^{-(x+y)}, \quad x, y \geq 0.$$

Тогда

$$f(x) = e^{-x} \cdot \int_0^\infty e^{-y} dy = e^{-x}.$$

ПРИМЕР Пусть X и Y – независимые случайные величины с одинаковыми плотностями распределения, [4]:

$$f(x) = f(y) = f(z) = \begin{cases} 2z, \text{ если } 0 \leq z \leq 1 \\ 0, \text{ во всех остальных случаях} \end{cases}$$

Вычислим вероятность $P(X + Y \leq 1)$, используя условия независимости:

$$f(x, y) = f(x) \cdot f(y) = \begin{cases} 4xy, \text{ если } 0 \leq x \leq 1, \ 0 \leq y \leq 1 \\ 0, \text{ во всех остальных случаях} \end{cases}$$

Тогда

$$P(X + Y \leq 1) = \iint\limits_{x+y \leq 1} f(x, y) dx dy = 4 \int_0^1 x \left(\int_0^{1-x} y dy \right) dx = 4 \int_0^1 x \frac{(1-x)^2}{2} dx = \frac{1}{6}.$$

ПРИМЕР Пусть двумерная плотность задана в виде, [4]:

$$f(x, y) = \begin{cases} 2e^{-(x+3y)}, \text{ если } x > 0 \text{ и } y > 0 \\ 0, \text{ во всех остальных случаях} \end{cases}$$

Поскольку область изменения X и Y представляет собой прямоугольник $(0, \infty) \times (0, \infty)$ и совместная плотность вероятности может быть записана как произведение двух функций

$$f(x, y) = \left(2e^{-x}\right) \cdot \left(e^{-3y}\right),$$

то случайные величины X и Y независимы.

3.2 Представления статистических данных

3.2.1 Простой статистический ряд

Простой статистический ряд удобно представить в виде таблицы (Таблица 4) как соответствие номера наблюдения i и результата наблюдения x_i.

Таблица 4:

Представление простого статистического ряда

номер наблюдения i	1	2	$\cdots$	n
результат наблюдения x_i	x_1	x_2	$\cdots$	x_n

3.2.2 Вариационный ряд

Если в простом статистическом ряде упорядочить все элементы x_i (например, по возрастанию):

$$x_1^* \leq x_2^* \leq x_3^* \leq \cdots \leq x_n^*,$$

то полученный ряд будет называться *вариационным рядом*.

Величина x_k^* называется *порядковая статистика*. Величина $J_n(x) = x_n^* - x_1^*$ называется *размах выборки*.

3.2.3 Эмпирическая функция распределения

По вариационному ряду можно построить *эмпирическую функцию распределения*:

$$F^*(x) = P^*(X < x) = \frac{n_x}{n},$$

где n_x – число значений величины X, которые меньше фиксированного числа x, а n – объем выборки (т.е. общее количество элементов выборки).

Величина X может принимать и одинаковые значения. Тогда пусть k – число разных значений величины X (очевидно, $k \leq n$). Пусть индекс $\nu = \{1, 2, \cdots, k\}$. Тогда в каждой точке x_ν эмпирическая функция распределения $F^*(x)$ будет претерпевать скачок, равный частоте:

$$p_\nu^* = \frac{m_\nu}{n},$$

где m_ν – число одинаковых значений величины X. Очевидно, $\sum_{\nu=1}^{k} p_\nu^* = 1$. Проиллюстрируем вышесказанное примером.

ПРИМЕР Приведем пример построения эмпирической функции распределения. Пусть для $i = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$ соответствующие $x_i = \{1, -2, 0, 1, -3, -1, 0, -1, 1, 3, 0, -1, 1, 2, 0, -1, 0, -2, 0, 1\}$. Из условия видно, что случайная величина принимает всего семь различных значений с определенной частотой (Таблица 5).

3.2.4 Полигон частот

Полигон частот – это сгруппированные данные выборки. Если объем выборки $\{x_1, x_2, x_3, \cdots, x_n\}$ большой ($n > 50$) и число одинаковых значений случайной величины велико ($m_\nu > 20$), то для упрощения дальнейшей обработки данных используют сгруппированные выборочные данные, строя полигон частот. Опишем алгоритм построения полигона частота, из которого станет ясно его определение.

Таблица 5:

Таблица для построения эмпирической функции распределения

x_ν	-3	-2	-1	0	1	2	3
m_ν	1	2	4	6	5	1	1

Таблица 6:

Количество попаданий значений случайной величины в построенные интервалы

Интервал J_j	$[\hat{x}_0, \hat{x}_1)$	$\cdots$	$[\hat{x}_{k-1}, \hat{x}_k]$
число попаданий	n_1	$\cdots$	n_k

АЛГОРИТМ ПОСТРОЕНИЯ ПОЛИГОНА ЧАСТОТ

1. Построить вариационный ряд данных (т.е. упорядочить выборку) и найти $x_{min} = x_1^*$ и $x_{max} = x_n^*$;

2. Весь размах $[x_1^*, x_n^*]$ разбить на k равных интервалов группирования. Число интервалов можно выбрать $k \approx \log_2 n + 1$. В практических задачах $7 \le k \le 10$. Иногда удобно взять интервалы разной длины, в зависимости от количества попадающих в них точек;

3. Отметить в порядке возрастания крайние точки интервалов: $\hat{x}_0, \hat{x}_1, \cdots, \hat{x}_{k-1}, \hat{x}_k$, а также середины интервалов $\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_k$;

4. Подсчитать количества выборочных данных $n_1, n_2, \cdots, n_k$, попавших в каждый интервал (Таблица 6).

5. Заменить величины n_j на *частоты* $p_j^* = n_j/n$ и получить статистический ряд.

   Совокупность точек $\{\tilde{x}_j, p_j^*\}$ и есть *полигон частот*.

3.2.5 Гистограмма

Гистограмма – это дискретный аналог[4] функции плотности вероятности, называемая также эмпирической плотностью вероятности $f^*(x)$.

[4]или выборочный аналог, т.к. основан на конечной дискретной выборке элементов – результатов наблюдений. В большинстве реальных задач обработки наблюдательных и экспериментальных данных работа всегда ведется именно с гистограммами.

Пусть p_j^* есть площадь прямоугольника с длиной основания $\Delta\hat{x}_j = \hat{x}_j - \hat{x}_{j-1}, j = 1, \cdots, k$. Тогда высота этого прямоугольника и определяется как эмпирическая плотность вероятности в точке $\tilde{x}_j$

$$f^*(\hat{x}_j) = \frac{p_j^*}{\Delta\hat{x}_j} = \frac{n_j}{n \cdot \Delta\hat{x}_j}.$$

Здесь, как и при построении полигона, $\hat{x}_0, \hat{x}_1, \cdots, \hat{x}_k$ – крайние точки интервалов, на которые разбивается вариационный ряд обрабатываемых данных. Совокупность таких прямоугольников для всех $\hat{x}_0, \hat{x}_1, \cdots, \hat{x}_k$ и составляет гистограмму, которая при большом количестве точек переходит в свой непрерывный аналог, в функцию плотности вероятности. Сумма площадей всех прямоугольников равна единице.

3.2.6 Кумулята

Кумулята – приближенная эмпирическая функция распределения. Точно так же, как гистограмма является дискретным аналогом функции плотности вероятности, так и кумулята является дискретным аналогом функции распределения.

$$F^*(\hat{x}_\nu) = P(X < \hat{x}_\nu) = \sum_{j=1}^{\nu} p_j^* = \sum_{j=1}^{\nu} \frac{n_j}{n}.$$

Здесь, как и выше, $\hat{x}_0, \hat{x}_1, \cdots, \hat{x}_\nu, \cdots, \hat{x}_k$ – крайние точки интервалов, на которые разбивается вариационный ряд обрабатываемых данных, $k \approx \log_2 n + 1$ – обычно рекомендуемое число интервалов разбиения, n – число элементов выборки, n_j – число данных измерений, повших в интервал $[\hat{x}_{j-1}, \hat{x}_j]$. Точка $\hat{x}_\nu$ – одна из точек-концов интервала: при вычислении кумуляты все предыдущие вероятности, очевидно, складываются, как и для непрерывной функции распределения.

3.2.7 Количество интервалов разбиения при группировке данных

В практических задачах рекомендуемое число интервалов k разбиения при группировке массива n данных есть

$$k \approx \log_2 n + 1.$$

Для оценки величины k можно использовать также метод *скользящего контроля* [4], заключающегося в минимизации оценки скользящего контроля J:

$$\min_{\Delta\hat{x}_j} \Big\{ J(\Delta\hat{x}_j) \Big\},$$

$$J(\Delta\hat{x}_j) = \sum_{\Delta\hat{x}_j} \Big(f^*(\hat{x}_j) \Big)^2 - \frac{2}{n} \sum_{i=1}^{n} f_{(-i)}^*(\hat{x}_j),$$

где $f_{(-i)}^*(\hat{x}_j)$ – гистограмма, построенная после удаления i-го наблюдения из массива данных. Минимизируемая функция, очевидно, требует пересчета гистограммы n раз. Для упрощения расчетов оценка скользящего контроля может быть представлена в виде, [4]:

$$J(\Delta\hat{x}_j) = \frac{2}{(n-1)\Delta\hat{x}_j} - \frac{n+1}{n-1} \sum_{j=1}^{k} \big(p_j^*\big)^2.$$

На практике удобно построить значения $\hat{J}(\Delta\hat{x}_j)$ для каждого $k \in [0, n]$ и определить минимум. Если $\hat{J}(\Delta\hat{x}_j)$ меняется незначительно для $k \in [k_1, k_2]$, то любое значение k из этого интервала можно принимать для расчета интервала разбиения.

3.2.8 Ядерная оценка плотности

В отличие от классической гистограммы, метод ядерной оценки плотности [4] представляет собой сглаженную оценку плотности распределения.

Ядро – гладкая функция $K(x)$, такая что

- $K \geq 0$
- $\int xK(x)dx = 0$
- $\sigma_K^2 \equiv \int x^2 K(x)dx > 0$

Ядерная оценка плотности для $h > 0$ есть

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} \cdot K\left(\frac{x - x_i}{h}\right).$$

Аналогично тому, как при построении гистограммы возникал вопрос об оптимальном выборе шага разбиения, в задаче ядерной оценки плотности возникает задача выбора оптимальной величины h. Так, при решении практических задач обработки данных следует выбирать такое h, которое минимизирует функцию

$$J(h) \approx \frac{1}{h \cdot n^2} \sum_i \sum_j K^*\left(\frac{x_i - x_j}{h}\right) + \frac{2}{n \cdot h} \cdot K(0),$$

где

$$K^*(x) = \int K(z - y)K(y)dy - 2K(x).$$

Приведем два примера используемых ядер.

Ядро Епанечникова

$$K(x) = \begin{cases} \frac{3}{4\sqrt{5}} \cdot \left(1 - \frac{x^2}{5}\right), \text{ если } |x| < \sqrt{5} \\ 0, \text{ во всех остальных случаях} \end{cases}$$

Гауссово ядро

$$K(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2}.$$

В этом случае величина

$$\int K(z - y)K(y)dy \sim N(0, 2).$$

4 Характеристики случайных величин и математические операции над случайными величинами

4.1 Математическое ожидание

Математическое ожидание – это характеристика среднего значения случайной величины (или, в общем случае, случайной функции).

В литературе математическое ожидание случайной величины X обозначается обычно $M[X]$, m_x или $E[X]$, а математическое ожидание функции случайной величины $E[g(X)]$ или $M[g(X)]$.

По определению, для дискретной случайной величины X, принимающей значения $\{x_1, x_2, \cdots, x_n\}$ с соответствующими вероятностями $\{p(x_1), p(x_2), \cdots, p(x_n)\}$, математическое ожидание есть

$$M[X] = \sum_{i=1}^{n} x_i p(x_i).$$

Для непрерывной случайной величины, обладающей заданной функцией плотности распределения $f(x)$, математическое ожидание есть

$$M[X] = \int_{-\infty}^{+\infty} x \cdot f(x) dx.$$

Таким образом, зная плотность распределения случайной величины, можно вычислить математическое ожидание этой случайной величины (а также и все остальные характеристики случайной величины, как будет показано ниже).

ПРИМЕР Не каждое распределение обладает конечным математическим ожиданием. Например, рассмотрим распределение Коши, плотность распределения которого задается функцией

$$f(x) = \frac{1}{\pi} \frac{1}{1+x^2}.$$

Вычислим математическое ожидание этого распределения, воспользовавшись правилом интегрирования по частям:

$$M[X] = \frac{2}{\pi} \int_0^{+\infty} \frac{x dx}{1+x^2} dx = \left[x \cdot \mathrm{tg}^{-1} x\right]\Big|_0^{\infty} - \int_0^{+\infty} \mathrm{tg}^{-1} x dx = \infty.$$

Таким образом, у распределения Коши не существует математического ожидания. Если много раз моделировать это распределение, то его среднее не будет стремиться принять какое-то определенное значение. Плотность распределения Коши обладает широкими крыльями, не спадающими к нулю на бесконечности, что означает возможность получить в наблюдениях экстремальные значения с достаточно большой вероятностью.

4.1.1 Свойства математического ожидания

Пусть X, Y – произвольные случайные величины, а C – неслучайная постоянная величина. Тогда математическое ожидание удовлетворяет следующим свойствам.

- $M[C] = C$
- $M[C \cdot X] = C \cdot M[X]$
- $M[X \pm Y] = M[X] \pm M[Y]$

- Пусть X, Y – непрерывные случайные вличины и $Y = g(X)$. Тогда

$$M[Y] = \int_{-\infty}^{+\infty} g(x) \cdot f(x) dx$$

Если случайные величины X и Y – независимые, то

- $M[X \cdot Y] = M[X] \cdot M[Y]$.

4.1.2 Условное математическое ожидание

Пусть X и Y – случайные величины. Условное математическое ожидание величины X при данном значении $Y = y$ определяется для дискретного случая:

$$M[X|Y = y] = \sum_{i=1}^{n} x_i \cdot p(X_i = x_i|Y = y) = \sum_{i=1}^{n} x_i \cdot \frac{p(X_i = x_i, Y = y)}{P(Y = y)}$$

и для непрерывного случая:

$$M[X|Y = y] = \int_{-\infty}^{+\infty} x \cdot \frac{f(x, y)}{f(y)} dx.$$

Математическое ожидание $M[X]$ случайной величины X – это неслучайное число, а условное математическое ожидание $M[X|Y = y]$ – это функция (в данном случае, переменной y).

Для функции случайных аргументов $g(X, Y)$

$$M[g(X, Y)|Y = y] = \sum_{i=1}^{n} g(x_i, y_i) \cdot p(X_i = x_i|Y = y) = \sum_{i=1}^{n} g(x_i, y_i) \cdot \frac{p(X_i = x_i, Y = y)}{P(Y = y)}$$

$$M[g(X, Y)|Y = y] = \int_{-\infty}^{+\infty} g(x, y) \cdot \frac{f(x, y)}{f(y)} dx.$$

соответственно для дискретного и непрерывного случаев.

Верно соотношение:

$$M[M[g(X, Y)|X]] = M[g(X, Y)].$$

4.2 Среднеквадратическое отклонение

Среднеквадратическое отклонение – это характеристика рассеяния относительно математического ожидания. Другими словами, среднеквадратическое отклонение характеризует, насколько сильно элементы выборки отклоняются от своего среднего значения.

В литературе среднеквадратическое отклонение случайной величины X обозначается обычно *s.d.*, σ_x, $\sigma[X]$.

По определению, среднеквадратическое отклонение дискретной случайной величины X, принимающей значения $\{x_1, x_2, \cdots, x_n\}$ с соответствующими вероятностями $\{p(x_1), p(x_2), \cdots, p(x_n)\}$, есть

$$\sigma[X] = \sqrt{M[(x - m_x)^2]} = \sqrt{\sum_{i=1}^{n} (x_i - m_x)^2 \cdot p(x_i)},$$

где

$$m_x = M[X] = \sum_{i=1}^{n} x_i p(x_i).$$

Для непрерывной случайной величины, обладающей заданной функцией плотности распределения $f(x)$, среднеквадратическое отклонение есть

$$\sigma[X] = \sqrt{\int_{-\infty}^{+\infty} (x - m_x)^2 \cdot f(x)dx},$$

где

$$m_x = M[X] = \int_{-\infty}^{+\infty} x \cdot f(x)dx.$$

4.3 Дисперсия

Дисперсией случайной величины называется квадрат ее среднеквадратического отклонения. В литературе дисперсия случайной величины X обозначается обычно $D[X]$ или σ_x^2.

По определению, дисперсия дискретной случайной величины X, принимающей значения $\{x_1, x_2, \cdots, x_n\}$ с соответствующими вероятностями $\{p(x_1), p(x_2), \cdots, p(x_n)\}$, есть

$$D[X] = M[(x - m_x)^2] = \sum_{i=1}^{n} (x_i - m_x)^2 \cdot p(x_i),$$

где

$$m_x = M[X] = \sum_{i=1}^{n} x_i p(x_i).$$

Для непрерывной случайной величины, обладающей заданной функцией плотности распределения $f(x)$, дисперсия есть

$$D[X] = \int_{-\infty}^{+\infty} (x - m_x)^2 \cdot f(x)dx,$$

где

$$m_x = M[X] = \int_{-\infty}^{+\infty} x \cdot f(x)dx.$$

4.3.1 Свойства дисперсии

Пусть X, Y – произвольные случайные величины, а C – неслучайная постоянная величина. Тогда дисперсия удовлетворяет следующим свойствам.

- $D[C] = 0$
- $D[C \cdot X] = C^2 \cdot D[X]$
- $D[X] = M[X^2] - (M[X])^2$
- $D[X \pm Y] = D[X] + D[Y] \pm 2M[(X - m_x) \cdot (Y - m_y)]$.

Последнее слагаемое называется *ковариацией* и равно нулю, если X и Y – независимые случайные величины.

4.3.2 Условная дисперсия

Величина, [4]

$$D[Y|X=x] = \int_{-\infty}^{+\infty} \big(y-\mu(x)\big)^2 \cdot \frac{f(x,y)}{f(x)} dy,$$

где

$$\mu(x) = M[Y|X=x]$$

называется условной дисперсией.

4.4 Меры положения и меры рассеяния

Пусть некоторая величина X наблюдается или измеряется некоторым прибором n раз. При статистической обработке выборки $\{x_1, x_2, \cdots, x_n\}$, если не оговорено особо, все значения x_i считаются равноправными, т.е. равновероятными. Таким образом, для оценки *среднего* значения (обозначается $\bar{x}$) искомой величины X применяется формула для математического ожидания с учетом того, что $p(x_i) = p = 1/n$:

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

Если каждое x_i обладает весом w_i, то определяется *взвешенное среднее*:

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}.$$

Среднее значение можно определить и по вариационному ряду (в данном случае выборке, приведенной в упорядоченное по возрастанию состояние) – оно, очевидно, должно лежать посередине. Более точно, в зависимости от четности или нечетности общего количества элементов выборки n:

$$m = x_{\frac{n+1}{2}},$$

если n – нечетное и

$$m = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right),$$

если n – четное. Число m называется *медианой*. Наконец, среднее может быть также оценено по наиболее часто встречающемуся элементу выборки, т.е. элементу x_l, при котором функция плотности максимальна. Такое x_l носит название *мода*.

Таким образом, *среднее, взвешенное среднее, медиана* и *мода* характеризуют примерное положение истинного значения искомой величины X и поэтому носят общее название *меры положения*.

Важно не только оценить среднее значение элементов выборки, но и указать, насколько сильно остальные элементы отклоняются от среднего значения, т.е., насколько велико рассеяние элементов. *Мерами рассеяния* или *рассеивания* служат вычисленные по выборке следующие характеристики: *среднеквадратическое отклонение* s, *среднее отклонение* d и *размах* r:

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

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

$$r = J_n(x) = x_n^* - x_1^*.$$

Здесь $x_n^* = x_{max}$, $x_1^* = x_{min}$ и все x_i предполагаются с весом 1.

Особо отметим важнейшую меру рассеяния – среднеквадратическое отклонение выборочного среднего (различают обозначения $s_{\bar{x}}$ или $s(\bar{x})$, если среднее $\bar{x}$ тоже оценивается по выборке, и $\sigma_{\bar{x}}$ или $\sigma(\bar{x})$, если среднее μ известно априори, т.е. это есть среднее генеральной совокупности):

$$s_{\bar{x}} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}.$$

$$\sigma_{\bar{x}} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(x_i - \mu\right)^2}.$$

Эта формула будет выведена ниже.

4.5 Коэффициент корреляции

Помимо определяемых по выборке мер положения и мер рассеяния, для двух случайных величин X и Y определим коэффициент корреляции q

$$q = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

где

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

есть выборочные средние случайных величин $X = \{x_1, x_2, \cdots, x_n\}$ и $Y = \{y_1, y_2, \cdots, y_n\}$.

Если $q = 0$, то X и Y не коррелированные (но могут быть зависимые). Если $q = \pm 1$, то между X и Y существует зависимость в виде прямой пропорциональности.

4.6 Моменты случайных величин

И математическое ожидание, и дисперсия случайной величины X представляют собой частный случаи моментов случайной величины. В общем случае различают *начальный момент k-того порядка*:

$$\alpha_k[X] = M[X^k]$$

и *центральный момент k-того порядка*:

$$\mu_k[X] = M[(X - \alpha_1[X])^k].$$

Таблица 7:

Статистический ряд случайной величины X

x_i	0	$\frac{\pi}{4}$	$\frac{\pi}{2}$
p_i	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$

Очевидно,

$$\alpha_1[X] = M[X] = \mu,$$
$$\mu_2[X] = D[X] = \sigma_x^2.$$

Моменты вычисляются по определению математического ожидания. Начальные моменты:

$$\alpha_k[X]) = \begin{cases} \sum_{i=1}^{n} x_i^k \cdot p(x_i) & \text{для дискретного распределения} \\ \int_{-\infty}^{\infty} x^k f(x)dx & \text{для непрерывного распределения} \end{cases}$$

Центральные моменты:

$$\mu_k[X]) = \begin{cases} \sum_{i=1}^{n} (x_i - \mu)^k \cdot p(x_i) & \text{для дискретного распределения} \\ \int_{-\infty}^{\infty} (x - \mu)^k f(x)dx & \text{для непрерывного распределения} \end{cases}$$

С помощью центрального момента и среднеквадратического отклонения вводится понятие *скошенности* (или *асимметрии*):

$$\gamma_1 = \frac{\mu_3[X]}{\sigma_x^3}$$

а с помощью центрального момента четвертого порядка и среднеквадратического отклонения – понятие *крутизны* (или *эксцесса*):

$$\gamma_2 = \frac{\mu_4[X]}{\sigma_x^4} - 3.$$

4.7 Распределение вероятности для функции случайных величин

4.7.1 Дискретная случайная величина

Пусть X – дискретная случайная величина и пусть $h(X)$ – функция этой случайной величины. Построим ряд распределения для функции случайной величины. Удобнее рассмотреть эту задачу на примере, [5].

Пусть $h(X) = \cos X$, а случайная величина X задана рядом распределения в виде таблицы (Таблица 7).

Тогда закон распределения $h(X)$ будет определяться Таблицей 8, в которой

$$h_i = h(x_i),$$

Таблица 8:

Статистический ряд функции случайной величины $h(X) = \cos X$

h_j	0	$\frac{\sqrt{2}}{2}$	1
p_j^H	$\frac{3}{6}$	$\frac{2}{6}$	$\frac{1}{6}$

Таблица 9:

Статистический ряд случайной величины Y

x_i	$-\frac{\pi}{2}$	$-\frac{\pi}{4}$	0	$\frac{\pi}{4}$	$\frac{\pi}{2}$
p_i	$\frac{1}{35}$	$\frac{3}{35}$	$\frac{6}{35}$	$\frac{10}{35}$	$\frac{15}{35}$

$$p_i^H = P(h(x_i) = h_i) = P(X = x_i)$$

Учитывая, что $x_1 = 0, x_2 = \pi/4, x_3 = \pi/2$ распишем последнее выражение поэлементно:

$$p_1^H = P(h(x_1) = h_1) = P(h(0) = \cos 0 = 1) = P(X = 0) = \frac{1}{6},$$

$$p_2^H = P(h(x_2) = h_2) = P\left(h\left(\frac{\pi}{4}\right) = \cos\frac{\pi}{4} = \frac{\sqrt{2}}{2}\right) = P(X = \frac{\pi}{4}) = \frac{2}{6},$$

$$p_3^H = P(h(x_3) = h_3) = P\left(h\left(\frac{\pi}{2}\right) = \cos\frac{\pi}{2} = 0\right) = P(X = \frac{\pi}{2}) = \frac{3}{6},$$

что и внесем в таблицу, расположив h_i в порядке возрастания. Отметим, что сказанное верно если существует только одно значение $X = x_k$, при котором $h(X) = h(x_k) = h_0$. Если существует несколько значений $X = x_\nu, x_{\nu+1}, \cdots, x_{k-1}, x_k$, при которых $h(X) = h_0$ (т.е. функция, обратная h, не однозначная), то

$$P(h(X) = h_0) = \sum_{j=\nu}^{k} P(X = x_j).$$

Проиллюстрируем сказанное примером, расширив выборку для дискретной случайной величины X из предыдущего примера (см. Таблицы 9-10) и переобозначив X на Y.

Таблица 10:

Статистический ряд функции случайной величины $h(Y) = \cos Y$

h_j	0	$\frac{\sqrt{2}}{2}$	1
p_j^H	$\frac{16}{35}$	$\frac{13}{35}$	$\frac{6}{35}$

4.7.2 Непрерывная случайная величина

Пусть теперь есть непрерывная случайная величина X, для которой известна плотность распределения $f(x)$, которая в дифференциальной форме записывается как, [5]:

$$f(x)dx = P(x \le X \le x + dx).$$

Ставится задача найти плотность распределения $g(h)$, такую, что:

$$g(h)dh = P(h \le H \le h + dh),$$

где

$$h = h(x).$$

Пусть $h(x)$ – однозначная функция. Тогда, по аналогии с дискретным случаем, можно найти малый интервал значений $h(x)$, соответствующий заданному малому интервалу значений X с известной вероятностью $f(x)dx$.

$$dx = \left|\frac{dx(h)}{dh}\right| dh,$$

где $x(h)$ – обратная функция, а $|\cdots|$ – модуль величины. Тогда

$$f(x)dx = f[x(h)]\left|\frac{dx(h)}{dh}\right| dh$$

$$g(h) = f[x(h)]\left|\frac{dx(h)}{dh}\right|.$$

ПРИМЕР Пусть $h(x) = \cos x$. Распределение вероятности для X

$$f(x)dx = a + bx,$$

где

$$0 \le x \le \pi/2.$$

Найдем плотность вероятности $g(h)$:

$$g(h)dh = f[x(h)]\left|\frac{dx(h)}{dh}\right| dh = \left[a + b\arccos h\right] \cdot \frac{dh}{\sqrt{1-h^2}},$$

$$0 \le h \le 1.$$

Окончательно

$$g(h) = \left[a + b \arccos h\right] \cdot \frac{1}{\sqrt{1-h^2}}, 0 \le h \le 1.$$

ПРИМЕР Приведем пример вычисления функции распределения. Вероятность обнаружить звезду в объеме dv равна $k \cdot dv$. Для каждой звезды найдется другая звезда – ее ближайший сосед, [6]. Найти функцию распределения расстояний до ближайшего соседа, а также среднее расстояние до ближайшего соседа и дисперсию расстояний.

Обозначим за X случайную величину, расстояние от звезды до ее ближайшего соседа. Тогда вероятность того, что сосед находится ближе расстояния x равно, по определению, функции распределения, $F(x) = P(X < x)$. Вероятность того, что ближайший сосед находится не ближе x равно, очевидно, $1 - F(x)$. Вероятность того, что ближайший сосед находится на расстоянии, заключенном между x и $x + dx$, есть $f(x)dx$, и равна произведению $1 - F(x)$ на вероятность того, что между сферами с радиусами x и $x \cdot dx$ имеется звезда. Таким образом,

$$f(x)dx = \left[1 - F(x)\right] \cdot k \cdot 4\pi \cdot x^2 dx.$$

Разделим обе части последнего уравнения на $k \cdot 4\pi \cdot x^2 dx$, потом продифференцируем по x и, учтя, что $F'(x) = f(x)$, получим

$$\frac{f'(x)}{f(x)} = \frac{2}{x} - 4\pi \cdot k \cdot x^2,$$

и после интегрирования

$$f(x) = cx^2 \cdot \exp\left\{-\frac{4}{3}\pi \cdot k \cdot x^3\right\}.$$

Произвольная постоянная c определяется из условия равенства интеграла плотности распределения единице на всей числовой прямой.

Окончательно находим

$$f(x) = 4\pi \cdot k \cdot x^2 \cdot \exp\left\{-\frac{4}{3}\pi \cdot k \cdot x^3\right\}.$$

Среднее расстояние до ближайшего соседа

$$\bar{x} = \int_0^\infty x \cdot f(x)dx = \left(\frac{3}{4\pi \cdot k}\right)^{1/3} \Gamma\left(\frac{4}{3}\right) \approx 0.554 \cdot k^{-1/3},$$

где

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} \cdot e^{-t} dt$$

есть гамма-функция (или эйлеров интеграл второго рода), значения которой известны.

Дисперсия расстояния до ближайшего соседа

$$\sigma^2 = \int_0^\infty (x - \bar{x})^2 \cdot f(x)dx = \left(\frac{3}{4\pi \cdot k}\right)^{2/3} \cdot \left[\Gamma\left(\frac{5}{3}\right) - \Gamma^2\left(\frac{4}{3}\right)\right] \approx 0.0405 \cdot k^{-2/3},$$

а среднеквадратическое отклонение

$$\sigma \approx 0.201 \cdot k^{-1/3}.$$

4.8 Неравенства для вероятностей случайных величин и их характеристик

В математической статистике неравенства, связывающие характеристики случайных величин, используются в тех случаях, когда расчет тех или иных характеристик сложен.

Неравенство Маркова

Пусть X – неотрицательная случайная величина. Пусть существует $M[X]$. Тогда для любого $t > 0$

$$P(X > t) \leq \frac{M[X]}{t}.$$

Действительно,

$$M[X] = \int_{-\infty}^{\infty} xf(x)dx = \int_{0}^{\infty} xf(x)dx = \int_{0}^{t} xf(x)dx + \int_{t}^{\infty} xf(x)dx \geq \int_{t}^{\infty} xf(x)dx \geq$$

$$\geq t \cdot \int_{t}^{\infty} f(x)dx = t \cdot P(X > t).$$

Неравенство Чебышева

Пусть X – случайная величина и $M[X] = \mu$, $D[X] = \sigma^2$. Тогда

$$P\big(|X - \mu| \geq t\big) \leq \frac{\sigma^2}{t^2}.$$

Действительно, используя неравенство Маркова, получаем

$$P\big(|X - \mu| \geq t\big) = P\big(|X - \mu|^2 \geq t^2\big) \leq \frac{M\big[(X - \mu)^2\big]}{t^2} = \frac{\sigma^2}{t^2}.$$

В частности, для $Z = (X - \mu)/\sigma$ и положив $t = k \cdot \sigma$, можно показать, что

$$P\big(|Z| \geq k\big) \leq \frac{1}{k^2}.$$

Неравенство Хефдинга

Пусть $X_1, X_2, \ldots, X_n$ – независимые случайные величины, такие что $M[X_i] = 0$ и $a_i \leq X_i \leq b_i$. Тогда для любых $\epsilon > 0$ и $t > 0$

$$P\left(\sum_{i=1}^{n} X_i \geq \epsilon\right) \leq e^{-t\epsilon} \cdot \prod_{i=1}^{n} \exp\left\{t^2 \cdot \frac{(b_i - a_i)^2}{8}\right\}.$$

Если $X_1, X_2, \ldots, X_n$ – независимые случайные величины, имеющие распределение Бернулли с параметром p, то

$$P\Big(\big|\bar{x} - p\big| > \epsilon\Big) \leq 2 \cdot e^{-2n\epsilon^2},$$

где $\bar{x}$ – среднее выборочное значение,

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

Неравенство Милла, [4]

Пусть $X \sim N(0,1)$. Тогда

$$P\big(|X| > t\big) \leq \sqrt{\frac{2}{\pi}} \frac{\exp\{-t^2/2\}}{t}.$$

Неравенство Коши-Шварца

Пусть две случайные величины X и Y имеют конечные дисперсии, тогда

$$M[XY] \leq \sqrt{M[X^2] \cdot M[Y^2]}.$$

5 Основные законы распределения случайной величины

5.1 Распределение точечной массы

Случайная величина X обладает распределением точечной массы в т. a ($X \sim \delta_a$), если $P(X = a) = 1$, т.е. функция распределения есть

$$F(x) = \begin{cases} 0, \text{ если } x < a \\ 1, \text{ если } x \geq a \end{cases}$$

Плотность распределения

$$f(x) = \begin{cases} 1, \text{ если } x = a \\ 0, \text{ во всех остальных случаях} \end{cases}$$

5.2 Биномиальное распределение

Пусть X – случайная величина, которая есть число появлений события A в n независимых экспериментах, произведенных при одинаковых условиях (т.н. *испытания Бернулли*). Тогда случайная величина X распределена по *биномиальному закону*

$$P(X = m) = C_n^m p^m q^{n-m},$$

где $q = 1 - p$, $m = 0, \cdots, n$ – число появлений события A.

Биномиальное распределение однозначно задается двумя параметрами: количеством элементов выборки n и вероятностью p появления события A. Математическое ожидание и дисперсия биномиального распределения равны соответственно:

$$M[X] = n \cdot p,$$

$$D[X] = n \cdot p \cdot q.$$

5.3 Распределение Пуассона

Распределение Пуассона есть предельный случай биномиального распределения при определенных условиях. Если число испытаний по схеме Бернулли стремится к бесконечности ($n \to \infty$) и при этом вероятность числа появлений события A стремится к нулю так, что произведение $n \cdot p$ остается конечным и постоянным, то биномиальное распределение переходит в распределение Пуассона.

Распределение Пуассона очень важно, в частности, для задач астрономии, потому что описывает распределение вероятности *редких событий*.

Если вероятность осуществления события A в интервале[5] δx равна $\lambda \delta x$, где λ – есть постоянная величина, то вероятность того, что в ограниченном интервале Δx событие A произойдет ровно k раз, дается распределением Пуассона:

$$p_k = \frac{(\lambda \cdot \Delta x)^k}{k!} e^{-\lambda \cdot \Delta x}.$$

[5] Это может быть интервал пространства, времени, а также длина, площадь, объем и др. в зависимости от условия задачи

Математическое ожидание и дисперсия распределения Пуассона равны друг другу:

$$M[X] = D[X] = \lambda \cdot \Delta x.$$

Распределение Пуассона можно использовать в качестве приближения для биномиального распределения (при малых $p_k \leq 0.10$),

$$\lambda \cdot \Delta x = n \cdot p.$$

ПРИМЕРЫ Число распадов радиоактивного вещества за время t; число космических частиц, попадающих на поверхность площади S за время t.

5.3.1 Понятие пуассоновского поля

Случайное поле точек называется *пуассоновским полем*, если выполняются следующие условия:

- точки распределяются в поле статистически равномерно со средней плотностью λ (величина на единицу площади или на единицу объема);
- точки попадают в непересекающиеся области независимо одна от другой;
- точки попадают в малый элемент площади (или объема) по одной, а не парами, тройками и т.д.

При выполнении этих условий число точек, попадающих в любую область g (плоскую или объемную) распределено по закону Пуассона:

$$p_k(g) = \frac{a^k}{k!} e^{-a},$$

где $a = S_g \cdot \lambda$ (для распределения на плоскости) и $a = V_g \cdot \lambda$ (для распределения в объеме).

ПРИМЕР Система ICRF (международная небесная система отсчета, сформированная по далеким источникам, преимущественно квазарам).

5.4 Геометрическое распределение

Геометрическим распределением называется закон распределения числа X независимых опытов с двумя исходами $\{A, \bar{A}\}$ в одинаковых условиях $P(A) = p, P(\bar{A}) = 1-p = q$ до первого появления A.

$$P(\{\textit{Событие } A \textit{ впервые появилось в опыте номер } m\}) = P(X = m) = q^{m-1}p.$$

Математическое ожидание и дисперсия случайной величины X есть, соответственно,

$$M[X] = \frac{1}{p},$$

$$D[X] = \frac{q}{p^2}.$$

5.5 Показательное распределение

Показательное или *экспоненциальное* распределение случайной величины X характеризуется плотностью распределения $f(x)$:

$$f(x) = \begin{cases} \lambda \cdot e^{-\lambda \cdot x} & \text{если } x \geq 0 \\ 0 & \text{если } x < 0 \end{cases}$$

Математическое ожидание и дисперсия случайной величины, имеющей показательное распределение, равны, соответственно,

$$M[X] = \frac{1}{\lambda},$$

$$D[X] = \frac{1}{\lambda^2}.$$

5.6 Равномерное распределение

Равномерное распределение случайной величины X характеризуется плотностью распределения $f(x)$:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{если } x \in [a, b] \\ 0 & \text{если } x < a \text{ и } x > b \end{cases}$$

Математическое ожидание и дисперсия случайной величины, имеющей равномерное распределение, равны, соответственно,

$$M[X] = \frac{a+b}{2},$$

$$D[X] = \frac{(b-a)^2}{12}.$$

5.7 Нормальное распределение

Нормальное распределение (или закон Гаусса) играет фундаментальную роль в теории ошибок, потому что

- нормальное распределение описывает распределение ошибок, возникающих при множестве малых независимых вкладов, носящих случайных характер;
- многие функции случайной величины – например, среднее значение или среднеквадратическое отклонение – распределены асимптотически нормально даже тогда, когда исходная случайная величина не обладает нормальным распределением.

5.7.1 Основные понятия

Плотность вероятности для случайной величины X, имеющей *нормальное распределение*:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\},$$

а функция распределения

$$F(x) = \int_{-\infty}^{x} f(t)dt = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left\{ \frac{-(t-\mu)^2}{2\sigma^2} \right\} dt.$$

Нормальное распределение полностью определяется двумя параметрами:μ и σ, которые есть математическое ожидание и среднеквадратическое отклонение, соответственно:

$$M[X] = \mu,$$
$$D[X] = \sigma^2.$$

Крутизна (или эксцесс) нормально распределенной величины X равен нулю:

$$\gamma_2 = \frac{\mu_4[X]}{\sigma^4[X]} - 3 = 0,$$

где $\mu_4[X]$ – центральный момент четвертого порядка, а $\sigma[X]$ – среднеквадратическое отклонение.

Величину X, распределенную по нормальному закону, обозначают

$$X \sim N(\mu, \sigma^2).$$

Для удобства работы с нормальной распределенными величинами и для подсчета необходимых вероятностей с помощью статистических таблиц (для того, чтобы не вычислять каждый раз интегралы функции распределения), вводят замену переменной

$$U = \frac{X - \mu}{\sigma}.$$

Тогда случайная величина U имеет т.н. *стандартное нормальное распределение*

$$U \sim N(0, 1).$$

Функция распределения для случайной величины U запишется следующим образом:

$$F(u) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{u} \exp\left\{ -\frac{v^2}{2} \right\} \sigma dv,$$

поскольку

$$u = \frac{x - \mu}{\sigma},$$
$$x = \sigma u + \mu,$$
$$dt = \sigma \cdot du,$$
$$\frac{1}{2}\frac{(t-\mu)^2}{\sigma^2} = \frac{u^2}{2}.$$

Функция распределения для стандартной нормальной величины обозначается $\Phi(u)$:

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{t^2}{2}} dt,$$

Эта функция задана таблично. Еще более удобно использовать т.н. *функцию Лапласа-Гаусса*, также заданную таблично, которая есть:

$$\Phi_0(u) = \frac{1}{\sqrt{2\pi}} \int_{0}^{u} e^{-\frac{t^2}{2}} dt.$$

Для функций $\Phi(u)$ и $\Phi_0(u)$ выполняются простые свойства, следующие из вида соответствующих интегралов:

- $\Phi(-u) = 1 - \Phi(u)$;
- $\Phi(u) = \frac{1}{2} + \Phi_0(u)$;
- $\Phi_0(-u) = -\Phi_0(u)$;
- $\Phi_0(0) = 0$;
- $\Phi_0(+\infty) = \frac{1}{2}$.

5.7.2 Центральная предельная теорема

Если в биномиальном распределении вероятность p фиксирована, число элементов выборки стремится к бесконечности, $n \to \infty$, то распределение такой случайной величины стремится к нормальному распределению.

ЦЕНТРАЛЬНАЯ ПРЕДЕЛЬНАЯ ТЕОРЕМА Пусть случайная величина X имеет среднее значение μ и дисперсию σ^2. Если дисперсия σ^2 конечна, то при стремлении объема выборки к бесконечности, $n \to \infty$, распределение *выборочного среднего* $\bar{x}$ будет стремиться к нормальному распределению со средним μ и дисперсией σ^2/n. Другими словами,

$$M[\bar{x}] = \mu,$$

$$D[\bar{x}] = \frac{\sigma^2}{n}.$$

Учитывая, что для любой дискретной случайной величины, в данном случае для $\bar{x}$

$$\sigma^2(\bar{x}) = \sum_{i=1}^{n} \Big(\bar{x}_i - M[\bar{x}]\Big)^2 \cdot p(\bar{x}_i) = \sum_{i=1}^{n} \Big(\bar{x}_i - M[\bar{x}]\Big)^2 \cdot \frac{1}{n},$$

получим для дисперсии выборочного среднего

$$D[\bar{x}] = \frac{1}{n^2} \sum_{i=1}^{n} (\bar{x}_i - \mu)^2.$$

В задачах обработки данных считается, что величина x_i (i-ая реализация случайной величины X) и величина $\bar{x}_i$ (i-ая реализация выборочного среднего случайной величины X или *среднее по выборке*) есть одно и то же, поскольку каждую x_i можно считать как некое среднее значение (*среднее по реализациям*). Одна реализация – это, например, одна серия наблюдений или один «проход» экспериментальной установки. Среднее по выборке равно среднему по реализациям и потому заменим в последней формуле $\bar{x}_i$ на x_i. Кроме того, отметим еще один факт, объяснение которому будет дано ниже, при обсуждении качества оценок случайных величин. Наиболее «качественная» оценка дисперсии произвольной случайной величины Y есть

$$D[Y] = \sum_{i=1}^{n} \Big(y_i - M[Y]\Big)^2 \cdot \frac{1}{n-1},$$

т.е. n в знаменателе заменяется на $n-1$ (хотя при большом объеме выборке эта поправка несущественна).

Учитывая все вышесказанное, получаем наиболее широко используемую формулу для оценки среднеквадратического отклонения выборочного среднего:

$$\sigma_{\bar{x}} = \sqrt{\frac{1}{n \cdot (n-1)} \sum_{i=1}^{n} (x_i - \mu)^2},$$

где x_i – результаты наблюдений или экспериментов, μ – среднее арифметическое результатов наблюдений, n – количество наблюдений. Эта формула является следствием центральной предельной теоремы.

5.7.3 Доказательство центральной предельной теоремы

Для доказательства центральной предельной теоремы нам понадобится понятие *производящей функции*. Производящая функция $\psi(t)$ ($t \in \Re$) есть преобразование Лапласа

$$\psi(t) = \psi_X(t) = M[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f(x) dx.$$

В частности, производящая функция нормально распределенной случайной величины $X \sim N(\mu, \sigma^2)$ есть

$$\psi_X(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right).$$

Первая производная производящей функции в нуле есть

$$\psi'(0) = \left[\frac{d}{dt} M[e^{tX}]\right]_{t=0} = M\left[\frac{d}{dt} e^{tX}\right]_{t=0} = M\left[X e^{tX}\right]_{t=0} = M[X].$$

Производная порядка k от производящей функции есть, соответственно,

$$\psi^k(0) = M[X^k].$$

Если случайная величина Y есть линейная функция случайной величины X с неслучайными коэффициентами a и b

$$Y = aX + b,$$

а $\psi_X(t)$ есть производящая функция случайной величины X, то производящая функция случайной величины Y есть

$$\psi_Y(t) = e^{bt} \cdot \psi_X(at)$$

Если $X_1, X_2, \ldots, X_n$ – независимые случайные величины и $Y = \sum_i X_i$, то

$$\psi_Y(t) = \prod_i \psi_i(t),$$

где $\psi_i(t)$ есть производящие функции случайных величин X_i.

Для доказательства центральной предельной теоремы возьмем случайные величины Y_i в виде

$$Y_i = \frac{X_i - \mu}{\sigma} \quad (i = 1, 2, \ldots, n),$$

где X_i – независимые случайные величины, обладающие средним μ и дисперсией σ^2. Обозначим

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i.$$

Пусть $\psi_{Y_i}(t) = \psi(t)$ – производящая функция случайной величины Y_i. Тогда производящая функция для $\sum_i Y_i$ есть $\psi(t)^n$, а производящая функция для Z_n есть

$$\xi_n(t) \equiv \left[\psi\left(\frac{t}{\sqrt{n}}\right)\right]^n.$$

Далее, воспользуемся вычисленными ранее производными в нуле производящей функции ($\psi'(0) = M[Y_1] = 0$, $\psi''(0) = M[Y_1^2] = D[Y_1] = 1$) и представим эту производящую функцию в виде ряда:

$$\psi(t) = \psi(0) + t\psi'(0) + \frac{t^2}{2!}\psi''(0) + \frac{t^3}{3!}\psi'''(0) + \cdots = 1 + 0 + \frac{t^2}{2!} + \frac{t^3}{3!}\psi'''(0) + \ldots$$

Тогда

$$\xi_n(t) \equiv \left[\psi\left(\frac{t}{\sqrt{n}}\right)\right]^n = \left[1 + \frac{t^2}{n \cdot 2!} + \frac{t^3}{n^{3/2} \cdot 3!}\psi'''(0) + \ldots\right]^n = \left[1 + \frac{\frac{t^2}{2!} + \frac{t^3}{n^{1/2} \cdot 3!}\psi'''(0) + \ldots}{n}\right]^n.$$

Воспользуемся известным из курса математического анализа пределом

$$\lim_{n\to\infty}\left[\left(1 + \frac{a_n}{n}\right)^n\right] = e^a,$$

где a – предел последовательности $\{a_n\}$.

Тогда

$$\xi(t) = \lim_{n\to\infty} \xi_n(t) = e^{t^2/2}.$$

Другими словами, $\xi(t)$ есть производящая функция величины, распределенной по стандартному нормальному закону $N(0, 1)$.

Окончательно,

$$Z_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i - \mu}{\sigma} = \frac{1}{\sqrt{n}}\frac{n \cdot \bar{x} - n \cdot \mu}{\sigma} = \frac{\sqrt{n} \cdot (\bar{x} - \mu)}{\sigma},$$

где $\bar{x}$ – выборочное среднее, и поскольку Z_n в пределе имеет распределение $N(0, 1)$, то

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

5.7.4 Правило «трех сигма»

Вычислим вероятность $P(X \in [\alpha, \beta))$ для частного случая, когда границы интервала симметричны относительно среднего значения случайной величины, т.е.

$$\alpha = \mu - l,$$

$$\beta = \mu + l.$$

Тогда, по определению функции распределения и функции Лапласа-Гаусса, получаем:

$$P(\mu - l \le X < \mu + l) = P\left(-\frac{l}{\sigma} \le \frac{X - \mu}{\sigma} < \frac{l}{\sigma}\right) = \Phi_0\left(\frac{l}{\sigma}\right) - \Phi_0\left(-\frac{l}{\sigma}\right) = 2\Phi_0\left(\frac{l}{\sigma}\right).$$

Таким образом, для нормально распределенной случайной величины X вероятность ее отклонения от среднего на величину l определяется как

$$P\Big(|X-\mu|<l\Big)=2\Phi_0\Big(\frac{l}{\sigma}\Big).$$

Число l, вообще говоря, любое положительное число. Однако особую важность представляют значения $l=\sigma,2\sigma,3\sigma$. Так, при $l=\sigma$:

$$P\Big(|X-\mu|<\sigma\Big)=2\Phi_0(1)=0.683\approx\frac{2}{3},$$

другими словами, примерно в двух третях случаев величина отклонения нормально распределенной случайной величины от своего среднего значения не превышает своего стандартного отклонения – это т.н. правило «одного сигма».

Аналогично определяется правило «двух сигма»:

$$P\Big(|X-\mu|<2\sigma\Big)=2\Phi_0(2)=0.954$$

и правило «трех сигма»:

$$P\Big(|X-\mu|<3\sigma\Big)=2\Phi_0(3)=0.997.$$

Согласно последнему равенству, все значения случайной величины X, распределенной по нормальному закону, с вероятностью 99.7% укладываются в интервале $[\mu-3\sigma,\mu+3\sigma)$[6]

5.7.5 Правила работы со статистическими таблицами нормального распределения

В математических справочниках и в приложениях учебников по теории вероятностей и математической статистике обычно приводятся таблицы для значений нормированной нормальной функции распределения (или функции распределения для стандартной нормальной величины) $\Phi(u)$ и для нормального интеграла вероятностей (или функции Лапласа-Гаусса) $\Phi_0(u)$:

$$\Phi(u)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u}e^{-\frac{t^2}{2}}dt, \tag{6}$$

$$\Phi_0(u)=\frac{1}{\sqrt{2\pi}}\int_{0}^{u}e^{-\frac{t^2}{2}}dt. \tag{7}$$

Таблица функции $\Phi(u)$ содержит вычисленный интеграл (6) для каждого u. Другими словами, таблица содержит *вероятности* того, что случайная величина

$$U=\frac{X-\mu}{\sigma}$$

[6]заметим, что формальное – следующее из определения функции распределения – невключение правого конца интервала не влияет на вычисление вероятностей, поскольку для непрерывного распределения вероятность в одной точке есть ноль.

(где X – случайная величина, обладающая нормальным законом распределения со средним μ и среднеквадратическим отклонением σ) принимает значения, меньше u. Соответствующие величины u заданы с интервалом в 0.1 в крайнем левом столбце таблицы. Если требуется найти $\Phi(u)$ для u, заданных с лучшей точностью, до 0.01, используется первая строка таблицы, где указаны сотые доли u. К примеру, для $u = 0.8$ значение $\Phi(u) = 0.7881$, для $u = 0.86$ значение $\Phi(u) = 0.8051$ (на пересечении строки $u = 0.8$ и столбца $u = 0.06$).

Область определения величины u – вся числовая ось, $\{-\infty, +\infty\}$, и смысл имеют любые значения в данном интервале. Обращаем внимание, что в таблице величина u меняется от 0 до, примерно, 4, потому что, во-первых, для отрицательных u, в силу симметрии стандартного нормального распределения, можно воспользоваться формулой:

$$\Phi(-u) = 1 - \Phi(u),$$

а во-вторых для больших u вероятность близка к единице: так, уже для $\Phi(3.79) = 0.9999$ и с ростом u только растет в силу своей монотонности.

Таблица для функции Лапласа-Гаусса $\Phi_0(u)$ содержит вычисленный интеграл (7) для каждого u и устроена полностью аналогично таблице для функции $\Phi(u)$. Все вероятности, найденные по ее таблице, могут быть получены из таблицы функции $\Phi(u)$ путем вычитания $1/2$, потому что

- $\Phi_0(u) = \Phi(u) - \frac{1}{2}$;
- $\Phi_0(-u) = -\Phi_0(u)$;
- $\Phi_0(0) = 0$;
- $\Phi_0(+\infty) = \frac{1}{2}$.

Приведем примеры задач на нормальное распределение.

ПРИМЕР Изготовлена цилиндрическая деталь диаметром D, [2]. Ошибки при ее изготовлении приводят к тому, что диаметр D есть случайная величина, распределенная по нормальному закону с параметрами: математическое ожидание $\mu = 40$ мм, среднеквадратическое отклонение $\sigma = 0.05$ мм. Деталь проходит технологический контроль, в результате которого признаны браком все детали с диаметром D таким, что $D < 39.85$ мм или $D > 40.05$ мм.

Определить вероятность того, что наугад выбранная для контроля деталь будет признана бракованной (событие A), и определить процент забракованных деталей.

Задача сводится к определению вероятности $P(A)$ попадания случайной величины D, распределенной по нормальному закону со средним $\mu = 40$ мм, среднеквадратическим отклонением $\sigma = 0.05$ мм, за пределы отрезка $[\alpha, \beta]$, $\alpha = 39.85$ мм и $\beta = 40.05$ мм, где событие $A = \{D < \alpha \text{ или } D > \beta\}$.

Решим задачу, используя противоположное событие $\bar{A} = \{D \in [\alpha, \beta]\}$. Тогда

$$P(A) = 1 - P(\bar{A}).$$

Вероятность $P(\bar{A})$ вычислим с использованием таблиц функции Лапласа-Гаусса $\Phi_0(u)$:

$$P(\bar{A}) = P\Big(D \in [39.85, 40.05]\Big) = \Phi_0\left(\frac{40.05 - 40.00}{0.05}\right) - \Phi_0\left(\frac{39.85 - 40.00}{0.05}\right) =$$

$$= \Phi_0(1) - \Phi_0(-3) = \Phi_0(1) + \Phi_0(3) = 0.341 + 0.499 = 0.840.$$

Вычислим средний процент забракованных деталей:

$$P(A) = 1 - 0.84 = 0.16 = 16\%$$

ПРИМЕР Максимальная ошибка высотомера $\Delta H_{max} = 30$ м. Найти вероятность того, что ошибка измерения высоты не превысит 10 м.

Используем правило «трех сигма» для нахождения величины среднеквадратического отклонения случайной величины Δh:

$$3\sigma_{\Delta h} = \Delta H_{max},$$

откуда $\sigma = 10$ м. Искомая вероятность равна

$$P(|\Delta h - m_{\Delta h}| < 10) = 2\Phi_0(1) = 0.683.$$

5.8 Распределения, близкие к нормальному

Существует ряд распределений, отличных от нормального, в силу физических свойств наблюдаемых величин.

Например, строго говоря, распределение параллаксов звезд, поскольку кривая распределения ограничена справа и слева, в отличие от нормального распределения (все параллаксы больше нуля и не существует очень больших параллаксов; кроме того, с уменьшением параллакса растет число звезд).

Еще один пример – распределение модулей скоростей группы движущихся астероидов, поскольку они неотрицательны и нет бесконечно больших скоростей.

Часто, для простоты, к таким распределениям все-таки применяют нормальный закон, оговаривая, на каком интервале и при каких дополнительных условиях нормальный закон хорошо объясняет наблюдательные данные. Однако полезно знать о других теоретических распределениях, близких к нормальному, но таковым все же не являющимися, что позволит аппроксимировать наблюдательные и экспериментальные данные более точными кривыми.

Рассмотрим плотность распределения

$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \cdot \Pi(x),$$

где $\Pi(x)$ – многочлен не выше 4-й степени по переменной x. Для определения пяти коэффициентов полинома 4-й степени можно записать пять уравнений, определяя моменты от нулевого до четвертого порядка. Если сделать замену переменной $u = x - \mu$, тогда для $\Pi(u)$:

$$a_0 = 1 + \frac{1}{8}\left[\frac{\mu_4}{\sigma^4} - 3\right];$$

$$a_1 = \frac{1}{2} \cdot \frac{1}{\sigma} \cdot \left[\frac{\mu_3}{\sigma^3}\right];$$

$$a_2 = -\frac{1}{4} \cdot \frac{1}{\sigma^2} \cdot \left[\frac{\mu_4}{\sigma^4} - 3\right];$$

$$a_3 = \frac{1}{6} \cdot \frac{1}{\sigma^3} \cdot \left[\frac{\mu_3}{\sigma^3}\right];$$

$$a_4 = \frac{1}{24} \cdot \frac{1}{\sigma^4} \cdot \left[\frac{\mu_4}{\sigma^4} - 3\right],$$

где

$$\gamma_1 = \frac{\mu_3}{\sigma^3}$$

есть асимметрия, а

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$$

есть эксцесс.

Величина $\mu_k = M[(x-\mu)^k]$ есть k-тый центральный момент случайной величины X, которая определяется по выборке.

Кроме полиномиальной «корректировки» нормального распределения возможны также обобщения закона нормального распределения путем введения переменного среднеквадратического отклонения $\sigma = \sigma(u)$. Заметим также, что если эмпирическое распределение содержит два максимума, то удобно представить его суммой двух распределений.

6 Точечные и интервальные оценки

Цель математической статистики – указать методы, с помощью которых по данным выборки можно получить оценки параметров генеральной совокупности. Более подробно, пусть производится наблюдение какой-либо случайной величины. Если наблюдения ведутся достаточно долго (в идеале бесконечное количество времени), то по результатам наблюдений можно точно вычислить такие параметры как, например, среднее, среднеквадратическое отклонение и среднеквадратическое отклонение среднего. Однако в реальности наблюдатель никогда не имеет дело с бесконечным набором наблюдений случайной величины (или с генеральной совокупностью). Поэтому параметры генеральной совокупности остаются недостижимыми. В распоряжении наблюдателя имеется только ограниченный набор данных (выборка) и только с помощью этого набора нужно и можно получить по возможности лучшее представление о параметрах генеральной совокупности.

Обозначим параметры генеральной совокупности: среднее μ, среднеквадратическое отклонение σ. Эти величины будем оценивать с помощью выборочного среднего $\bar{x}$, выборочного среднеквадратического отклонения s и среднеквадратического отклонения выборочного среднего $s(\bar{x})$.

Оценки могут быть *точечными* и *интервальными*. Точечная оценка определяется одним числом, например, точечной оценкой среднего генеральной совокупности μ является среднее арифметическое элементов выборки. Интервальная оценка указывает доверительный интервал для точечной оценки, т.е., насколько хороша эта оценка. Например, с ростом числа элементов выборки интервальная оценка должна становиться *уже*, поскольку чем больше выборка, тем ближе оценка к истинному значению параметра. Интервальная оценка записывается как

$$J_\theta = \{\theta^* - \epsilon, \theta^* + \epsilon\},$$

где θ – какой-то из оцениваемых параметров генеральной совокупности, например, μ или σ. Величина θ^* есть точечная оценка параметра θ, ϵ есть *точность оценки* (зависящая, в том числе, от размера выборки).

Вероятность того, что оценка равна оцениваемому параметру на уровне точности ϵ есть

$$\gamma = P(|\theta^* - \theta| < \epsilon)$$

и называется *доверительной вероятностью* или *надежностью* оценки.

6.1 Оценка вероятности случайного события

Пусть нужно получить точечную оценку вероятности $P(A)$, где A – случайное событие. Обозначим $P(A) = p$. Точечная оценка для p есть частота появления события A, т.е.

$$p^* = \frac{X}{n},$$

где X – число опытов, в которых событие A произошло, а n – число всех проведенных опытов (количество элементов выборки). Пусть происходит повторение серий из n опытов каждая. Тогда величина X есть случайная величина. Запишем X в виде

$$X = \sum_{i=1}^{n} X_i(A),$$

где

$$X_i(A) = \begin{cases} 1 & \text{если событие } A \text{ появилось (с вероятностью } p) \\ 0 & \text{если событие } A \text{ не появилось (с вероятностью } q = 1 - p). \end{cases}$$

Тогда

$$M[X_i(A)] = 1 \cdot p + 0 \cdot q = p,$$
$$D[X_i(A)] = (1-p)^2 \cdot p + (0-p)^2 \cdot q = p \cdot q.$$

В качестве точечной оценки для p будем рассматривать величину p^*:

$$p^* = \frac{1}{n} \sum_{i=1}^{n} X_i(A),$$

$$M[p^*] = \frac{1}{n} \sum_{i=1}^{n} M[X_i(A)] = \frac{1}{n} \sum_{i=1}^{n} p = p,$$
$$D[p^*] = \frac{1}{n^2} \sum_{i=1}^{n} D[X_i(A)] = \frac{1}{n^2} \sum_{i=1}^{n} pq = \frac{pq}{n}$$

Согласно центральной предельной теореме

$$p^* \sim N\Big(M[p^*], D[p^*]\Big) = \Big(p, \frac{pq}{n}\Big).$$

Теперь построим интервальную оценку для p:

$$P\Big(|p - p^*| < \epsilon\Big) = \gamma.$$

Для нормально распределенной случайной величины $p \sim N(p^*, D[p^*])$ вероятность ее отклонения от своей оценки p^* на величину ϵ определяется с помощью функции Лапласа-Гаусса следующим образом

$$P\Big(|p - p^*| < \epsilon\Big) = 2\Phi_0\Big(\frac{\epsilon}{\sqrt{D[p^*]}}\Big) = 2\Phi_0\Big(\frac{\epsilon\sqrt{n}}{\sqrt{p^* \cdot q^*}}\Big).$$

Напомним, что ϵ – точность оценки, γ – *доверительная вероятность* (или *надежность*, или *достоверность*) оценки. Величина $\alpha = 1 - \gamma$ называется *уровнем значимости* или *процентной точкой* (иногда обозначается в процентах, например, $\alpha = 5\%$). Величина p^* – точечная оценка параметра p, а величина $q^* = 1 - p^*$.

Обозначим

$$u_\gamma = \frac{\epsilon\sqrt{n}}{\sqrt{p^* \cdot q^*}}.$$

Тогда

$$\gamma = 2\Phi_0(u_\gamma)$$

и u_γ, которая называется *квантиль уровня* γ есть функция, обратная к функции Лапласа-Гаусса:

$$u_\gamma = \Phi_0^{-1}\Big(\frac{\gamma}{2}\Big).$$

Получаем полезную связь между точностью оценки, надежностью оценки и числом элементов выборки:

$$\gamma = 2\Phi_0\left(\frac{\epsilon\sqrt{n}}{\sqrt{p^* \cdot q^*}}\right),$$

$$\epsilon = \frac{u_\gamma\sqrt{p^* \cdot q^*}}{\sqrt{n}},$$

$$n = \frac{u_\gamma^2 p^* \cdot q^*}{\epsilon^2}.$$

Окончательно, искомый доверительный интервал для p^* есть

$$J_{p^*} = \{p^* - \epsilon, p^* + \epsilon\},$$

который с вероятностью γ накрывает истинное (всегда неизвестное) значение параметра p.

ПРИМЕР Вычисление точечной и интервальной оценки. Произведено десять испытаний однотипных авиационных двигателей, в семи из которых были достигнуты требуемые показатели тяговооруженности. Определить точечную и интервальную оценки вероятности события $A = \{$ требуемые показатели тяговооруженности достигнуты $\}$ при заданной надежности $\gamma = 0.95$.

Точечная оценка искомой вероятности есть частота события A:

$$p^* = \frac{7}{10} = 0.70.$$

Зная надежность (доверительную вероятность) γ, можно вычислить точность этой точечной оценки ϵ, т.е. построить доверительный интервал точечной оценки. Для этого сначала вычислим квантиль u_γ с помощью функции Лапласа-Гаусса $\Phi_0(u)$:

$$u_\gamma = \Phi_0^{-1}\left(\frac{\gamma}{2}\right).$$

или с помощью функции распределения стандартной нормальной величины, $\Phi(u)$

$$u_\gamma = \Phi^{-1}\left(\frac{1+\gamma}{2}\right).$$

Для заданной надежности $\gamma = 0.95$

$$u_{0.95} = \Phi_0^{-1}(0.475) = \Phi^{-1}(0.975) = 1.96.$$

(ищем в таблице значений функции Лапласа-Гаусса $\Phi_0(u)$ величину 0.475 и смотрим, какой величине u (десятой и сотой части) соответствует эта вероятность; либо, соответственно, ищем в таблице значений функции распределения стандартной нормальной величины $\Phi(u)$ величину 0.975 и смотрим, какой величине u (десятой и сотой части) соответствует эта вероятность).

Найденное значение квантили $u_\gamma = 1.96$ подставляем в выражение для точности оценки:

$$\epsilon = \frac{u_\gamma\sqrt{p^* \cdot q^*}}{\sqrt{n}} = \frac{1.96\sqrt{0.70 \cdot (1-0.70)}}{\sqrt{10}} = 0.28.$$

Таким образом, неизвестная вероятность p с надежностью 95% лежит в доверительном интервале:

$$J_{p^*} = \{0.70 - 0.28, 0.70 + 0.28\} = \{0.42, 0.98\}.$$

ПРИМЕР Ранее рассматривалось неравенство Хефдинга. Для независимых случайных величин, распределенных по закону Бернулли, $X_1, X_2, \ldots, X_n$ с параметром p

$$P\left(\left|\bar{x}_n - p\right| > \epsilon\right) \leq 2 \cdot e^{-2n\epsilon^2},$$

где $\bar{x}$ – среднее выборочное значение. Это неравенство позволяет, в частности, построить доверительный интервал для биномиального параметра p:

$$P\left(\left|\bar{x} - p\right| > \epsilon\right) \leq \alpha,$$

где

$$\epsilon_n = \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}.$$

6.1.1 Геометрическая интерпретация доверительного интервала оценки вероятности

В общем виде результат предыдущего примера запишем как

$$|p - p^*| < u_\gamma \sqrt{\frac{p \cdot q}{n}}.$$

Возведя обе части этого неравенства в квадрат, получим область, которая на плоскости в координатах (p, p^*) есть внутренняя часть эллипса, [2]:

$$(p - p^*)^2 = \frac{u_\gamma^2}{n} \cdot p \cdot (1 - p). \tag{8}$$

Прямая, параллельная оси Op и проходящая через фиксированную точку p^* (в примере $p^* = 0.70$), пересечет эллипс в двух точках. Длина этого сечения есть доверительный интервал J_{p^*}. Точки пересечения в общем виде вычисляются из квадратного уравнения границы эллипса:

$$p_{1,2} = \frac{p^* + \dfrac{u_\gamma^2}{2n}}{1 + \dfrac{u_\gamma^2}{n}} \pm \frac{1}{1 + \dfrac{u_\gamma^2}{n}} \cdot \sqrt{\left(p^* + \frac{u_\gamma^2}{2n}\right)^2 - (p^*)^2 \cdot \left(1 + \frac{u_\gamma^2}{n}\right)}$$

При большом объеме выборки u_γ^2/n и u_γ^2/n^2 стремятся к нулю быстрее, чем $p^* \cdot (1 - p^*)/n$, поэтому

$$p_{1,2} = p^* \mp u_\gamma \cdot \sqrt{\frac{p^* \cdot (1 - p^*)}{n}},$$

что равносильно тому, как если бы в правой части уравнения (8) стояла в точности точечная оценка p^*.

Чем больше размер выборки n, тем меньше доверительный интервал и, значит, тем *уже* эллипс.

6.2 Оценка математического ожидания

6.2.1 Точечная оценка математического ожидания

В качестве точечной оценки математического ожидания принимают выборочное среднее (среднее арифметическое всех элементов выборки):

$$\mu^* = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Сразу же возникает вопрос, можно ли в качестве точечной оценки математического ожидания выбрать что-то другое, например, моду или медиану, которые тоже являются характеристиками положения среднего. Оказывается, среднее арифметическое является лучшей оценкой и это можно доказать, например, с помощью *метода максимальноˊго правдоподобия (ММП)* или *метода наименьших квадратов (МНК)*.

6.2.2 Использование метода максимального правдоподобия для поиска точечной оценки математического ожидания

На примере поиска точечной оценки математического ожидания случайной величины, рассмотрим, что такое *метод максимальноˊго правдоподобия*.

Предположим, что вид закона распределения генеральной совокупности известен, но неизвестны параметры, конкретизирующие этот закон (например, известно, что генеральная совокупность распределена по нормальному закону, но неизвестно ни его среднее μ, ни его дисперсия σ^2). Пусть из генеральной совокупности извлечена некоторая выборка данных и по ней нужно оценить μ и σ^2. Метод максимального правдоподобия заключается в том, что выбираются такие оценки параметров, которые дают максимальное значение плотности вероятности (другими словами, вероятность которых максимальна).

Если $x_i (i = 1, 2, \cdots, n)$, где все элементы *не зависят друг от друга*, составляет выборку с плотностью вероятности $f(x_1, x_2, \cdots, x_n, \mu^*, \sigma^*)$, то совместная плотность вероятность есть

$$l(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i, \mu^*, \sigma^*),$$

которая и называется функцией правдоподобия. Обозначим L:

$$L = \ln\left[l(x_1, x_2, \cdots, x_n)\right].$$

Тогда для того, чтобы μ^* и σ^* давали максимальное значение плотности вероятности, необходимо:

$$\frac{\partial L}{\partial \mu^*} = \frac{\partial}{\partial \mu^*}\sum_{i=1}^{n} \ln\left[f(x_i, \mu^*, \sigma^*)\right] = 0,$$

$$\frac{\partial L}{\partial \sigma^*} = \frac{\partial}{\partial \sigma^*}\sum_{i=1}^{n} \ln\left[f(x_i, \mu^*, \sigma^*)\right] = 0.$$

Для нормального распределения

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot \exp -\frac{(x-\mu)^2}{2\sigma^2}.$$

Мы хотим оценить только математическое ожидание, поэтому только его обозначим «со звездочкой», μ^*. Поскольку все x_i – независимые, то

$$f(x_1, x_2, \cdots, x_n) = f(x_1, \mu^*, \sigma) \cdot f(x_2, \mu^*, \sigma) \cdots f(x_n, \mu^*, \sigma) =$$

$$= \frac{1}{\left[\sigma \cdot \sqrt{2\pi}\right]^n} \cdot \exp\left\{ -\sum_{i=1}^{n} \frac{(x_i - \mu^*)^2}{2\sigma^2} \right\} = l(x, \mu^*).$$

$$L = \ln l = \ln\left(\sigma \cdot \sqrt{2\pi}\right)^{-n} - \sum_{i=1}^{n} \frac{(x_i - \mu^*)^2}{2\sigma^2}.$$

$$\frac{\partial}{\partial \mu^*} L = -\frac{1}{2\sigma^2} \cdot 2 \cdot (-1) \cdot \sum_{i=1}^{n} (x_i - \mu^*) = 0.$$

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu^*) = 0.$$

Таким образом, искомая оценка и есть среднее арифметическое:

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

что и требовалось доказать.

6.2.3 Использование метода наименьших квадратов для поиска точечной оценки математического ожидания

Рассмотрим метод наименьших квадратов также для поиска точечной оценки математического ожидания случайной величины. Суть метода заключается в поиске такой оценки, которая минимизирует сумму квадратов отклонений отдельных реализаций случайной величины от искомой оценки. Минимизация осуществляется по всей выборке. Другими словами, ищется минимум

$$\sum_{i=1}^{n} (x_i - \mu^*)^2$$

по μ^*. В точке минимума первая производная по μ с необходимостью равно 0:

$$\frac{d}{d\mu^*} \left(\sum_{i=1}^{n} (x_i - \mu^*)^2 \right) = -2 \sum_{i=1}^{n} (x_i - \mu^*) = 0,$$

откуда

$$\mu^* = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Таким образом, метод наименьших квадратов и метод максимального правдоподобия дают одну и ту же точечную оценку математического ожидания: среднее арифметическое всех элементов выборки. Кроме того, при использовании метода наименьших квадратов не делалось никакого предположения относительно закона распределения случайной величины – таким образом, среднее арифметическое является хорошей точечной оценки для выборки, обладающей произвольным законом распределения.

6.2.4 Интервальная оценка математического ожидания

Теперь найдем интервальную оценку математического ожидания, определив, тем самым, от чего зависит качество этой оценки.

Пусть X_i – случайные результаты наблюдений, независимые, одинаково распределенные нормальные случайные величины, такие, что:

$$M[X_i] = \mu$$

$$D[X_i] = \sigma^2.$$

Тогда величина

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

есть точечная оценка математического ожидания с параметрами

$$M\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum M[x_i] = \frac{1}{n}\cdot n\cdot \mu = \mu,$$

$$D\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n^2}\sum D[x_i] = \frac{1}{n^2}\cdot n\cdot \sigma^2 = \frac{\sigma^2}{n}.$$

Таким образом, случайна величина

$$\mu^* = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right),$$

а доверительный интервал для μ^*

$$P\left(\left|\mu - \mu^*\right| < \epsilon\right),$$

$$J_{\mu^*} = \{\mu^* - \epsilon, \mu^* + \epsilon\},$$

где точность оценки есть

$$\epsilon = \frac{u_\gamma \cdot \sigma}{\sqrt{n}},$$

$$u_\gamma = \frac{\epsilon\cdot\sqrt{n}}{\sigma} = \Phi_0^{-1}\left(\frac{\gamma}{2}\right).$$

Важно отметить, что среднеквадратическое отклонение σ *известно*, т.е. точность измерений должна быть задана. Если дисперсия задачи априори не известна и ее нужно находить по выборке, то имеет место задача оценки математического ожидания с неизвестной дисперсией. Точечные оценки в обоих случаях совпадают, однако незнание дисперсии во втором случае приводит к тому, что в интервальной оценке вместо квантили u_γ появится квантиль $t_{n,\gamma}$, имеющая т.н. *t-распределение Стьюдента с n степенями свободы* (имеющее соответствующие таблицы для своей функции распределения $T(t)$, и переходящее в нормальное при большом объеме выборки).

Итак, если дисперсия выборки априори *не известна*, то ее оценивают по выборке:

$$(\sigma^*)^2 = \frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i - \mu^*)^2,$$

где μ^*, как и раньше, оценивается по выборке своим средним арифметическим:

$$\mu^* = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Поясним, откуда берется $(n-1)$ в выражении для оценки дисперсии.

- Для вывода этой формулы вычислим для выборки математическое ожидание суммы квадратов отклонений отдельных выборочных значений от их среднего арифметического:

$$M\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right] = M\left[\sum_{i=1}^{n} x_i^2 - n\cdot\bar{x}^2\right] = \sum_{i=1}^{n} M[x_i^2] - n\cdot M[\bar{x}^2].$$

По свойству дисперсии и учитывая, что $M[x_i] = M[\bar{x}]$:

$$M[x_i^2] = \sigma^2 + (M[x_i])^2,$$

$$M[\bar{x}^2] = \sigma_{\bar{x}}^2 + (M[x_i])^2.$$

Как было показано ранее,

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

и

$$\sigma_{\bar{x}}^2 = \sum_{i=1}^{n}\frac{\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Таким образом,

$$M[\bar{x}^2] = \frac{\sigma^2}{n} + (M[x_i])^2,$$

$$M\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right] = n\cdot(\sigma^2 + (M[x_i])^2) - n\cdot\left(\frac{\sigma^2}{n} + (M[x_i])^2\right) = (n-1)\cdot\sigma^2.$$

Вместо среднего значения суммы квадратов отклонений подставим в точную формулу то ее значение, которое получается для одной выборки. Тогда приближенная формула для вычисления дисперсии одного измерения:

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

Обычно, чтобы не путать с априорно заданной точной дисперсией, приближенную дисперсию обозначают $(\sigma^*)^2$ или s^2. Отметим, что при выводе этой формулы не делалось никаких предположений о законе распределения случайной величины.

Оценка математического ожидания при неизвестной дисперсии имеет вид

$$P\left(\left|\mu^* - \mu\right| < \tilde{\epsilon}\right),$$

$$J_{\mu^*} = \{\mu^* - \tilde{\epsilon}, \mu^* + \tilde{\epsilon}\},$$

где точность оценки есть

$$\tilde{\epsilon} = \frac{t_{n,\gamma}\cdot\sigma^*}{\sqrt{n}},$$

где $t_{n,\gamma}$ определяется (по аналогии с функцией распределения стандартной нормальной величины, $\Phi(u)$) по статистическим таблицам распределения Стьюдента:

$$t_{n,\gamma} = T^{-1}\left(n, \frac{1+\gamma}{2}\right)$$

или, через процентный уровень значимости α,

$$t_{n,\gamma} = T^{-1}\left(n, 1 - \frac{1+\gamma}{2}\right) = T^{-1}\left(n, \frac{\alpha}{2}\right).$$

Здесь n – число степеней свободы, о величине которого речь пойдет ниже.

6.2.5 t-распределение Стьюдента

Распределение Стьюдента и использование таблиц этого распределения разберем на примере, [5].

ПРИМЕР Задана выборка в виде вариационного ряда – измерения роста 10-ти человек:

$$160, 160, 167, 170, 173, 176, 178, 178, 181, 181.$$

Нужно проверить, действительно ли средний рост большой группы людей равен $\mu^* = 167$ см.

Пусть случайная величина X есть величина роста. Пусть эта случайная величина распределена по нормальному закону со средним $\mu^* = 167$ и неизвестной дисперсией σ^*.

Вычислим характеристики выборки. Выборочное среднее есть среднее арифметическое всех элементов выборки: $\bar{x} = 172.4$. Это точечная оценка среднего. Теперь нужно вычислить интервальную оценку среднего при неизвестной дисперсии, чтобы определить, попадает ли в этот интервал ожидаемая величина $\mu^* = 167$.

Введем величину

$$z = \frac{(\mu^* - \bar{x}) \cdot \sqrt{n}}{\sigma},$$

распределенную по стандартному нормальному закону $N(0, 1)$.

Если бы дисперсия σ^2 была известна, можно было бы воспользоваться таблицами стандартного нормального распределения и проверить, является ли величина z значимо больше 0. Но поскольку величина дисперсии не известна, надо сначала ее оценить при помощи выборочной дисперсии s^2

$$(\sigma^*)^2 = s^2 = \sum_{i=1}^{10} \frac{(x_i - 172.4)^2}{9} = 62.9,$$

т.е.

$$s = 7.93.$$

Оценка среднеквадратического отклонения для величины $\bar{x}$ есть

$$\frac{s}{\sqrt{n}} = \sqrt{\frac{62.9}{10}} = 2.51.$$

По аналогии с z введем величину t:

$$t = \frac{(\mu^* - \bar{x}) \cdot \sqrt{n}}{s}.$$

Величина t служит критерием проверки, и нам необходимо вычислить ее распределение для $\mu^* = 167$.

Если переписать выражение для t в виде

$$t = \frac{\mu^* - \bar{x}}{\sigma/\sqrt{n}} \Big/ \sqrt{\frac{s^2}{\sigma^2}},$$

то числитель оказывается распределенным по стандартному нормальному закону $N(0,1)$, а квадратный корень из знаменателя имеет также известное распределение, называемое χ^2 *(«хи-квадрат»)* с $(n-1)$ степенями свободы. Более подробно, величина

$$u = \frac{s^2}{\sigma^2} = \sum_{j=1}^{n-1} y_j^2.$$

имеет распределение $\chi^2_{n-1}/(n-1)$, где $Y = \{y_1, y_2, \cdots, y_{n-1}\}$ – распределенная по стандартному нормальному закону случайная величина с независимыми компонентами.

Таким образом, величина t есть функция двух случайных величин, чье распределение известно, а значит, может быть вычислена по известным правилам. Величина t имеет распределение, называемое распределением Стьюдента с числом степеней свободы $(n-1)$ (таким же, как и у соответствующего χ^2-распределения).

В нашем примере доверительная вероятность или надежность оценки есть

$$P\left(|\mu^* - \bar{x}| < \tilde{\epsilon}\right) = P\left(|\mu^* - \bar{x}| < \frac{t_{(n-1),\gamma} \cdot s}{\sqrt{n}}\right) = P\left(\frac{|\mu^* - \bar{x}|}{s/\sqrt{n}} < t_{(n-1),\gamma}\right) = \gamma.$$

Подставляя найденные величины

$$\mu^* = 167.0, \bar{x} = 172.4, s/\sqrt{n} = 2.51, n = 10,$$

получаем

$$P\left(\frac{|167.0 - 172.4|}{2.51} < t_{(9),\gamma}\right) = P\left(172.4 - 2.51 \cdot t_{(9),\gamma} < 167.0 < 172.4 + 2.51 \cdot t_{(9),\gamma}\right) = \gamma.$$

Осталось выбрать надежность γ, вычислить с помощью таблиц $t_{(9),\gamma}$ и проверить выполнение неравенства.

Согласно стандартным рекомендациям, зададимся доверительной вероятностью $\gamma = 0.9$ (или уровнем значимости $\alpha/2 = 5\%$). Табличное значение $t_{(9),5\%} = 1.8331$, т.е. с вероятностью 0.9 должно быть $|t_{(9),5\%}| \leq 1.8331$. Тогда предполагаемое среднее значение μ^* должно с вероятностью 0.9 лежать в интервале $\{167.8, 177.0\}$. Но в нашем случае это не так. Можно прийти к тому же выводу, сравнив табличное значение $t_{(n-1),\gamma}$ с вычисленной статистикой t:

$$\frac{|167.0 - 172.4|}{2.51} = 2.15 > 1.8331.$$

Следовательно, идея принять средний рост 167 см для данной выборки оказалась неудачной.

Можно было бы выбрать и другую доверительную вероятность.

Однако надо иметь в виду, что чем больше доверительная вероятность, тем меньше уровень значимости и, следовательно, тем менее точен результат. Так, к примеру, для уровня значимости 0.05% доверительный интервал станет очень большим (в нашем случае он станет $\{160.4, 184.4\}$) и, хотя он и покроет значение 167.0, никакой практической ценности иметь не будет. Наоборот, чем меньше доверительная вероятность, тем выше уровень значимости и тем *уже* доверительный интервал.

6.3 Оценка дисперсии

6.3.1 Точечная оценка дисперсии

Точечная оценка дисперсии, вычисленная по выборке:

$$(\sigma^*)^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \mu^*)^2 = \frac{1}{n-1} \cdot \left(\sum_{i=1}^{n} x_i^2 - n \cdot (\mu^*)^2 \right),$$

где μ^* тоже оценивается по выборке своим средним арифметическим:

$$\mu^* = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Для больших объемов выборки ($n > 30$) можно пользоваться точечной оценкой дисперсии (получаемой, например, методом максимального правдоподобия аналогично вычислению оценки для математического ожидания):

$$(\sigma^*)^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \mu^*)^2 = \frac{1}{n} \cdot \left(\sum_{i=1}^{n} x_i^2 - n \cdot (\mu^*)^2 \right),$$

Существует большое количество точечных оценок для дисперсии и среднеквадратического отклонения (см., например, [7]), обладающих разными точностями и полезные в прикладных задачах для быстрых расчетов. Приведем, для примера, простую линейную оценку Даутона для средеквадратического отклонение (которая при малых выборках $n \leq 10$ дает 94% эффективности по сравнению с оценкой максимального правдоподобия):

$$\sigma^* = \frac{1.77245}{n \cdot (n-1)} \sum_{i=1}^{n} x_i \cdot \left(2 \cdot i - n - 1 \right).$$

6.3.2 Интервальная оценка дисперсии

Пусть X_i – независимые нормально распределенные случайные величины *с известным математическим ожиданием* и неизвестной истиной дисперсией, [2]: $M[X_i] = \mu_x$, $D[X_i] = \sigma_x^2$.

В качестве точечной оценки дисперсии случайного параметра X примем статистику

$$(\sigma^*)^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \mu_x)^2.$$

Для построения *интервальной оценки дисперсии* используется *статистика* χ^2:

$$\chi^2 = \frac{n \cdot (\sigma^*)^2}{(\sigma_x)^2} = \sum_{i=1}^{n} \left(\frac{x_i - \mu_x}{\sigma_x} \right)^2.$$

Величина

$$u_i = \frac{x_i - \mu_x}{\sigma_x} \sim N\left(0, 1\right).$$

Таким образом, сумма квадратов случайных величин, распределенных по стандартному нормальному закону, и обладает распределением «хи-квадрат» с n степенями свободы

(т.е., с n независимыми элементами, составляющими эту величину). Плотность вероятности χ^2-распределения есть

$$f(x) = \left(2^{n/2} \cdot \Gamma(n/2)\right)^{-1} \cdot x^{n/2-1} \cdot \exp\{-x/2\},$$

$$X \sim \chi^2(n), 0 < x < +\infty.$$

Вероятность попадания случайной величины $X \sim \chi^2(n)$ в интервал $[x_1, x_2]$ есть

$$P(X \in [\alpha, \beta]) = \int_\alpha^\beta z(x)dx.$$

Поскольку плотность вероятности $f(x)$ распределения χ^2 не является симметричной относительно оси ординат (в отличие от симметричной плотности вероятности стандартного нормального закона), то доверительный интервал для оценки дисперсии J_{σ^*} построим так, чтобы вероятность попадания случайной величины слева и справа от концов отрезка $[\alpha, \beta]$ была одинаковой и равной $(1-\gamma)/2$, где γ, как и раньше, доверительная вероятность или надежность оценки (в данном случае, надежность оценки дисперсии):

$$P(x < \alpha) = P(x > \beta) = \frac{1-\gamma}{2}.$$

Точки α и β, которые есть

$$\alpha = \chi^2(n, p_\alpha),$$
$$\beta = \chi^2(n, p_\beta),$$

определяются по таблице χ^2-распределения

$$P(\chi^2 > \beta) = \int_\beta^{+\infty} f(x)dx = \frac{1-\gamma}{2} = p_\beta,$$

$$P(\chi^2 < \alpha) = 1 - \int_\alpha^{+\infty} f(x)dx = \frac{1-\gamma}{2} = 1 - p_\alpha.$$

Доверительный интервал для дисперсии есть

$$J_{\sigma^*} = \left\{\frac{n \cdot (\sigma^*)^2}{\beta}, \frac{n \cdot (\sigma^*)^2}{\alpha}\right\},$$

где

$$(\sigma^*)^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \mu_x)^2$$

есть точечная оценка дисперсии, а μ_x априори известно. Этот доверительный интервал накрывает неизвестную искомую дисперсию с заданной доверительной вероятностью γ. Точность оценки дисперсии ϵ есть

$$\epsilon = \frac{1}{2} \cdot \left(\frac{n \cdot (\sigma^*)^2}{\alpha} - \frac{n \cdot (\sigma^*)^2}{\beta}\right) = \frac{n \cdot (\sigma^*)^2}{2} \cdot \frac{\beta - \alpha}{\beta \cdot \alpha}.$$

ПРИМЕР Производилась оценка дисперсии случайного параметра X по результатам 20 испытаний, [2]. Результат статистической обработки значений $x_1, x_2, \cdots, x_{20}$ оказался равным $(\sigma^*)^2 = 16$, и среднее значение X было априори известно. Определить

интервальную оценку истинного неизвестного значения дисперсии σ^2 при заданной надежности $\gamma = 0.95$.

Для определения границ α и β для вычисления доверительного интервала χ^2, определим вероятности p_α, p_β:

$$p_\alpha = \frac{1+\gamma}{2} = 0.975,$$
$$p_\beta = \frac{1-\gamma}{2} = 0.025.$$

По таблице χ^2-распределения найдем

$$\alpha = \chi^2(20, 0.975) = 9.591,$$
$$\beta = \chi^2(20, 0.025) = 34.170.$$

В таблице первый левый столбец – число степеней свободы n, верхняя строка содержит величину границы интервала, в процентах, в порядке убывания (которые вычисляются по заданной надежности или, что то же самое, доверительной вероятности γ).

Теперь вычислим границы доверительного интервала для оценки дисперсии:

$$J_{\sigma^*} = \left\{ \frac{20 \cdot 16}{34.170}, \frac{20 \cdot 16}{9.59} \right\} = \{9.36, 33.37\}$$

Истинное (неизвестное) значение дисперсии σ^2 с вероятностью 95% накрывается отрезком $\{9.36, 33.37\}$. Доверительный интервал не является симметричным относительно оценки $(\sigma^*)^2 = 16$. Поэтому точность оценки дисперсии ϵ можно определить только приближенно по формуле

$$\epsilon = \frac{1}{2} \cdot \left(\frac{n \cdot (\sigma^*)^2}{\alpha} - \frac{n \cdot (\sigma^*)^2}{\beta} \right) = \frac{33.37 - 9.36}{2} = 12.0.$$

Рассмотрим теперь, как измениться интервальная оценка дисперсии, если *математическое ожидание априори не известно.*

Если математическое ожидание случайной величины X неизвестно, то его, как и дисперсию, нужно определять по имеющейся выборке:

$$\mu^* = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

В этом случае за точечную оценку дисперсии принимается

$$(\sigma^*)^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \mu^*)^2.$$

Для построения интервальной оценки дисперсии построим статистику χ^2:

$$\chi^2 = \frac{(n-1) \cdot (\sigma^*)^2}{\sigma^2} = \sum_{i=1}^{n-1} \left(\frac{x_i - \mu^*}{\sigma_x} \right)^2,$$

где, как и в предыдущем случае, σ_x^2 – неизвестное истинное значение дисперсии. Число степеней свободы такой статистики $(n-1)$. Интервальная оценка дисперсии при неизвестном математическом ожидании строится аналогично рассмотренной выше, с тем отличием, что концы доверительного интервала $\tilde{\alpha}$, $\tilde{\beta}$ определяются теперь как

$$\tilde{\alpha} = \chi^2(n-1, p_\alpha); \tilde{\beta} = \chi^2(n-1, p_\beta),$$

а концы доверительного интервала вычисляются по формулам:

$$J_{\sigma^*} = \left\{ \frac{(n-1)\cdot(\sigma^*)^2}{\tilde{\beta}}, \frac{(n-1)\cdot(\sigma^*)^2}{\tilde{\alpha}} \right\}$$

Точность оценки дисперсии при неизвестном математическом ожидании вычисляется как

$$\epsilon = \frac{(n-1)\cdot(\sigma^*)^2}{2} \cdot \frac{\beta - \alpha}{\beta \cdot \alpha}.$$

Доверительный интервал для дисперсии, получаемый в условиях незнания математического ожидания, получается шире, чем доверительный интервал для дисперсии при известном математическом ожидании. Точность в первом случае хуже, чем во втором.

6.4 Сравнение дисперсий двух выборок нормальной генеральной совокупности

Предположим, имеется две нормально распределенные случайные выборки $X = \{x_1, x_2, \cdots, x_n\}$ и $Y = \{y_1, y_2, \cdots, y_m\}$. Пусть $X \sim N(\mu_x, \sigma_x^2)$ и $Y \sim N(\mu_y, \sigma_y^2)$.

Для проверки того, можно ли на некотором уровне точности считать равными дисперсии двух выборок, $\sigma_x^2 = \sigma_y^2$, используется т.н. *F-статистика Фишера* (или *Фишера-Снедекора*):

$$\tilde{F}(n-1, m-1) = \frac{s_1^2}{s_2^2} = \frac{\chi^2(n-1)/(n-1)}{\chi^2(m-1)/(m-1)},$$

Отношение оценок дисперсий всегда берется большая к меньшей, чтобы отношение было больше единицы. Если ставится задача проверить, равны или нет дисперсии двух выборок, то нужно вычислить по выборке величину $\tilde{F}(n-1, m-1)$ и проверить, попадает ли она в критический интервал, который определяется по функции $F(n-1, m-1)$ из таблицы Фишера, где размеры выборок определяют число степеней свободы этой табличной функции:

$$\left\{ F_1\left(\frac{\alpha}{2}; n-1, m-1\right), F_2\left(\frac{\alpha}{2}; n-1, m-1\right) \right\},$$

причем

$$F_1\left(\frac{\alpha}{2}; n-1, m-1\right) = \frac{1}{F\left(\frac{\alpha}{2}; m-1, n-1\right)},$$

$$F_2\left(\frac{\alpha}{2}; n-1, m-1\right) = F\left(\frac{\alpha}{2}; n-1, m-1\right).$$

Другими словами, проверяется

$$F_1 < \tilde{F} < F_2.$$

Если это не выполняется, то дисперсии признаются разными.

Рассмотрим пример на сравнение двух дисперсий и работу с таблицей распределения Фишера.

ПРИМЕР Проводится сравнение точности работы двух типов высотомеров, [2].

В ходе проверок фиксировались отклонения показаний высотометров от точного значения высоты. Результаты отклонений приведены в Таблице (11). Требуется сравнить дисперсии двух выборок (т.е. сравнить точности двух высотомеров). Доверительный интервал $\gamma = 90\%$ (или, что то же самое, уровень значимости $\alpha = 0.10$).

Таблица 11:

Точность работы двух высотомеров

Номер наблюдения, i	Отклонение высотомера $No.1$ (в метрах)	Отклонение высотомера $No.2$ (в метрах)
1	-8	-20
2	-14	-10
3	0	-3
4	14	11
5	-38	-4
6	2	12
7	50	-3
8	1	17
9	10	42
10	15	
11	0	
12	22	

Вычислим точечные оценки математических ожиданий и дисперсий величин $X = \{x_1, x_2, \cdots, x_{12}\}$ и $Y = \{y_1, y_2, \cdots, y_9\}$:

$$\mu_x^* = \frac{1}{12}\sum_{i=1}^{12} x_i = 4.5,$$

$$\mu_y^* = \frac{1}{9}\sum_{i=1}^{9} y_i = 4.7,$$

$$(\sigma_x^*)^2 = \frac{1}{12-1}\sum_{i=1}^{12}(x_i - \mu_x)^2 = 451.9,$$

$$(\sigma_y^*)^2 = \frac{1}{9-1}\sum_{i=1}^{9}(y_i - \mu_y)^2 = 331.7,$$

Значение статистики Фишера есть отношение большей оценки дисперсии к меньшей:

$$\tilde{F} = \frac{451.9}{331.7} = 1.36.$$

Число степеней свободы табличной статистики есть $\hat{n} = 12 - 1 = 11$, $\hat{m} = 9 - 1 = 8$.

$$F_2(0.05; 11, 8) = F(0.05; 11, 8) = 3.31,$$

$$F_1(0.05; 11, 8) = \frac{1}{F(0.05; 8, 11)} = \frac{1}{2.95} = 0.34.$$

Окончательно получаем верное неравенство

$$0.34 < 1.36 < 3.31,$$

следовательно, дисперсии двух выборок равны с доверительной вероятность 90%.

7 Перенос ошибок

Часто результат эксперимента представляет собой некоторую функцию от нескольких различных случайных величин X_r. Предположим, что каждое наблюдаемое значение x_r принадлежит генеральной совокупности со средним μ_r и дисперсией σ_r^2. Теория переноса ошибок позволяет определить значение среднеквадратического отклонения, которое следует приписать величине $y = y(x_1, x_2, \cdots, x_m)$.

Для набора m случайных величин X_r $(r = 1, 2, \cdots, m)$ *матрица ошибок* $M_E(x)$ определяется следующим образом:

$$M_E(x) = \begin{pmatrix} M[(x_1-\mu_1)^2] & M[(x_1-\mu_1)(x_2-\mu_2)] & \cdots & M[(x_1-\mu_1)(x_r-\mu_r)] \\ M[(x_2-\mu_2)(x_1-\mu_1)] & M[(x_2-\mu_2)^2] & \cdots & M[(x_2-\mu_2)(x_r-\mu_r)] \\ \vdots & \vdots & \ddots & \vdots \\ M[(x_r-\mu_r)(x_1-\mu_1)] & M[(x_r-\mu_r)(x_2-\mu_2)] & \cdots & M[(x_r-\mu_r)^2] \end{pmatrix}$$

Матрица $M_E(x)$ – симметричная, а ее диагональные элементы есть дисперсии соответствующих величин:

$$[M_E(x)]_{rr} = \sigma_r^2.$$

Если величины x_r не коррелированные, то матрица ошибок диагональная.

Пусть теперь величины y_r $(r = 1, 2, \cdots, m)$ есть линейные функции переменных $x_s (s = 1, 2, \cdots, n)$:

$$y_r = a_{r1}x_1 + a_{r2}x_2 + \cdots + a_{rn}x_n = \sum_{s=1}^{n} a_{rs}x_s$$

или

$$y = Ax.$$

где a_{rs} есть постоянные неслучайные величины, элементы матрицы A.

Если произведено несколько измерений каждой величины x_s, то оценка величины y_r получается при замене x_s на их среднее $\bar{x}_s$. Если $M_E(x)$ – матрица ошибок величин x_s, тогда матрица ошибок для функций y_r будет

$$M_E(y) = A \cdot M_E(x) \cdot A^T.$$

Докажем последнее равенство

- Пусть, для простоты записи, $\mu_1 = \cdots = \mu_r = \mu_x$, то есть математическое ожидание всех X_r одинаковое. Рассмотрим для простоты случай двух случайных величин x_1 и x_2, и две функции y_1 и y_2:

  $$y_1 = a_{11}x_1 + a_{12}x_2,$$

  $$y_2 = a_{21}x_1 + a_{22}x_2.$$

  Матрица

  $$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

  Матрица ошибок есть

  $$M_E(x) = \begin{pmatrix} M[(x_1-\mu_x)^2] & M[(x_1-\mu_x)(x_2-\mu_x)] \\ M[(x_2-\mu_x)(x_1-\mu_x)] & M[(x_2-\mu_x)^2] \end{pmatrix}.$$

Тогда

$$B = A \cdot M_E(x) \cdot A^T =$$

$$= \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} M[(x_1-\mu_x)^2] & M[(x_1-\mu_x)(x_2-\mu_x)] \\ M[(x_2-\mu_x)(x_1-\mu_x)] & M[(x_2-\mu_x)^2] \end{pmatrix} \begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{pmatrix}.$$

Распишем первый элемент полученной квадратной матрицы:

$$b_{11} = \Big(a_{11}M[(x_1-\mu_x)^2] + a_{12}M[(x_1-\mu_x)(x_2-\mu_x)] \Big) \cdot a_{11} +$$

$$+ \Big(a_{11}M[(x_1-\mu_x)(x_2-\mu_x)] + a_{12}M[(x_2-\mu_x)^2] \Big) \cdot a_{12} =$$

$$= (a_{11})^2 M[(x_1-\mu_x)^2] + (a_{12})^2 M[(x_2-\mu_x)^2] + 2a_{11}a_{12}M[(x_1-\mu_x)(x_2-\mu_x)].$$

Теперь для матрицы

$$M_E(y) = \begin{pmatrix} M[(y_1-\mu_{y1})^2] & M[(y_1-\mu_{y1})(y_2-\mu_{y2})] \\ M[(y_2-\mu_{y2})(y_1-\mu_{y1})] & M[(y_2-\mu_{y2})^2] \end{pmatrix},$$

где

$$\mu_{y1} = (a_{11}+a_{12})\mu_x,$$
$$\mu_{y2} = (a_{21}+a_{22})\mu_x,$$

распишем первый элемент:

$$M_E(y)_{11} = M[(a_{11}x_1 + a_{12}x_2 - (a_{11}+a_{12})\mu_x)^2] = M[(a_{11}(x_1-\mu_x) + a_{12}(x_2-\mu_x))^2] =$$

$$= (a_{11})^2 M[(x_1-\mu_x)^2] + (a_{12})^2 M[(x_2-\mu_x)^2] + 2a_{11}a_{12}M[(x_1-\mu_x)(x_2-\mu_x)] = b_{11}.$$

Аналогично поэлементно доказывается равенство матрицы B и $M_E(y)$.

Полученный результат можно легко обобщить на случай нелинейных функций $y_r = f_r(x_1, \cdots, x_s, \cdots, x_n)$, предположив, что функция f_r мало меняется в области, ограниченной среднеквадратическим отклонением от ее среднего значения, или, другими словами, что с точностью до членов первого порядка ее разложение в ряд Тэйлора имеет вид

$$y_r = f_r(x_1, \cdots, x_n) = f_r(\bar{x}_1, \cdots, \bar{x}_n) + \sum_{s=1}^{n} \left\{ \left(x_s - \bar{x}_s \right) \frac{\partial f_r}{\partial x_s}\bigg|_{\bar{x}} \right\}.$$

В этом разложении $\bar{x}_s$ есть среднее значение x_s, а $x = \{\bar{x}_1, \cdots, \bar{x}_n\}$. Тогда оценка величины y_r есть

$$y_r^* = f_r(\bar{x}_1, \cdots, \bar{x}_n),$$

а элементами матрицы ошибок являются

$$[M_E(y)]_{rs} = \sum_{i=1}^{n}\sum_{j=1}^{n} \left\{ (x_i - \bar{x}_i)(x_j - \bar{x}_j) \frac{\partial f_r}{\partial x_i}\bigg|_{\bar{x}} \frac{\partial f_s}{\partial x_j}\bigg|_{\bar{x}} \right\}.$$

Можно доказать, по аналогии с линейным случаем, что

$$M_E(y) = F \cdot M_E(x) \cdot F^T,$$

где F – матрица, элементы которой равны

$$[F]_{rs} = \frac{\partial f_r}{\partial x_s}\bigg|_{\bar{x}}.$$

7.0.1 Отношение двух случайных величин

Для примера применения полученных формул рассмотрим одномерный случай *отношения двух случайных величин* x_1 и x_2.

Среднее значение этого отношения есть

$$\bar{y} \approx \frac{\bar{x}_1}{\bar{x}_2}.$$

Матрица ошибок для $\bar{x}_1$ и $\bar{x}_2$ есть

$$M_E(x) = \begin{pmatrix} s_1^2 & s_1 s_2 q_{12} \\ s_1 s_2 q_{12} & s_2^2 \end{pmatrix},$$

где q_{12} – оценка коэффициента корреляции x_1 и x_2.

Матрица ошибок F одномерной функции $\bar{y}$ есть, с одной стороны, $s_{\bar{y}}^2$, а с другой стороны – произведение трех матриц (в силу одномерности функции y матрица F есть строка, а матрица F^T есть столбец):

$$M_E(y) = \begin{pmatrix} \frac{1}{\bar{x}_2} & -\frac{\bar{x}_1}{\bar{x}_2^2} \end{pmatrix} \begin{pmatrix} s_1^2 & s_1 s_2 q_{12} \\ s_1 s_2 q_{12} & s_2^2 \end{pmatrix} \begin{pmatrix} \frac{1}{\bar{x}_2} \\ -\frac{\bar{x}_1}{\bar{x}_2^2} \end{pmatrix}$$

После перемножения получаем ошибку величины $\bar{y}$:

$$s_{\bar{y}}^2 = \frac{\bar{x}_1^2}{\bar{x}_2^2} \left\{ \frac{s_1^2}{\bar{x}_1^2} + \frac{s_2^2}{\bar{x}_2^2} - 2 q_{12} \frac{s_1}{\bar{x}_1} \frac{s_2}{\bar{x}_2} \right\}.$$

7.0.2 Произведение двух случайных величин

Из рассуждений, полностью аналогичных предыдущим, среднее значение произведения двух случайных величин x_1 и x_2 есть

$$\bar{y} \approx \bar{x}_1 \bar{x}_2,$$

а ошибка величины $\bar{y}$ есть

$$s_{\bar{y}}^2 \approx \bar{x}_1^2 \bar{x}_2^2 \left\{ \frac{s_1^2}{\bar{x}_1^2} + \frac{s_2^2}{\bar{x}_2^2} + 2 q_{12} \frac{s_1}{\bar{x}_1} \frac{s_2}{\bar{x}_2} \right\}.$$

Важно отметить, что если x_1 и x_2 – независимые нормально распределенные величины, то строгим выражением для дисперсии их произведения является

$$s_{\bar{y}}^2 = s_1^2 \bar{x}_2^2 + s_2^2 \bar{x}_1^2 + s_1^2 s_2^2$$

и это выражение может сильно отличаться от предыдущего приближенного выражения при $q_{12} = 0$, если ошибки $\bar{x}_1^2$ и $\bar{x}_2^2$ велики.

7.0.3 Дисперсия произвольной функции от n независимых случайных величин

Еще одним частным случаем вычисления матрицы ошибок является вычисление дисперсии функции многих независимых случайных переменных. Если $y = f(x_i), i = 1, \cdots, n$, то

$$s_{\bar{y}}^2 = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i} \right)^2 \Bigg|_{\bar{x}_i} s^2(\bar{x}_i).$$

Существует еще один способ определения характеристик функции случайных величин, $y = y(x_1, x_2, \cdots, x_n)$, но он основан на редко реализуемом в практических задачах предположении, что нам точно известны плотности распределения случайных аргументов этой функции, [8].

Рассмотрим эту задачу более подробно на примере функции двух случайных аргументов и найдем, для примера, функцию распределения суммы $x_1 + x_2$. Поскольку x_1, x_2 – независимы, то вероятность того, что x_1 лежит в интервале $[a_1, b_1]$, а x_2 лежит в интервале $[a_2, b_2]$, равна произведению

$$\int_{a_1}^{b_1} f_1(u)du \cdot \int_{a_2}^{b_2} f_2(v)dv = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f_1(u)f_2(v)dudv.$$

Следовательно, пара случайных величин (x_1, x_2) имеет плотность распределения $f(u, v) = f_1(u)f_2(v)$.

Значение функции распределения $H(t)$ случайной величины $x_1 + x_2$ в точке t равно вероятности того, что $x_1 + x_2 < t$:

$$H(t) = P(x_1 + x_2 < t).$$

Имеет место *ТЕОРЕМА*, [8]:

- Если совокупность случайных величин $\{x_1, x_2, \cdots x_n\}$ обладает плотностью вероятности $f(x_1, x_2, \cdots, x_n)$, то вероятность попадания случайной точки X с координатами $\{x_1, x_2, \cdots x_n\}$ в произвольную область G равна интегралу от функции f по этой области:

$$P(X \in G) = \int_G f(u_1, u_2, \cdots, u_n)du_1du_2 \cdots du_n.$$

Используя эту теорему, видим, что

$$H(t) = \iint\limits_{u+v<t} f_1(u)f_2(v)dudv = \int_{-\infty}^{+\infty} du \int_{-\infty}^{t-u} f_1(u)f_2(v)dv = \int_{-\infty}^{+\infty} du \int_{-\infty}^{t} f_1(u)f_2(w-u)dw,$$

где введена новая переменная интегрирования $w = u + v$.

Поскольку плотности вероятностей неотрицательны, можно изменить порядок интегрирования:

$$H(t) = \int_{-\infty}^{t} dw \int_{-\infty}^{+\infty} f_1(u)f_2(w-u)du.$$

Эта функция распределения обладает плотностью вероятности, которая и есть плотность вероятности суммы двух независимых случайных величин с плотностями $f_1(t), f_2(t)$:

$$h(t) = \int_{-\infty}^{+\infty} f_1(u)f_2(t-u)du.$$

Зная плотность вероятности суммы двух случайных величин, можно вычислить все характеристики этой суммы: среднее, дисперсию и любые другие интересующие моменты высших порядков.

8 Элементы линейной алгебры

Приведем некоторые сведения из линейной алгебры, необходимые для решения систем линейных уравнений.

8.1 Система линейных уравнений

В общем случае система m линейных уравнений с n неизвестными (или кратко, *линейная система*) (далее просто «система») имеет вид, [9] – [10]:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \cdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \end{cases}$$

Величины $x_1, x_2, \cdots, x_n$ – неизвестные, которые нужно вычислить, решив матричное уравнение. Заданная матрица системы (или *основная матрица*) есть

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

Заданный вектор-столбец свободных членов есть $(b_1, b_2, \cdots, b_m)$.

Если все $b_j = 0$, то система называется *однородной*. Если хотя бы одно значение $b_j \neq 0$, то система называется *неоднородной*. Система называется *квадратной*, если $n = m$.

Решением системы называется совокупность n чисел $c_1, c_2, \cdots, c_n$, которые при подстановки в систему на место неизвестных $x_1, x_2, \cdots, x_n$ обращает все уравнения этой системы в тождества.

Не всякая система имеет решение. Например,

$$\begin{cases} x_1 + x_2 = 1 \\ x_1 + x_2 = 2 \end{cases}$$

решения не имеет.

Система называется *совместной*, если она имеет хотя бы одно решение, и *несовместной*, если нет ни одного решения. Совместная система называется *определенной*, если у нее есть единственное решение и *неопределенной*, если у нее есть хотя бы два разных решения.

Условие наличие у системы хотя бы одного решения формулируется *ТЕОРЕМОЙ КРОНЕКЕРА-КАПЕЛЛИ*

8.1.1 Теорема Кронекера-Капелли

Для того, чтобы линейная система

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \cdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \end{cases}$$

являлась совместной (т.е., имела хотя бы одно решение), необходимо и достаточно, чтобы ранг расширенной матрицы этой системы был равен рангу ее основной матрицы.

Расширенная матрица системы есть:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & b_m \end{pmatrix}.$$

Матрица обладает *рангом* r, если у матрицы A есть минор порядка r, не равный нулю, а всякий минор порядка $r+1$ равен нулю. *Минор порядка* r – это определитель r-го порядка с элементами, лежащими на пересечении любых k строк и любых k столбцов матрицы A.

8.2 Метод Крамера решения системы линейных уравнений

Решение системы линейных уравнений существует и единственно, когда определитель матрицы системы не равен нулю. В этом случае единственное решение ищется *по формулам Крамера*:

$$x_i = \frac{\Delta_i}{\Delta},$$

где Δ_i – определитель матрицы, получающейся из основной матрицы системы заменой j-го столбца на столбец свободных членов, Δ – определитель основной матрицы системы.

Разберем метод Крамера на примере.

ПРИМЕР

- Найти решение квадратной системы линейных уравнений:

$$\begin{cases} x_1 + 2x_2 + 3x_3 + 4x_4 = 30 \\ -x_1 + 2x_2 - 3x_3 + 4x_4 = 10 \\ x_2 - x_3 + x_4 = 3 \\ x_1 + x_2 + x_3 + x_4 = 10 \end{cases}$$

Матрица системы (или основная матрица) есть

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ -1 & 2 & -3 & 4 \\ 0 & 1 & -1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

Распишем подробно вычисление определителя Δ_1 по формуле расчета с помощью алгебраических дополнений. Напомним, *алгебраическое дополнение* для элемента матрицы A a_{ij} – это число $\Delta_{ij} = (-1)^{i+j} M_{ij}$, где M_{ij} – минор, определитель матрицы, получаемой вычеркиванием из матрицы A i-ой строки и j-го столбца. Для вычисления определителя матрицы A с помощью алгебраических дополнений используется метод разложения по строке (или столбцу) по формуле:

$$\Delta = detA = \sum_{j=1}^{n} a_{ij}\Delta_{ij}.$$

используем это разложение для вычисления Δ_1:

$$\Delta_1 = \begin{vmatrix} 30 & 2 & 3 & 4 \\ 10 & 2 & -3 & 4 \\ 3 & 1 & -1 & 1 \\ 10 & 1 & 1 & 1 \end{vmatrix} = 30 \cdot \begin{vmatrix} 2 & -3 & 4 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \end{vmatrix} - 2 \cdot \begin{vmatrix} 10 & -3 & 4 \\ 3 & -1 & 1 \\ 10 & 1 & 1 \end{vmatrix} + 3 \cdot \begin{vmatrix} 10 & 2 & 4 \\ 3 & 1 & 1 \\ 10 & 1 & 1 \end{vmatrix} - 4 \cdot \begin{vmatrix} 10 & 2 & -3 \\ 3 & 1 & -1 \\ 10 & 1 & 1 \end{vmatrix} =$$

$$= 30 \cdot [2 \cdot (-2) + 3 \cdot 0 + 4 \cdot 2] - 2 \cdot [10 \cdot (-2) + 3 \cdot (-7) + 4 \cdot 13] + 3 \cdot [10 \cdot 0 - 2 \cdot (-7) + 4 \cdot (-7)] -$$
$$-4 \cdot [10 \cdot 2 - 2 \cdot 13 - 3 \cdot (-7)] = -4.$$

Определитель основной матрицы:

$$\Delta = \begin{vmatrix} 1 & 2 & 3 & 4 \\ -1 & 2 & -3 & 4 \\ 0 & 1 & -1 & 1 \\ 1 & 1 & 1 & 1 \end{vmatrix} = -4.$$

Определители $\Delta_2, \Delta_3, \Delta_4$:

$$\Delta_2 = \begin{vmatrix} 1 & 30 & 3 & 4 \\ -1 & 10 & -3 & 4 \\ 0 & 3 & -1 & 1 \\ 1 & 10 & 1 & 1 \end{vmatrix} = -8, \Delta_3 = \begin{vmatrix} 1 & 2 & 30 & 4 \\ -1 & 2 & 10 & 4 \\ 0 & 1 & 3 & 1 \\ 1 & 1 & 10 & 1 \end{vmatrix} = -12,$$

$$\Delta_4 = \begin{vmatrix} 1 & 2 & 3 & 30 \\ -1 & 2 & -3 & 10 \\ 0 & 1 & -1 & 3 \\ 1 & 1 & 1 & 10 \end{vmatrix} = -16.$$

Тогда $x_1 = \Delta_1/\Delta = 1, x_2 = \Delta_2/\Delta = 2, x_3 = \Delta_3/\Delta = 3, x_4 = \Delta_4/\Delta = 4$.

8.3 Метод Гаусса решения системы линейных уравнений

Помимо метода Крамера, существует другой метод решения системы линейных уравнений, называемый *методом Гаусса*. Основная идея метода в том, что со строками и столбцами матрицы можно производить три *элементарные операции*, которые не изменяют ранга матрицы:

1. перестановку двух строк (столбцов),
2. умножение строки (столбца) на любой, отличный от нуля, множитель,
3. прибавление к одной строке (столбцу) произвольной линейной комбинации других строк (столбцов).

Любую матрицу можно привести к диагональному виду с помощью этих трех элементарных операций согласно следующему алгоритму.

1. Перестановкой строк (столбцов) сделать $a_{11} \neq 0$ (если это необходимо).
2. Умножить первую строку на a_{11}^{-1}.
3. Вычесть из j-го столбца первый столбец, умноженный на a_{1j}.

4. Вычесть из i-ой строки первую строку, умноженную на a_{i1}; получится матрица вида:

$$\tilde{A} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \tilde{a}_{22} & \cdots & \tilde{a}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \tilde{a}_{m2} & \cdots & \tilde{a}_{mn} \end{pmatrix}.$$

5. Все предыдущие шаги осуществляем с матрицей

$$\begin{pmatrix} \tilde{a}_{22} & \cdots & \tilde{a}_{2n} \\ \vdots & \ddots & \vdots \\ \tilde{a}_{m2} & \cdots & \tilde{a}_{mn} \end{pmatrix}.$$

и т.д. до получения диагональной матрицы.

Как метод Гаусса, так и метод Крамера могут привести к большим погрешностям, если значения коэффициентов и свободных членов заданы приближенно или когда производится округление в процессе вычисления. В первую очередь это относится к случаю, когда основная матрица линейной системы является *плохо обусловленной* (см. (13.4.1)), т.е., когда малым изменениям элементов этой матрицы отвечают большие изменения элементов обратной матрицы. В таком случае решение линейной системы $x = A^{-1}b$ окажется *неустойчивым*. Для решения неустойчивых линейных систем существуют методы *регуляризации А.Н. Тихонова*, *итерационные методы Якоби* и методы *сингулярного матричного разложения*, [10].

9 Понятие о равноточных и неравноточных измерениях

Равноточность результатов (наблюдательных или экспериментальных выборочных данных) означает, что все эти результаты $x_1, x_2, \cdots, x_n$ получены с одинаковой точностью. Если все x_i равноточны, то их среднеквадратические отклонения равны: $s_i = s$ (s^2 – дисперсия, оцененная по выборке).

Напомним основные характеристики равноточной выборки и обобщим их для случая неравноточных измерений.

- *Наиболее вероятное значение определяемой величины есть среднее арифметическое всех элементов выборки*:
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

- *Наиболее вероятное значение средней квадратичной ошибки одного измерения*:
$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}.$$

- *Средняя квадратичная ошибка среднего арифметического*:
$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n(n-1)}}.$$

В реальных задачах часто бывают случаи, когда для как можно более надежного определения какой-то величины собирают измерения разного происхождения, т.е. выполненные на разных приборах, при разных условиях, разными методами и т.д. такие измерения носят название *неравноточных*.

Простейший случай неравноточных измерений – собрание не прямых измерений, а выводов из равноточных измерений, число которых различно в различных выводах.

Другими словами, пусть имеет m_1 равноточных измерений и из них выведено наиболее вероятное значение x_1. Далее, из m_2 равноточных измерений выводится наиболее вероятное значение x_2 и т.д. Получается набор наиболее вероятных величин $\{x_1, x_2, \cdots, x_n\}$ и из этого набора нужно вывести наиболее вероятное значение x_k.

Для каждого x_k среднеквадратическое отклонение есть, по построению
$$s_k = \frac{s}{\sqrt{m_k}},$$
где s – среднеквадратическая ошибка одного измерения.

Если x_k – равноточные, то $s_k = s$. Если x_k – неравноточные, то вместо $s_1, s_2, \cdots, s_n$ вводят числа $p_1, p_2, \cdots, p_n$, называемые *весами измерений*:
$$p_k = \frac{s_0^2}{s_k^2},$$
где s_0^2 – любое положительное число (s_0 называется среднеквадратическое отклонение на единицу веса).

Из определения введенных весов следуют их свойства.

- Веса неравноточных измерений обратно пропорциональны своим дисперсиям.
- Веса неравноточных измерений – относительные числа.

Основные характеристики для случая неравноточных измерений.

- *Наиболее вероятное значение определяемой величины есть среднее весовое или среднее взвешенное всех элементов выборки*:

$$\bar{x}_p = \frac{1}{p}\sum_{k=1}^{n} p_k x_k,$$

где

$$p = \sum_{k=1}^{n} p_k.$$

- *Наиболее вероятное значение средней квадратичной ошибки измерения с весом единица*:

$$s_p = \sqrt{\frac{\sum_{k=1}^{n} p_k (x_k - \bar{x}_p)^2}{n-1}}.$$

- *Средняя квадратичная ошибка среднего весового или среднего взвешенного*:

$$s_{\bar{x}_p} = \frac{s_0}{\sqrt{p}},$$

где

$$p = \sum_{k=1}^{n} p_k.$$

9.1 Условные и нормальные уравнения

В реальных задачах часто бывает так, что подлежащие определению величины нельзя наблюдать непосредственно. Вместо них из наблюдений можно определить только функции неизвестных, [1].

Рассмотрим пример.

Пусть наблюдения дают значения x_k и y_k величин x и y соответственно. Предполагается, что x и y связаны зависимостью:

$$y = \Theta_0 + \Theta_1 x + \Theta_2 x^2,$$

где $\Theta_0, \Theta_1, \Theta_2$ – подлежащие определению коэффициенты. Каждое наблюдение (x_k, y_k) дает уравнение с тремя неизвестными:

$$y_k = \Theta_0 + \Theta_1 x_k + \Theta_2 x_k^2, k = 1, 2, \cdots, n$$

В общем виде задача ставится так: вместо подлежащих определению величин $\Theta_0, \Theta_1, \Theta_2, \cdots$ из наблюдений получаются величины y_k, которые есть функции от неизвестных $\Theta_0, \Theta_1, \Theta_2, \cdots$. Каждое наблюдение дает *условное уравнение* вида

$$f_k(\Theta_0, \Theta_1, \Theta_2, \cdots, x_k) = y_k.$$

Если бы в процессе наблюдений не было случайных ошибок или ошибки были бы так малы, что ими можно было бы пренебречь, то было бы достаточно иметь столько наблюдений сколько и неизвестных. Однако в реальных задачах это не так.

Для того, чтобы можно было надеяться на частичную взаимную компенсацию ошибок, берут число наблюдений (т.е. число условных уравнений) гораздо больше, чем количество неизвестных. Тогда получается алгебраическая (как правило нелинейная) система условных уравнений со случайными правыми частями.

Поскольку в системе есть случайные ошибки, то система, очевидно, несовместна даже при точных функциональных связях. Это означает, что не существует таких $\Theta_0^*, \Theta_1^*, \Theta_2^*, \cdots$, которые удовлетворяли бы всем условным уравнениям одновременно, т.е.

$$f_k(\Theta_0^*, \Theta_1^*, \Theta_2^*, \cdots) - y_k = \epsilon_k \neq 0$$

Величина $f_k(\Theta_0, \Theta_1, \Theta_2, \cdots) - y_k$ называется *невязка*.

Если дана система равноточных условных уравнений, то будем искать неизвестные так, чтобы сумма квадратов невязок была наименьшей, в чем заключается *принцип Лежандра*.

9.2 Принцип Лежандра и метод наименьших квадратов (МНК)

Образуем сумму квадратов невязок

$$S = \sum_{k=1}^{n} \Big[f_k(\Theta_0, \Theta_1, \Theta_2, \cdots) - y_k \Big]^2.$$

Необходимое условие минимума S:

$$\frac{\partial S}{\partial \Theta_0} = \frac{\partial S}{\partial \Theta_1} = \frac{\partial S}{\partial \Theta_2} = \cdots = 0.$$

Полученные уравнения называются *нормальными уравнениями*.

9.3 Обобщенный принцип Лежандра и взвешенный МНК

Принцип Лежандра можно обобщить на неравноточные условные уравнения. Кроме того, можно привести неравноточные уравнения к равноточным.

Пусть известны средние квадратичные ошибки $s_1, s_2, \cdots, s_n$ и найдены веса $p_1, p_2, \cdots, p_n$.

Наиболее вероятная совокупность значений получается при минимизации (*обобщенный принцип Лежандра*)

$$S_p = \sum_{k=1}^{n} p_k \Big[f_k(\Theta_0, \Theta_1, \Theta_2, \cdots) - y_k \Big]^2.$$

Из обобщенного принципа Лежандра легко получить правило приведения неравноточных условных уравнений к равноточным:

$$S_p = \sum_{k=1}^{n} p_k \cdot \epsilon_k^2 = \sum_{k=1}^{n} \Big(\epsilon_k \sqrt{p_k} \Big)^2 = \sum_{k=1}^{n} \tilde{\epsilon}_k^2.$$

10 Линеаризация условных уравнений и представление результата решения условных уравнений

Составить нормальные уравнения можно при любом виде условных уравнений, но решать, особенно в нелинейном случае, довольно трудно. Кроме того, полученные решения нормальных уравнений не будут обязательно линейными по случайной величине y_k. Это затрудняет вычисление средних и среднеквадратических отклонений неизвестных (т.е вычисление точечных оценок и допустимых интервалов для этих неизвестных).

Очевидно, гораздо удобнее работать с линейными уравнениями, когда неизвестные зависят от случайных величин линейным образом.

Приводить к линейному виду можно удачной заменой переменных. Например, пусть условные уравнения имеют вид

$$\alpha_k \sin(\Theta_0 + \Theta_1) + \beta_k \sin(\Theta_0 - \Theta_1) + \gamma_k e^{-2\Theta_2} = y_k,$$

где $\Theta_0, \Theta_1, \Theta_2$ – неизвестные, которые нужно определить, $\alpha_k, \beta_k, \gamma_k$ – заданные неслучайные величины, y_k – случайные величины.

Сделаем подстановку:

$$\sin(\Theta_0 + \Theta_1) = x,$$
$$\sin(\Theta_0 - \Theta_1) = y,$$
$$e^{-2\Theta_2} = z.$$

После подстановки получим линейную систему условных уравнений:

$$\alpha_k x + \beta_k y + \gamma_k z - y_k = 0.$$

Решив эту систему (методом Крамера или методом Гаусса), получим точечные оценки неизвестных $\bar{x}, \bar{y}, \bar{z}$. Поскольку эти величины есть линейные функции случайных величин y_k (которые обычно предполагаются нормально распределенными и независимыми), то и сами оценки неизвестных обладают тем же распределением (нормальным).

Искомые неизвестные

$$\bar{\Theta}_0 = \frac{\arcsin \bar{x} + \arcsin \bar{y}}{2}$$
$$\bar{\Theta}_1 = \frac{\arcsin \bar{x} - \arcsin \bar{y}}{2}$$
$$\bar{\Theta}_2 = \frac{\ln \bar{z}}{2}.$$

Наконец, находим законы распределения $\bar{\Theta}_0, \bar{\Theta}_1, \bar{\Theta}_2$ и потом вычисляем их дисперсии.

Однако такую замену можно сделать далеко не всегда. Рассмотрим общей метод линеаризации условных уравнений, предположив только, что коэффициенты условных уравнений точные, а случайные ошибки y_k малы по модулю.

Рассмотрим эту задачу на примере.

$$\Theta_0 \cdot \sin\left(\frac{2\pi t}{\Theta_3} + \Theta_1\right) + \Theta_2 = w$$

Определить $\Theta_0, \Theta_1, \Theta_2$ и Θ_3, считая, что величина w содержит случайные ошибки и все измерения равноточные.

Случайная величина w задана таблично для разных моментов времени (Таблица (12)).

Таблица 12:

Статистический ряд случайной величины w_k в моменты времени t_k

t_k	0.0	0.52	1.04	1.56	2.08	2.60	3.12	3.64	4.16	4.68	5.20	5.72	6.24
w_k	2.02	1.86	1.49	1.02	0.51	0.14	-0.03	0.15	0.53	0.97	1.46	1.90	1.98

Подставим в заданный закон табличные значения t_k и w_k и получим систему из 13-ти нелинейных условных уравнений

$$\Theta_0 \cdot \sin\left(\frac{2\pi t_k}{\Theta_3} + \Theta_1\right) + \Theta_2 = w_k$$

Найдем (любым способом) предварительные приближенные значения. Период $\Theta_3^0 = 6.30$, потому что при $t = 6.24(< 6.30)$ еще не получено исходное значение (2.02), т.е. период колебания должен быть немного больше. Величины $\Theta_2^0 = 1.08$ и $\Theta_0^0 = 0.90$, т.к. максимальное значение w близко к 2 и среди значений есть близкое к 0, что означает, что Θ_2 и Θ_0 есть величины одного порядка; пусть Θ_2^0 есть среднее арифметическое всех w_k. Далее, примем $\Theta_1^0 = 1.50$ – это следует из грубого сравнения w с обычной синусоидой.

Выбрав начальные приближения искомых неизвестных параметров, положим

$$\Theta_0 = \Theta_0^0 + x_1,$$

$$\Theta_1 = \Theta_1^0 + y_1,$$

$$\Theta_2 = \Theta_2^0 + z_1$$

$$\Theta_3 = \Theta_3^0 + u_1.$$

Подставим эти выражения в условные уравнения и разложим функции (f_k) в ряды по степеням x_1, y_1, z_1, u_1 и ограничимся в разложениях первыми степенями этих поправок:

$$f_k(\Theta_0^0, \Theta_1^0, \Theta_2^0, \Theta_3^0) - y_k + \left.\frac{\partial f_k}{\partial \Theta_0}\right|_0 x_1 + \left.\frac{\partial f_k}{\partial \Theta_1}\right|_0 y_1 + \left.\frac{\partial f_k}{\partial \Theta_2}\right|_0 z_1 + \left.\frac{\partial f_k}{\partial \Theta_3}\right|_0 u_1 = 0,$$

где

$$f_k = \Theta_0 \cdot \sin\left(\frac{2\pi t_k}{\Theta_3} + \Theta_1\right) + \Theta_2$$

$$y_k = w_k.$$

$$x_1 \cdot \sin\left(\frac{2\pi t_k}{\Theta_3^0} + \Theta_1^0\right) + \Theta_0^0 y_1 \cdot \cos\left(\frac{2\pi t_k}{\Theta_3^0} + \Theta_1^0\right) - u_1 \Theta_0^0 \cdot \frac{2\pi t_k}{\left(\Theta_3^0\right)^2} \cdot \cos\left(\frac{2\pi t_k}{\Theta_3^0} + \Theta_1^0\right) + z_1 +$$

$$+ \left[\Theta_0^0 \cdot \sin\left(\frac{2\pi t_k}{\Theta_3^0} + \Theta_1^0\right) + \Theta_2^0 - w_k\right] = 0.$$

$$k = 1, 2, \cdots, 13.$$

Перепишем (переобозначим) систему в виде

$$a_k x_1 + b_k y_1 + c_k z_1 + d_k u_1 + y_k = 0,$$

где

$$a_k = \sin \tau_k,$$
$$\tau_k = \frac{2\pi t_k}{\Theta_3^0} + \Theta_1^0,$$
$$b_k = \Theta_0^0 \cdot \cos \tau_k,$$
$$c_k = 1,$$
$$d_k = -\Theta_0^0 \cdot \cos\left(\tau_k\right) \cdot \frac{2\pi t_k}{\left(\Theta_3^0\right)^2},$$
$$y_k = \Theta_0^0 \cdot \sin \tau_k + \Theta_2^0 - w_k.$$

Теперь необходимо решить систему условных уравнений, использовав принцип Лежандра, и определить

$$x_1, y_1, z_1, u_1,$$

а также среднеквадратические ошибки этих величин

$$s_{x1}, s_{y1}, s_{z1}, s_{u1}.$$

Пусть дана система линейных условных уравнений

$$a_k x + b_k y + c_k z + d_k u + y_k = 0,$$

где неизвестные:

$$x, y, z, u,$$

неслучайные числа, изменяющиеся от уравнения к уравнению:

$$a_k, b_k, c_k, d_k,$$

а случайные ошибки содержаться только в y_k.

Невязки есть

$$\epsilon_k = a_k x + b_k y + c_k z + d_k u + y_k.$$

Согласно принципу Лежандра, нужно минимизировать

$$S = \sum_{k=1}^{n} \epsilon_k^2 = \sum_{k=1}^{n} (a_k x + b_k y + c_k z + d_k u + y_k)^2.$$

Считая условные уравнения равноточными, запишем необходимое условие минимума:

$$\begin{cases} \frac{\partial S}{\partial x} = 2 \sum_{k=1}^{n} \left(a_k x + b_k y + c_k z + d_k u + y_k\right) \cdot a_k = 0 \\ \\ \frac{\partial S}{\partial y} = 2 \sum_{k=1}^{n} \left(a_k x + b_k y + c_k z + d_k u + y_k\right) \cdot b_k = 0 \\ \\ \frac{\partial S}{\partial z} = 2 \sum_{k=1}^{n} \left(a_k x + b_k y + c_k z + d_k u + y_k\right) \cdot c_k = 0 \\ \\ \frac{\partial S}{\partial u} = 2 \sum_{k=1}^{n} \left(a_k x + b_k y + c_k z + d_k u + y_k\right) \cdot d_k = 0 \end{cases}$$

Эти условия приводят к нормальным уравнениям

$$\begin{cases} x\sum_{k=1}^n a_k^2 + y\sum_{k=1}^n a_k b_k + z\sum_{k=1}^n a_k c_k + u\sum_{k=1}^n a_k d_k + \sum_{k=1}^n a_k y_k = 0 \\ x\sum_{k=1}^n b_k a_k + y\sum_{k=1}^n b_k^2 + z\sum_{k=1}^n b_k c_k + u\sum_{k=1}^n b_k d_k + \sum_{k=1}^n b_k y_k = 0 \\ x\sum_{k=1}^n c_k a_k + y\sum_{k=1}^n c_k b_k + z\sum_{k=1}^n c_k^2 + u\sum_{k=1}^n c_k d_k + \sum_{k=1}^n c_k y_k = 0 \\ x\sum_{k=1}^n d_k a_k + y\sum_{k=1}^n d_k b_k + z\sum_{k=1}^n d_k c_k + u\sum_{k=1}^n d_k^2 + \sum_{k=1}^n d_k y_k = 0 \end{cases}$$

Аналогично составляются нормальные уравнения при любом числе неизвестных.

Решение системы нормальных уравнений методом Крамера (удобно для систем не выше 4-го порядка):

$$\bar{x} = \frac{\Delta_x}{\Delta}, \bar{y} = \frac{\Delta_y}{\Delta}, \bar{z} = \frac{\Delta_z}{\Delta}, \bar{u} = \frac{\Delta_u}{\Delta},$$

где Δ – определитель основной матрицы системы. Для переменной $\bar{x}$

$$\Delta_x = -\begin{vmatrix} [ay] & [ab] & [ac] & [ad] \\ [by] & [bb] & [bc] & [bd] \\ [cy] & [cb] & [cc] & [cd] \\ [dy] & [db] & [dc] & [dd] \end{vmatrix} = -\sum_{k=1}^n y_k \cdot \begin{vmatrix} a_k & [ab] & [ac] & [ad] \\ b_k & [bb] & [bc] & [bd] \\ c_k & [cb] & [cc] & [cd] \\ d_k & [db] & [dc] & [dd] \end{vmatrix}$$

где обозначено

$$[ab] = \sum_{k=1}^n a_k b_k.$$

Обозначим

$$\Delta_k = \begin{vmatrix} a_k & [ab] & [ac] & [ad] \\ b_k & [bb] & [bc] & [bd] \\ c_k & [cb] & [cc] & [cd] \\ d_k & [db] & [dc] & [dd] \end{vmatrix}.$$

Тогда

$$\bar{x} = \frac{\Delta_x}{\Delta} = -\sum_{k=1}^n \frac{\Delta_k}{\Delta} y_k.$$

Тогда

$$s_{\bar{x}}^2 = \sum_{k=1}^n \frac{\Delta_k^2}{\Delta^2} s_k^2. \tag{9}$$

Докажем утверждение (9).

Действительно, с учетом взаимной независимости величин y_k дисперсия величины $\bar{x}$ по определению есть

$$D[\bar{x}] \equiv s_{\bar{x}}^2 = D\Big[-\sum_{k=1}^n \frac{\Delta_k}{\Delta} y_k\Big] = \sum_{k=1}^n \left(\frac{\Delta_k}{\Delta}\right)^2 D[y_k] = \sum_{k=1}^n \frac{\Delta_k^2}{\Delta^2} s_k^2.$$

Для равноточных данных

$$s_{\bar{x}}^2 = \sum_{k=1}^n \frac{\Delta_k^2}{\Delta^2} s_0^2.$$

Можно вывести, что

$$\sum_{k=1}^{n} \Delta_k^2 = \Delta\Delta_{11},$$

где Δ_{11} есть алгебраическое дополнение первого диагонального элемента основной матрицы системы. Для доказательства достаточно заметить, что

$$\sum_{k=1}^{n} \Delta_k^2 = \sum_{k=1}^{n} \Delta_k \cdot \Delta_k = \begin{vmatrix} \sum a_k\Delta_k & [ab] & [ac] & [ad] \\ \sum b_k\Delta_k & [bb] & [bc] & [bd] \\ \sum c_k\Delta_k & [cb] & [cc] & [cd] \\ \sum d_k\Delta_k & [db] & [dc] & [dd] \end{vmatrix},$$

причем определители, содержащие одинаковые столбцы, равны нулю

$$\sum b_k\Delta_k = \sum c_k\Delta_k = \sum d_k\Delta_k = 0,$$

а

$$\sum a_k\Delta_k = \Delta.$$

Раскладывая четырехмерный определитель по первому элементу, получаем искомое соотношение

$$\sum_{k=1}^{n} \Delta_k^2 = \Delta \cdot (-1)^{1+1}\Delta_{11}.$$

Тогда для равноточных данных

$$s_{\bar{x}}^2 = \frac{\Delta_{11}}{\Delta} \cdot s_0^2.$$

и

$$p_{\bar{x}} = \frac{s_0^2}{s_{\bar{x}}^2} = \frac{\Delta}{\Delta_{11}}$$

Аналогично для других неизвестных:

$$\bar{y} = \frac{\Delta_y}{D}, \bar{z} = \frac{\Delta_z}{\Delta}, \bar{u} = \frac{\Delta_u}{\Delta}$$

и

$$p_{\bar{y}} = \frac{\Delta}{\Delta_{22}}, p_{\bar{z}} = \frac{\Delta}{\Delta_{33}}, p_{\bar{u}} = \frac{\Delta}{\Delta_{44}},$$

где Δ – определитель основной матрицы линейной системы нормальных уравнений, а Δ_{kk} – соответствующие алгебраические дополнения k-го диагонального элемента основной матрицы.

После решения нормальных уравнений получаются наиболее вероятные значения неизвестных: $\bar{x}, \bar{y}, \bar{z}, \bar{u}$. Их подстановка в условные уравнения даст невязки, удовлетворяющие условию минимума суммы квадратов. Эти невязки называются *остаточными погрешностями* и обозначаются ϵ_k.

Сумма квадратов остатков есть

$$\bar{S} =$$

$$= \sum_{k=1}^{n} \epsilon_k^2 = \sum_{k=1}^{n} \Big(a_k\bar{x}+b_k\bar{y}+c_k\bar{z}+d_k\bar{u}+y_k \Big)^2 = \bar{x}\sum_{k=1}^{n} a_k y_k + \bar{y}\sum_{k=1}^{n} b_k y_k + \bar{z}\sum_{k=1}^{n} c_k y_k + \bar{u}\sum_{k=1}^{n} d_k y_k + \sum_{k=1}^{n} y_k^2.$$

наиболее вероятное значение среднеквадратической ошибки на единицу веса s_0:

$$s_0 = \sqrt{\frac{\bar{S}}{n-m}},$$

где m – число неизвестных.

Тогда среднеквадратические ошибки для неизвестных есть

$$s_{\bar{x}} = \frac{s_0}{\sqrt{p_{\bar{x}}}},$$
$$s_{\bar{y}} = \frac{s_0}{\sqrt{p_{\bar{y}}}},$$
$$s_{\bar{z}} = \frac{s_0}{\sqrt{p_{\bar{z}}}},$$
$$s_{\bar{u}} = \frac{s_0}{\sqrt{p_{\bar{u}}}}.$$

Окончательно решение задачи (в первом приближении) записывается как

$$x = \bar{x} \pm s_{\bar{x}},$$
$$y = \bar{y} \pm s_{\bar{y}},$$
$$z = \bar{z} \pm s_{\bar{z}},$$
$$u = \bar{u} \pm s_{\bar{u}}.$$

Для рассматриваемого случая

$$\bar{x} = 0.096; \bar{y} = 0.092; \bar{z} = -0.075; \bar{u} = -0.026.$$
$$p_{\bar{x}} = 6.2; p_{\bar{y}} = 0.40; p_{\bar{z}} = 4.5; p_{\bar{u}} = 0.113.$$
$$\bar{S} = 0.006925.$$
$$s_0^2 = \frac{0.007}{13-4} = 0.000778; s_0 = 0.028.$$
$$s_{\bar{x}}^2 = \frac{0.000778}{6.2} = 0.00013, s_{\bar{x}} = 0.011;$$
$$s_{\bar{y}}^2 = \frac{0.000778}{0.40} = 0.00020, s_{\bar{y}} = 0.045;$$
$$s_{\bar{z}}^2 = \frac{0.000778}{4.5} = 0.00017, s_{\bar{z}} = 0.013;$$
$$s_{\bar{u}}^2 = \frac{0.000778}{0.113} = 0.0069, s_{\bar{u}} = 0.083.$$

Окончательно решение первого приближения есть

$$x = 0.096 \pm 0.011,$$
$$y = 0.092 \pm 0.045,$$
$$z = -0.075 \pm 0.013,$$
$$u = -0.026 \pm 0.083.$$

После того, как найдено решение первого приближения, можно построить второе приближение и т.д., аналогичным образом строя системы условных уравнений, сводя их к системе нормальных уравнений и решая. Процесс может быть остановлен тогда, когда в двух последовательных итерациях с заданной точностью получают одинаковые значения. Так, точечные оценки решения второго приближения есть $\{\bar{x}, \bar{y}, \bar{z}, \bar{u}\} = \{0.004, -0.013, -0.001, -0.006\}$, а величина $\bar{S} = 0.006864$.

11 Однофакторный дисперсионный анализ

Различают три типа связи между случайными величинами и, соответственно, три группы методов. *Дисперсионный анализ* устанавливает наличие возмущающего фактора, который влияет на статистическую совокупность выборочных данных. Степень влияния внешних факторов можно определить методами *корреляционного анализа*. Конкретная математическая модель влияния устанавливается *регрессионным анализом*, [7].

Различают *однофакторный* и многофакторный дисперсионный анализ. Суть однофакторного дисперсионного анализа состоит в том, чтобы установить наличие изменения дисперсии выборочных данных при изменении уровней влияния какого-то одного внешнего фактора. Если при изменении этого фактора дисперсия выборки будет значимо изменяться,то этот фактор должен быть признан значимым в своем влиянии на среднее значение наблюдаемой величины. Дисперсионный анализ дает возможность только установить наличие значимо влияющего фактора, но не дает возможность количественно оценить силу его влияния и тем более не дает математическую модель этого фактора.

Рассмотрим задачу однофакторного дисперсионного анализа.

Для анализа необходимо иметь несколько выборок случайных данных. полученных из одной генеральной совокупности. Перед началом анализа надо проверить, что распределение исходных элементов выборки подчиняется нормальному распределению (методы такой проверки будет рассмотрены в последнем разделе). Кроме того, надо проверить, чтобы дисперсии выборок были одинаковыми (проверка равенства дисперсий двух выборок производится по F-критерию Фишера).

Анализируется влияние фактора $A = \{A_1, A_2, \cdots, A_k\}$ (Таблица (13)). В каждом столбце выборка $\{x_{i1}, x_{i2}, \cdots, x_{in}\}$ характеризует изменение данных под влиянием фактора A уровня A_i. Точечная оценка дисперсии такой i-ой выборки (при неизвестном математическом ожидании, которое также должно оцениваться по этой выборке) есть

$$s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_{ij} - \frac{1}{n}\sum_{j=1}^{n} x_{ij}\right)^2.$$

Пусть проверено, что для всех $i = 1, 2, \cdots, k$ все $s_i^2 = const$.

Эти k выборок (каждая из которых отвечает своему уровню критерия A) можно рассматривать как объединенную выборку, объем которой есть $\sum_{i=1}^{k} n = n \cdot k$. Среднее этой объединенной выборки есть

$$\bar{x} = \frac{1}{n \cdot k}\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij},$$

а дисперсия

$$s^2 = \frac{1}{n \cdot k - 1}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}\right)^2 = \frac{1}{n \cdot k - 1}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij} - \frac{1}{n \cdot k}\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}\right)^2 =$$

$$= \frac{1}{n \cdot k - 1}\left[(n \cdot k - k) \cdot s_0^2 + (k-1) \cdot s_A^2\right],$$

где

$$s_0^2 = \frac{1}{n \cdot k - k}\sum_{i=1}^{k}(n-1) \cdot s_i^2$$

Таблица 13:

Уровни фактора $A = \{A_1, A_2, \cdots, A_i, \cdots, A_k\}$ *при однофакторном дисперсионном анализе.*

Номер наблюдения	A_1	A_2	$\cdots$	A_i	$\cdots$	A_k
1	x_{11}	x_{21}	$\cdots$	x_{i1}	$\cdots$	x_{k1}
2	x_{12}	x_{22}	$\cdots$	x_{i2}	$\cdots$	x_{k2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
j	x_{1j}	x_{2j}	$\cdots$	x_{ij}	$\cdots$	x_{kj}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{1n}	x_{2n}	$\cdots$	$\vdots$	$\cdots$	x_{kn}

есть суммарная дисперсия (или рассеяние внутри выборки), а

$$s_A^2 = \frac{1}{k-1}\sum_{i=1}^{k} n \cdot \left(\frac{1}{n}\sum_{j=1}^{n} x_{ij} - \frac{1}{n\cdot k}\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}\right)^2,$$

есть дисперсия между выборками (т.е. сумма квадратов отклонений средних по выборке от среднего объединенной выборки).

- Докажем тождество (*связь дисперсии объединенной выборки с дисперсией внутри одной выборки и с дисперсией между выборками*):

$$s^2 = \frac{1}{n\cdot k - 1}\Big[(n\cdot k - k)\cdot s_0^2 + (k-1)\cdot s_A^2\Big].$$

Для доказательства и в методических целях распишем выражения s^2, s_0^2, s_A^2:

$$s^2\cdot(n\cdot k - 1) = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij} - \frac{1}{n\cdot k}\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}\left\{x_{ij}^2 - \frac{2}{n\cdot k}x_{ij}\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij} + \right.$$

$$\left. + \frac{1}{(n\cdot k)^2}\left(\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}\right)^2\right\} = \sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2 - \frac{2}{n\cdot k}\left(\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}\right)^2 + \frac{n\cdot k}{(n\cdot k)^2}\left(\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}\right)^2 =$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2 - \frac{1}{n\cdot k}\left(\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}\right)^2.$$

$$s_0 = \frac{1}{n\cdot k - k}\sum_{i=1}^{k}(n-1)\cdot\frac{1}{n-1}\sum_{j=1}^{n}\left(x_{ij} - \frac{1}{n}\sum_{j=1}^{n} x_{ij}\right)^2 =$$

$$= \frac{1}{k\cdot(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij} - \frac{1}{n}\sum_{j=1}^{n} x_{ij}\right)^2 = \frac{1}{k\cdot(n-1)}\left\{\sum_{i=1}^{k}\sum_{j=1}^{n}\left[x_{ij}^2 - \frac{2}{n}x_{ij}\sum_{j=1}^{n} x_{ij} + \right.\right.$$

$$+\frac{1}{n^2}\left(\sum_{j=1}^{n} x_{ij}\right)^2\Bigg]\Bigg\} = \frac{1}{k\cdot(n-1)}\Bigg\{\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2 - \frac{2}{n}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij}\sum_{j=1}^{n} x_{ij}\right)+$$

$$+\frac{1}{n^2}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(\sum_{j=1}^{n} x_{ij}\right)^2\Bigg\} = \frac{1}{k\cdot(n-1)}\Bigg\{\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2 - \frac{1}{n}\sum_{i=1}^{k}\left(\sum_{j=1}^{n} x_{ij}\right)^2\Bigg\}.$$

$$s_A = \frac{1}{k-1}\sum_{i=1}^{k} n\cdot\left(\frac{1}{n}\sum_{j=1}^{n} x_{ij} - \frac{1}{n\cdot k}\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}\right)^2 = \frac{1}{n\cdot(k-1)}\sum_{i=1}^{k}\Bigg(\sum_{j=1}^{n} x_{ij}-$$

$$-\frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}\Bigg)^2 = \frac{1}{n\cdot(k-1)}\Bigg\{\sum_{i=1}^{k}\left(\sum_{j=1}^{n} x_{ij}\right)^2 - \frac{2}{k}\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}+$$

$$+\sum_{i=1}^{k}\frac{1}{k^2}\left(\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}\right)^2\Bigg\} = \frac{1}{n\cdot(k-1)}\Bigg\{\sum_{i=1}^{k}\left(\sum_{j=1}^{n} x_{ij}\right)^2 - \frac{1}{k}\left(\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}\right)^2\Bigg\}.$$

таким образом, нужно доказать, что

$$\frac{1}{n\cdot k-1}\cdot\left[\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2 - \frac{1}{n\cdot k}\left(\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}\right)^2\right] =$$

$$= \frac{1}{n\cdot k-1}\cdot\Bigg[(n\cdot k-k)\cdot\left[\frac{1}{k\cdot(n-1)}\left\{\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2 - \frac{1}{n}\sum_{i=1}^{k}\left(\sum_{j=1}^{n} x_{ij}\right)^2\right\}\right]+$$

$$+(k-1)\cdot\left[\frac{1}{n\cdot(k-1)}\left\{\sum_{i=1}^{k}\left(\sum_{j=1}^{n} x_{ij}\right)^2 - \frac{1}{k}\left(\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}\right)^2\right\}\right]\Bigg].$$

Раскрывая скобки, получим

$$\frac{1}{nk-1}\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2 - \frac{1}{nk(nk-1)}\left(\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}\right)^2 =$$

$$= \frac{k(n-1)}{k(n-1)(nk-1)}\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2 - \frac{k(n-1)}{nk(n-1)(nk-1)}\sum_{i=1}^{k}\left(\sum_{j=1}^{n} x_{ij}\right)^2+$$

$$+\frac{k-1}{n(k-1)(nk-1)}\sum_{i=1}^{k}\left(\sum_{j=1}^{n} x_{ij}\right)^2 - \frac{k-1}{nk(k-1)(nk-1)}\left(\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}\right)^2,$$

что и требовалось доказать.

Проверка влияния фактора A на изменение средних сводится к сравнению двух дисперсий, s_A^2 и s_0^2. Влияние фактора A признается значимым, если значимо отношение s_A^2/s_0^2. Отношение s_A^2/s_0^2 признается значимым с доверительной вероятностью $\gamma = 1-\alpha$, если

$$\frac{s_A^2}{s_0^2} > F(\alpha; k-1, k(n-1)),$$

где $F(\alpha; k-1, k(n-1))$ – γ-квантиль (или α-процентная точка) F-распределения Фишера с $(k-1), k(n-1)$ степенями свободы.

12 Корреляционный анализ

Корреляционный анализ предполагает изучение зависимости между случайными величинами с одновременной количественной оценкой степени неслучайности их совместного изменения, [7].

Зависимость между случайными величинами X и Y характеризуется *коэффициентом корреляции*, точное значение которого есть

$$\rho = \frac{M[(X - m_x) \cdot (Y - m_y)]}{\sqrt{D[X] \cdot D[Y]}}.$$

Коэффициент корреляции показывает, насколько зависимость между случайными величинами X и Y близка к строго линейной. Если X и Y имеют нормальное распределения, то $\rho = 0$ для них означает отсутствие линейной связи. Равенство $|\rho| = 1$ означает наличие строгой линейной связи.

12.1 Оценка коэффициента корреляции

В случае работы с реальными данными для двух случайных величин $X = \{x_1, x_2, \cdots, x_n\}$ и $Y = \{y_1, y_2, \cdots, y_n\}$ *выборочный коэффициент корреляции* $r = q = \rho^*$ (приводятся различные обозначения коэффициента корреляции, встречаемые в литературе) есть

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

где

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

Если $r = 0$, то X и Y не коррелированные (но могут быть зависимые). Если $|r| = 1$, то между X и Y существует зависимость в виде прямой пропорциональности.

При малом объеме выборки, $n < 15$, коэффициент корреляции лучше оценивать по формуле, [7]:

$$\tilde{r} = r \cdot \left[1 + \frac{1 - r^2}{2(n-3)}\right].$$

При большом объеме выборки, $n > 200$, выборочный коэффициент корреляции r имеет нормальное распределение

$$r \sim N(m_r, \sigma_r^2),$$

$$m_r = r, \sigma_r^2 = \frac{1 - r^2}{n - 1}.$$

12.2 Исследование значимости корреляции

На практике наибольшую важность представляет задача исследования значимости корреляции, т.е. исследуется, насколько сильно коэффициент корреляции отличен от нуля. Для этой цели вычисляется выборочное значение коэффициента корреляции r и

Таблица 14:

Случайные величины X и Y, исследуемые на корреляционную зависимость

x_i	2	4	1	7	3	11	14	15	21	4
y_i	7	6	4	11	2	21	31	23	40	15

сравнивается с табличным критическим значением r_γ. Для выборки большого объема, $n > 200$, критического значение коэффициента корреляции хорошо аппроксимируется u_γ-квантилью нормального распределения:

$$r_\gamma = \frac{1}{\sqrt{n-1}} \cdot u_\gamma.$$

Рассмотрим пример определения значимости корреляционной зависимости между двумя выборками, [7].

- В результате наблюдений над случайными величинами X и Y получена совокупность данных из 10-ти элементов для каждой случайной величины (см. Таблица (14)) Необходимо проверить, есть ли корреляция между X и Y с доверительной вероятностью $\gamma = 0.95$

  Для решения находим характеристики выборок:

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 8.2, \sum_{i=1}^{10}(x_i - \bar{x})^2 = 405.6, \bar{y} = \frac{1}{10}\sum_{i=1}^{10} y_i = 16.0, \sum_{i=1}^{10}(y_i - \bar{y})^2 = 1422.0,$$

$$\sum_{i=1}^{10}(x_i - \bar{x}) \cdot (y_i - \bar{y}) = 723.0.$$

  Далее получаем оценки коэффициента корреляции

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{723}{\sqrt{405.6 \cdot 1422}} = 0.952.$$

  Используя приближенную оценку, лучшую для малых выборок

$$\tilde{r} = r \cdot \left[1 + \frac{1 - r^2}{2(n-3)}\right] = 0.952 \cdot \left(1 + \frac{1 - 0.952^2}{2 \cdot 7}\right) = 0.958.$$

  Используя u_γ-квантиль нормального распределения, применим формулу для оценки критического значения коэффициента корреляции для больших выборок

$$r_\gamma = r_{0.95} = \frac{1}{\sqrt{n-1}} \cdot u_\gamma = \frac{1.96}{3} = 0.653.$$

Если учесть, что рассматриваемая выборка мала ($n = 10$), то

$$r_{0.95} = 0.632.$$

Последнее значение можно получить из таблицы точных критических значений r_γ, [7].

В любом случае, $\tilde{r} = 0.958 > 0.653$, и корреляция признается значимой с доверительной вероятностью $\gamma = 0.95$.

13 Регрессионный анализ

Методы дисперсионного и корреляционного анализа позволяют определить, есть ли связь между случайными величинами и позволяют оценить силу этой связи. Теперь необходимо уметь определять конкретный вид функциональной зависимости между случайными величинами – в этом и заключается задача регрессионного анализа.

Пусть исследуется связь между двумя случайными выборками $X = \{x_1, x_2, \cdots, x_n\}$ и $Y = \{y_1, y_2, \cdots, y_n\}$. *Регрессией y по x* называется зависимость средних значений случайной величины Y от средних значений случайной величины X. Методы нахождения этой зависимости и обязательные оценки статистических свойств этой зависимости – задача *регрессионного анализа.*

По выборочным данным можно, очевидно, найти только оценку истиной регрессии, которая будет содержать ошибки, связанные со случайностью и ограниченностью выборки.

В основе регрессионного анализа лежит метод наименьших квадратов (МНК). Согласно МНК, в качестве уравнения регрессии $y = f(x)$ выбирается функция, которая дает минимум сумме квадратов разностей

$$S = \sum_{i=1}^{n} \Big[y_i - f(x_i) \Big]^2 .$$

Как правило, общий вид функции $f(x)$ определяется заранее, а методом наименьших квадратов определяются коэффициенты функции $f(x)$, минимизирующие S. Количественная мера рассеяния значений y_i вокруг регрессии $f(x)$ является дисперсия:

$$D = \frac{1}{n-k} \sum_{i=1}^{n} \Big[y_i - f(x_i) \Big]^2 ,$$

где k – число коэффициентов, входящих в аналитическое выражение регрессии (например, если $f(x)$ – многочлен степени m, то $k = m + 1$).

В зависимости от вида функции $f(x)$ различают *линейную регрессию*:

$$f(x) = a + b \cdot x$$

и *нелинейную регрессию*:

$$f(x) = a_0 + a_1 \cdot x + a_2 \cdot x^2 + \cdots .$$

В задачах нелинейной регрессии часто используют разного рода линеаризующие преобразования (например, замену переменных). При невозможности или неэффективности линеаризации регрессия строится с помощью многочленов специального вида – ортогональных полиномов (например, полиномов Чебышева).

- *Общая схема построения линейной регрессии:*

  1. Задание определенной регрессионной модели с неизвестными коэффициентами вида $f(x) = a + b \cdot x$;
  2. Нахождение выборочной оценки истинной регрессии по данным $\{x_1, x_2, \cdots, x_n\}$ и $\{y_1, y_2, \cdots, y_n\}$, т.е. нахождение неизвестных коэффициентов a и b методом МНК из условия минимума выражения $\sum_{i=1}^{n} \Big[y_i - f(x_i) \Big]^2$;

3. Оценка статистической значимости выборочной регрессии;
4. Нахождение доверительного интервала выборочной регрессии (включающего в себя с заданной вероятностью истинную регрессию);
5. Анализ регрессионных остатков, исследование на наличие выбросов.

13.1 Постановка задачи линейного регрессионного анализа

Модель зависимости двух случайных выборок в линейном регрессионном анализе –

$$y = f(x) = \alpha + \beta \cdot x,$$

где α и β – истинные коэффициенты регрессии. Их выборочные оценки будем обозначать a и b соответственно.

Условие минимума по α и β функционала

$$S = \sum_{i=1}^{n} \Big[y_i - f(x_i) \Big]^2$$

дает систему двух уравнений:

$$\begin{cases} \frac{\partial S}{\partial \alpha} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \Big(\alpha + \beta \cdot x_i \Big) = 0, \\ \frac{\partial S}{\partial \beta} = \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \Big(\alpha + \beta \cdot x_i \Big) \cdot x_i = 0, \end{cases}$$

из которой следует система

$$\begin{cases} n \cdot \alpha + \beta \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \\ \alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i. \end{cases}$$

Решение системы дает искомые оценки неизвестных коэффициентов:

$$b = \frac{n \sum\limits_{i=1}^{n} x_i y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{n \sum\limits_{i=1}^{n} x_i^2 - \Big(\sum\limits_{i=1}^{n} x_i \Big)^2}$$

$$a = \frac{\sum\limits_{i=1}^{n} y_i - b \sum\limits_{i=1}^{n} x_i}{n}.$$

Для проверки правильности вычислений можно использовать соотношения:

$$\bar{y} = a + b \cdot \bar{x},$$

где

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i; \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

Важной особенностью регрессионных уравнений является следующая. Регрессия y по x: $y = \alpha + \beta \cdot x$ не эквивалентна в общем случае регрессии x по y: $x = \alpha^* + \beta^* \cdot y$.

Если s_x, s_y – среднеквадратические отклонения случайных величин X, Y соответственно, т.е.

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}\Big(x_i - \bar{x}\Big)^2,$$

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\Big(y_i - \bar{y}\Big)^2,$$

то регрессии $y = f(x)$ и $x = \phi(y)$ можно записать следующим образом:

$$y = \bar{y} + r \cdot \frac{s_y}{s_x} \cdot \Big(x - \bar{x}\Big),$$

$$x = \bar{x} + r \cdot \frac{s_x}{s_y} \cdot \Big(y - \bar{y}\Big),$$

где r – выборочный коэффициент корреляции.

Регрессии y по x и x по y совпадают только в одном случае, когда существует корреляция между y и x с коэффициентом корреляции, по модулю точно равным единице. Если $r = 0$, то прямые регрессии y по x и x по y перпендикулярны друг другу и тогда $\beta = r \cdot s_y/s_x$, $\beta^* = r \cdot s_x/s_y$.

Если $s_x = s_y$, то коэффициенты корреляции и регрессии совпадают.

Рассмотрим статистический анализ найденных оценок коэффициентов a и b линейной регрессии.

13.2 Статистический анализ параметров линейной регрессии

Для того, чтобы линейная модель оказалась удовлетворительной для описания зависимости двух случайных величин x и y, прежде всего необходимо проверить, не равен ли коэффициент β нулю, т.е. нужно проверить значимость его отклонения от нуля (в противном случае равенство нулю коэффициента при x означает, что модель линейной регрессии не подходит).

Для проверки значимости отклонения β от нуля используется статистика t-распределения Стьюдента. Значение β является значимым с доверительной вероятностью γ (процентной точкой $\alpha = 1 - \gamma$), если

$$|b| > t_{(n-2),\alpha/2\%} \cdot s_\beta.$$

Доверительный интервал с доверительной вероятностью γ (процентной точкой $\alpha = 1 - \gamma$) для истинного коэффициента β определяется как

$$b - s_\beta \cdot t_{(n-2),\alpha/2\%} \le \beta \le b + s_\beta \cdot t_{(n-2),\alpha/2\%},$$

где

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \Big(\sum_{i=1}^{n} x_i\Big)^2}.$$

Здесь число степеней свободы t-распределения есть $(n-2)$ (поскольку оцениваются два неизвестных коэффициента) и по таблице t-распределения следует искать $t = t_{(n-2),\alpha/2\%}$. Далее,

$$s_\beta = \frac{s}{s_x \cdot \sqrt{n-1}},$$

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}\Big(y_i - a - bx_i\Big)^2,$$

$$(s^2 = \frac{1}{n-2}\cdot S)$$

(S – сумма квадратов невязок)

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}\Big(x_i - \bar{x}\Big)^2,$$

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

Аналогично предыдущему, доверительный интервал с доверительной вероятностью γ (или α-процентной точкой) для истинного коэффициента α (не путайте обозначение коэффициента и обозначение процентной точки) определяется как

$$a - s_\alpha \cdot t_{(n-2),\alpha/2\%} \le \alpha \le a + s_\alpha \cdot t_{(n-2),\alpha/2\%},$$

где

$$s_\alpha = s \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)\cdot s_x^2}}.$$

Довольно громоздкий вывод выражений для s_α и s_β основан на требовании, чтобы

$$\frac{b-\beta}{s_\beta} \sim T(n-2) \text{ и } \frac{a-\alpha}{s_\alpha} \sim T(n-2)$$

и проводится с помощью представления распределения Стьюдента как отношения нормального распределения к квадратному корню из χ^2-распределения. Другими словами, путем деления на среднеквадратическое отклонение оценки параметра линейной регресси в числителе формируется величина, обладающая нормальным законом распределения, а в знаменателе формируется величина, обладающая распределением $\sqrt{\chi^2}$, [12].

ПРИМЕР Оценки параметров линейной регрессии. Пусть задана совокупность данных (Таблица (15)). Для этих данных нужно найти точечную и интервальную оценки коэффициентов α и β регрессии $y = \alpha + \beta \cdot x$. Принять доверительную вероятность $\gamma = 0.95$. Обозначим для компактности вычислений $\sum_{i=1}^{n} = \sum_{i=1}^{10} = \sum$.

Сначала вычислим точечные оценки β и α. Для

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \Big(\sum x_i\Big)^2}$$

Таблица 15:

Представление закона распределения случайной величины в виде таблицы – статистического ряда распределения

x_i	1.2	2.4	2.8	4.2	5.9	6.8	8.1	9.2	10.1	11.0
y_i	7	12	17	24	29	38	46	45	54	68

вычислим необходимые суммы[7]:

$$\sum x_i = 61,7; \left(\sum x_i\right)^2 = 3806.89; \sum x_i^2 = 486.99; \sum y_i = 340; \sum x_i y_i = 2695.1.$$

Тогда

$$b = \frac{10 \cdot 2695.1 - 61.7 \cdot 340}{10 \cdot 486.99 - 3806.89} = 5.6189,$$

$$a = \frac{\sum y_i - b\sum x_i}{n} = \frac{340 - 5.6189 \cdot 61.7}{10} = -0.668.$$

Построим доверительные интервалы для α и β. Предварительно вычислим:

$$\bar{x} = 6.17; s_x^2 = \frac{1}{n-1}\sum\left(x_i - \bar{x}\right)^2 = 11.8112; s_x = 3.3467.$$

Далее вычислим дисперсию

$$s^2 = \frac{1}{n-2}\sum\left(y_i - \hat{y}_i\right)^2,$$

где

$$\hat{y}_i = a + b \cdot x_i.$$

Для нашей задачи

$$\hat{y}_i = \{6.075, 12.818, 15.065, 22.932, 32.484, 37.541, 44.846, 51.027, 56.084, 61.141\}.$$

Тогда

$$s^2 = \frac{1}{8}\sum\left(y_i - \hat{y}_i\right)^2 = 13.4755.$$

$$s_\beta = \frac{s}{s_x \cdot \sqrt{n-1}} = 0.3656,$$

$$s_\alpha = s \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)\cdot s_x^2}} = 2.486.$$

[7]При обработки реальных данных перед началом вычислений рекомендуется написать соответствующие простые подпрограммы.

Для доверительной вероятности 0.95 имеем

$$t_{2.5\%}(8) = 2.306.$$

Если искомое линейное приближение верно, то коэффициент b должны быть, очевидно, значимо отличен от нуля. Это проверяется также t-статистикой:

$$|b| = 5.6189 > t_{2.5\%}(8) \cdot s_\beta = 2.306 \cdot 0.3656 = 0.843.$$

Можно также проверить, можно ли, к примеру, округлить коэффициент β до 5:

$$|5.619 - 5.0| = 0.619 < t_{2.5\%}(8) \cdot s_\beta = 0.843.$$

Поскольку модуль разности меньше соответствующей t-статистики, то разность двух значений неотделима от нуля. Следовательно, можно принять $\beta = 5$.

Теперь найдем доверительный интервал для β:

$$5.619 - 2.306 \cdot 0.3656 = 4.776 \leq \beta \leq 6.462 = 5.619 + 2.306 \cdot 0.3656.$$

Аналогично исследуем оценки для коэффициента α. Сначала проверим, можно ли с хорошей точностью считать этот коэффициент равным нулю:

$$|a| = 0.668 < t_{2.5\%}(8) \cdot s_\alpha = 2.306 \cdot 2.485 = 5.73.$$

Неравенство выполняется, следовательно, на заданном уровне точности α не отличима от нуля. Двусторонний доверительный интервал для α:

$$-0.668 - 2.306 \cdot 2.485 = -6.398 \leq \alpha \leq 5.602 = -0.668 + 2.306 \cdot 2.485.$$

Окончательно получаем, что с доверительной вероятностью 0.95 уравнение регрессии есть

$$y = 5 \cdot x.$$

13.3 Оценка остаточной дисперсии и сравнение двух линейных регрессий

Разберем пример сравнения двух линейных регрессий, в который включен анализ остаточных дисперсий.

В результате двух независимых экспериментов получены результаты ($n_1 = 10, n_2 = 6$), указанные в Таблицах (16)-(17). Проверим, являются ли статистически неразличимыми линейные регрессионные модели, полученные по обеим выборкам. Принять доверительную вероятность 0.95.

Первое, что нужно сделать, это выписать регрессионные модели для обеих выборок. Как и раньше, для удобства обозначим для первой выборки

$$\sum_{i=1}^{n_1} = \sum_{i=1}^{10} = \sum_{1}$$

а для второй выборки

$$\sum_{i=1}^{n_2} = \sum_{i=1}^{6} = \sum_{2}$$

Таблица 16:

Статистический ряд распределения x_{1i}, y_{1i}

x_{1i}	2	4	6	9	11	16	17	20	25	31
y_{1i}	9	19	22	41	49	61	69	83	98	128

Таблица 17:

Статистический ряд распределения x_{2i}, y_{2i}

x_{2i}	12	16	21	23	28	31
y_{2i}	54	68	87	93	112	130

Тогда для первой выборки:

$$\sum_1 x_{1i} = 141; \bar{x}_{1i} = 14.1; \left(\sum_1 x_{1i}\right)^2 = 19881; \sum_1 x_{1i}^2 = 2789; \sum_1 y_{1i} = 579;$$

$$\bar{y}_{1i} = 57.9; \sum_1 x_{1i} y_{1i} = 11361.$$

Коэффициенты регрессии для первой выборки:

$$b_1 = \frac{n_1 \sum_1 x_i y_i - \sum_1 x_i \sum_1 y_i}{n_1 \sum_1 x_i^2 - \left(\sum_1 x_i\right)^2} = \frac{10 \cdot 11361 - 141 \cdot 579}{10 \cdot 2789 - 19881} = 3.992;$$

$$a_1 = \frac{\sum_1 y_i - b_1 \sum_1 x_i}{n_1} = \frac{579 - 3.992 \cdot 141}{10} = 1.613.$$

Кроме того, далее понадобиться выборочная дисперсия для x_{1i}:

$$s^2_{\bar{x}_1} = \frac{1}{n_1 - 1} \sum_1 \left(x_{1i} - \bar{x}_1\right)^2 = 88.989$$

$$s_{\bar{x}_1} = 9.433.$$

Аналогичные вычисления проведем для второй выборки.

$$\sum_2 x_{2i} = 131; \bar{x}_{2i} = 21.83; \left(\sum_2 x_{2i}\right)^2 = 17161; \sum_2 x_{2i}^2 = 3115; \sum_2 y_{2i} = 544;$$

$$\bar{y}_{2i} = 90.667; \sum_2 x_{2i} y_{2i} = 12868.$$

Коэффициенты регрессии для второй выборки:

$$b_2 = \frac{n_2 \sum_2 x_i y_i - \sum_2 x_i \sum_2 y_i}{n_2 \sum_2 x_i^2 - \left(\sum_2 x_i\right)^2} = \frac{6 \cdot 12868 - 131 \cdot 544}{6 \cdot 3115 - 17161} = 3.888;$$

$$a_2 = \frac{\sum_2 y_i - b_2 \sum_2 x_i}{n_2} = \frac{544 - 3.888 \cdot 131}{6} = 5.78.$$

Выборочная дисперсия для x_{2i}:

$$s_{\bar{x}_2}^2 = \frac{1}{n_2 - 1} \sum_2 \left(x_{2i} - \bar{x}_2\right)^2 = 50.976$$

$$s_{\bar{x}_2} = 7.139.$$

Теперь вычислим дисперсии рассеяния значений y_{1i} и y_{2i} вокруг своих линий регрессии.

$$s_1^2 = \frac{1}{n_1 - 2} \sum_1 \left(y_{1i} - a_1 - b_1 \cdot x_{1i}\right)^2 = \frac{1}{10 - 2} \sum_1 \left(y_{1i} - 1.613 - 3.992 \cdot x_{1i}\right)^2 = 10.056.$$

$$s_2^2 = \frac{1}{n_2 - 2} \sum_2 \left(y_{2i} - a_2 - b_2 \cdot x_{2i}\right)^2 = \frac{1}{6 - 2} \sum_2 \left(y_{2i} - 5.78 - 3.888 \cdot x_{1i}\right)^2 = 7.027.$$

Для того, чтобы проверить, неразличимы ли регрессии, надо проверить выполнение трех условий (при заданной доверительной вероятности):

1. $s_1^2 = s_2^2$ (равенство остаточных дисперсий),
2. $a_1 = a_2$,
3. $b_1 = b_2$.

Сначала проверяется равенство остаточных дисперсий с помощью F-критерия Фишера. Если

$$\frac{s_1^2}{s_2^2} < F\left(\alpha; n_1 - 2, n_2 - 2\right),$$

то остаточные дисперсии признаются одинаковыми. Для нашей задачи

$$\frac{s_1^2}{s_2^2} = \frac{10.056}{7.027} = 1.431,$$

а

$$F\left(\alpha; n_1 - 2, n_2 - 2\right) = F\left(5\%; 8, 4\right) = 6.04.$$

При работе с таблицей учитываем, что $k_1 = 8$ соответствует большей дисперсии ($s_1^2 = 10.056$), а $k_2 = 4$ соответствует меньшей дисперсии ($s_2^2 = 7.027$). Поскольку $1.431 < 6.04$, то остаточные дисперсии s_1^2 и s_2^2 признаются статистически неразличимыми и, следовательно, можно переходить к сравнению коэффициентов регрессий.

Для сравнения b_1 и b_2 используется статистика t_b:

$$t_b = \frac{b_1 - b_2}{s^* \cdot \sqrt{\dfrac{1}{(n_1 - 1) \cdot s_{\bar{x}_1}^2} + \dfrac{1}{(n_2 - 1) \cdot s_{\bar{x}_2}^2}}},$$

где

$$s_{\bar{x}_1}^2 = \frac{1}{n_1 - 1} \sum_1 \Big(x_{1i} - \bar{x}_1 \Big)^2,$$

$$s_{\bar{x}_2}^2 = \frac{1}{n_2 - 1} \sum_2 \Big(x_{2i} - \bar{x}_2 \Big)^2,$$

$$\bar{x}_1 = \frac{1}{n_1} \sum_1 x_{1i},$$

$$\bar{x}_2 = \frac{1}{n_2} \sum_2 x_{2i},$$

$$s^* = \sqrt{\frac{(n_1 - 2) \cdot s_1^2 + (n_2 - 2) \cdot s_2^2}{n_1 + n_2 - 4}},$$

$$s_1^2 = \frac{1}{n_1 - 2} \sum_1 \Big(y_{1i} - a_1 - b_1 \cdot x_{1i} \Big)^2,$$

$$s_2^2 = \frac{1}{n_2 - 2} \sum_2 \Big(y_{2i} - a_2 - b_2 \cdot x_{2i} \Big)^2.$$

Если

$$|t_b| \le t_{(n_1 + n_2 - 4), \alpha/2\%},$$

то сравниваемые угловые коэффициенты регрессий b_1 и b_2 считаются равными и далее нужно переходить к сравнению коэффициентов a_1 и a_2.

Для нашей задачи

$$s^* = \sqrt{\frac{8 \cdot 10.056 + 4 \cdot 7.0271}{12}} = 3.008,$$

$$t_b = \frac{3.992 - 3.888}{3.008 \cdot \sqrt{\frac{1}{9 \cdot 88.988} + \frac{1}{5 \cdot 50.967}}} = 0.0078.$$

Из таблицы t-распределения Стьюдента находим

$$t_{12, 2.5\%} = 2.1788.$$

Поскольку $0.0078 < 2.1788$, то коэффициенты b_1 и b_2 признаются статистически равными с доверительной вероятностью 0.95.

Теперь проверим статистическое равенство коэффициентов a_1 и a_2. Для их сравнения используется статистика

$$t_a = \frac{\bar{b} - \tilde{b}}{\tilde{s}},$$

где

$$\bar{b} = \frac{(n_1 - 1) \cdot s^2_{\bar{x}_1} b_1 + (n_2 - 1) \cdot s^2_{\bar{x}_2} b_2}{(n_1 - 1) \cdot s^2_{\bar{x}_1} + (n_2 - 1) \cdot s^2_{\bar{x}_2}},$$

$$\tilde{b} = \frac{\bar{y}_1 - \bar{y}_2}{\bar{x}_1 - \bar{x}_2},$$

$$\tilde{s} = s^* \cdot \sqrt{\frac{1}{(n_1 - 1) \cdot s^2_{\bar{x}_1} + (n_2 - 1) \cdot s^2_{\bar{x}_2}} + \frac{1}{(\bar{x}_1 - \bar{x}_2)^2} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

Все остальные величины были определены выше.

Для рассматриваемой задачи:

$$\bar{b} = \frac{9 \cdot 88.989 \cdot 3.992 + 5 \cdot 50.967 \cdot 3.888}{9 \cdot 88.989 + 5 \cdot 50.967} = 3.967;$$

$$\tilde{b} = \frac{57.9 - 90.667}{14.1 - 21.83} = 4.239,$$

$$\tilde{s} = 3.008 \cdot \sqrt{\frac{1}{9 \cdot 88.989 + 5 \cdot 50.967} + \frac{1}{(14.1 - 21.83)^2} \cdot \left(\frac{1}{10} + \frac{1}{6}\right)} = 0.2212.$$

Статистика t_a:

$$t_a = \frac{3.967 - 4.239}{0.2212} = -1.23.$$

Из таблицы t-распределения Стьюдента, как мы уже находили, $t_{12,2.5\%} = 2.1788$. Поскольку $|-1.23| < 2.1788$, то коэффициенты a_1 и a_2 также признаются статистически равными с доверительной вероятностью 0.95.

Окончательно, обе регрессии

$$y = f_1(x) = 1.613 + 3.992 \cdot x$$

и

$$y = f_2(x) = 5.78 + 3.888 \cdot x$$

признаются статистически идентичными.

13.4 Полиномиальная регрессия

Если линейные уравнения регрессии плохо описывают статистические данные, то необходимо применять другие, более сложные модели. В первую очередь, из общего вида предполагаемой зависимости, делается попытка отыскать линеаризующее преобразование (аналогично тому, как было показано на примере при рассмотрении условных и нормальных уравнений). Это удается далеко не всегда – и кроме того, для применения методов регрессионного анализа, необходимо, чтобы функция от нормально распределенной случайной величины также оказалась нормально распределенной, [5].

Рассмотрим универсальный метод построения нелинейной регрессии. Большинство нелинейных регрессионных моделей могут быть представлены как линейные по неизвестным параметрам:

$$y = y(x, \{\Theta_i\}) = \Theta_0 f_0(x) + \Theta_1 f_1(x) + \cdots + \Theta_r f_r(x).$$

Здесь $x = \{x_1, x_2, \cdots, x_n\}, y = \{y_1, y_2, \cdots, y_n\}$ – результаты наблюдений, для которых ищется в общем случае нелинейная регрессионная связь, $\Theta_0, \Theta_1, \cdots, \Theta_r$ – неизвестные и требующие оценки параметры модели, $f_0(x), f_1(x), \cdots, f_r(x)$ – заданные функции наблюдений $\{x_i\}$.

В дальнейшем будем рассматривать разложения функции $f(x)$ только по полиномам[8] (в ряд Тейлора):

$$f_0(x) = 1, f_1(x) = x, f_2(x) = x^2, \cdots, f_r(x) = x^r.$$

При этом

$$y = \Theta_0 + \Theta_1 x + \Theta_2 x^2 + \cdots + \Theta_r x^r. \tag{10}$$

Отметим, что если ограничится первой степенью по x (т.е. $r = 1$), и оценивать четыре параметра $(\Theta_0, \Theta_1, \Theta_2, \Theta_3)$, то мы придем к задаче, разобранной в п. 10 при рассмотрении линеаризации системы условных уравнений и составлении системы нормальных уравнений. Различие только в том, что рассматриваемая здесь модель уже полиномиальная. В обоих случаях для нахождения неизвестных параметров модели используется метод наименьших квадратов (МНК).

В общем случае случайная величина $y_i (i = 1, 2, \cdots, n)$ может быть представлена как

$$y_i = \Theta_0 + \Theta_1 x_i + \Theta_2 x_i^2 + \cdots + \Theta_r x_i^r + \epsilon_i,$$

где ϵ_i – ошибки (невязки), представляющие собой (для определенности) случайные величины с одинаковой дисперсией, хотя распределение этих ошибок может не быть нормальным. Как и раньше, неизвестные параметры $\Theta_0, \Theta_1, \cdots, \Theta_r$ модели будем искать минимизацией по этим переменным суммы квадратов невязок[9]:

$$S = \sum_{i=1}^{n} \epsilon_i^2,$$

где

$$S = \sum_{i=1}^{n} \left(y_i - \Theta_0 - \Theta_1 x_i - \cdots - \Theta_r x_i^r \right)^2.$$

Обозначим, как и раньше,

$$\sum_{i=1}^{n} = \sum$$

[8] Существует много видов разложений функции $f(x)$, для которых применим нижеследующий формализм, например, разложение в ряд Фурье, когда $f_0(x) = 1/2, f_1(x) = \sin x, f_2(x) = \cos x, \cdots, f_{2r-1}(x) = \sin rx, f_{2r}(x) = \cos rx$.

[9] Здесь и далее, в сравнении с рассматриваемой ранее МНК-схемой нахождения решения условных уравнений, знаки «-» и «+» перед неизвестными параметрами вводятся для удобства дальнейших вычислений, но все Θ_i должны быть одного знака

Необходимое условие минимума есть

$$\begin{cases} \dfrac{\partial S}{\partial \Theta_0} = -2\sum\Big(y_i - \Theta_0 - \Theta_1 x_i - \cdots - \Theta_r x_i^r\Big) = 0 \\ \dfrac{\partial S}{\partial \Theta_1} = -2\sum x_i\Big(y_i - \Theta_0 - \Theta_1 x_i - \cdots - \Theta_r x_i^r\Big) = 0 \\ \cdots \\ \dfrac{\partial S}{\partial \Theta_r} = -2\sum x_i^r\Big(y_i - \Theta_0 - \Theta_1 x_i - \cdots - \Theta_r x_i^r\Big) = 0 \end{cases}.$$

Эти уравнения, как известно, можно переписать в виде нормальных уравнений, который для данного случая будут иметь вид:

$$\begin{cases} n\Theta_0 + \Theta_1\sum x_i + \cdots + \Theta_r\sum x_i^r = \sum y_i \\ \Theta_0\sum x_i + \Theta_1\sum x_i^2 + \cdots + \Theta_r\sum x_i^{r+1} = \sum x_i y_i \\ \cdots \\ \Theta_0\sum x_i^r + \Theta_1\sum x_i^{r+1} + \cdots + \Theta_r\sum x_i^{2r} = \sum x_i^r y_i \end{cases}.$$

Запишем решение в матричной форме.

Введем матрицу системы (которая в различной литературе носит также названия: *основная, конструкционная, структурная*[10]):

$$A = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^r \\ 1 & x_2 & x_2^2 & \cdots & x_2^r \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_n & x_n^2 & \cdots & x_n^r \end{pmatrix}$$

и

$$\Theta^T = \Big(\Theta_0, \Theta_1, \Theta_2, \cdots, \Theta_r\Big),$$

$$\epsilon^T = \Big(\epsilon_1, \epsilon_2, \cdots, \epsilon_n\Big).$$

В матричном обозначении

$$S = \sum_{i=1}^{n}\epsilon_i^2 = \epsilon^T\cdot\epsilon = \Big(y - A\Theta\Big)^T\cdot\Big(y - A\Theta\Big) = y^Ty - 2\Theta^TA^Ty + \Theta^TA^TA\Theta.$$

Согласно МНК нужно дифференцировать последнее равенство по всем параметрам Θ_i и приравнять результат к нулю:

$$-2A^Ty + 2A^TA\Theta = 0,$$

[10]Если $n = r$, то такая матрица называется матрицей Вандермонда

которую можно переписать в виде

$$\Big(A^T A\Big)\Theta = A^T y.$$

Решение последнего имеет вид:

$$\tilde{\Theta} = \Big(A^T A\Big)^{-1} A^T y. \tag{11}$$

13.4.1 Ортогональные полиномы и преимущества их использования

Теперь рассмотрим частный случай:

$$y = y(x, \Theta_0, \Theta_1) = \Theta_0 + \Theta_1 x.$$

Сделаем замену переменной

$$\xi = \xi(x, \Phi_0, \Phi_1) = \Phi_0 + \Phi_1\Big(x - \bar{x}\Big),$$

где как и раньше

$$\bar{x} = \frac{1}{n}\sum x_i.$$

Рассмотрим значительные преимущества такой замены переменных. Аналогично предыдущим выводам, в матричной форме уравнения МНК имеют вид

$$y = B\Phi + \epsilon,$$

где новая конструкционная матрица B есть

$$B = \begin{pmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{pmatrix}$$

Поскольку полученная модель – линейная (если положить $f_0(x) = 1, f_1(x) = x - \bar{x}$), то ее решение имеет вид, полностью аналогичный (11):

$$\tilde{\Phi} = \Big(B^T B\Big)^{-1} B^T y, \tag{12}$$

где

$$B^T B = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 - \bar{x} & x_2 - \bar{x} & \cdots & x_n - \bar{x} \end{pmatrix} \cdot \begin{pmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{pmatrix} = \begin{pmatrix} n & \sum x_i - n\bar{x} \\ \sum x_i - n\bar{x} & \sum \big(x_i - \bar{x}\big)^2 \end{pmatrix} =$$

$$= \begin{pmatrix} n & 0 \\ 0 & \sum \big(x_i - \bar{x}\big)^2 \end{pmatrix}.$$

Тогда

$$(B^TB)^{-1} = \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \left[\sum \left(x_i - \bar{x}\right)^2\right]^{-1} \end{pmatrix}.$$

Обратим внимание, что матрица B^TB – диагональная, и, следовательно, может быть легко обращена без ошибок, вызванных округлением. Это особенно важно, когда нужно проводить последовательную аппроксимацию функции $f(x)$ полиномами все более высокого порядка. Остановимся на этом моменте более подробно. Обычный полином в общем случае имеет вид (10):

$$y = \Theta_0 + \Theta_1 x + \Theta_2 x^2 + \cdots + \Theta_r x^r.$$

Такая модель приводит к матрице

$$A^TA = \begin{pmatrix} n & \sum x_i & \sum x_i^2 & \cdots & \sum x_i^r \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & \cdots & \sum x_i^{r+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_i^r & \sum x_i^{r+1} & \sum x_i^{r+2} & \cdots & \sum x_i^{2r} \end{pmatrix}, \tag{13}$$

которая плохо обусловлена при больших r. Плохая обусловленность может даже ухудшаться с ростом n. В итоге при вычислении A^TA могут возникать значительные ошибки округления. Это происходит по следующей причине. Предположим, все $\{x_i\}$ заключены в интервале от 0 до 1. И пусть множество $n\{x_i\}$ равномерно расширяется при росте $n \to \infty$. Тогда

$$\sum x_i^r \to n \int_0^1 x^r dx = \frac{n}{r+1}$$

и, следовательно,

$$A^TA \approx n \cdot \begin{pmatrix} 1 & 1/2 & 1/3 & \cdots & 1/(r+1) \\ 1/2 & 1/3 & 1/4 & \cdots & 1/(r+2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1/(r+1) & 1/(r+2) & 1/(r+3) & \cdots & 1/(2r+1) \end{pmatrix} = nH_{r+1},$$

где H_p – гильбертова матрица ранга p, которая плохо обусловлена при больших p, [10]. Чтобы избежать таких неустойчивостей решения, вместо модели (10) вводится эквивалентная ей модель, записанная в виде ортогональных полиномов:

$$\xi = \xi(x, \Phi_0, \Phi_1, \cdots, \Phi_r) = \Phi_0 + \Phi_1 G_1(x) + \Phi_2 G_2(x) + \cdots + \Phi_r G_r(x),$$

где полиномы $G_j(x) = k_j^{(0)} + k_j^{(1)}x + \cdots + k_j^{(j-1)}x^{j-1} + x^j$ (для рассмотренного выше линейного случая $G_1(x) = 1, G_2(x) = x - \bar{x}$). В общем случае коэффициенты $k_j^{(0)}, k_j^{(1)}, \cdots, k_j^{(j-1)}$ определяются из системы (метод ортогонализации Грама-Шмидта)

$$\begin{cases} \sum G_0(x_i) \cdot G_j(x_i) = 0, \\ \sum G_1(x_i) \cdot G_j(x_i) = 0, \\ \cdots \\ \sum G_{j-1}(x_i) \cdot G_j(x_i) = 0 \end{cases},$$

здесь, как и раньше, обозначено

$$\sum = \sum_{i=1}^{n},$$

а

$$G_0(x) = 1.$$

Свойство ортогональности означает, что

$$\int_0^1 G_j(x)G_m(x)dx = 0(j \neq m),$$

откуда следует, что все недиагональные элементы матрицы B^TB обращаются в ноль:

$$B^TB = \begin{pmatrix} \sum G_0^2(x_i) & \sum G_0(x_i)G_1(x_i) & \cdots & \sum G_0(x_i)G_r(x_i) \\ \sum G_1(x_i)G_0(x_i) & \sum G_1^2(x_i) & \cdots & \sum G_1(x_i)G_r(x_i) \\ \vdots & \vdots & \ddots & \vdots \\ \sum G_r(x_i)G_0(x_i) & \sum G_r(x_i)G_1(x_i) & \cdots & \sum G_r^2(x_i) \end{pmatrix} =$$

$$= \begin{pmatrix} \sum G_0^2(x_i) & 0 & \cdots & 0 \\ 0 & \sum G_1^2(x_i) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sum G_r^2(x_i) \end{pmatrix}_{(r+1)(r+1)}$$

Обращение диагональной матрицы приводит к меньшим ошибкам округления. Кроме того, в диагональном представлении гораздо легче сравнивать аппроксимации r и $r+1$ порядка, поскольку требуется вычислить только один элемент $\sum G_{r+1}^2(x_i)$.

13.4.2 Ортогональные нормированные полиномы и преимущества их использования

Следующий шаг в задаче упрощения аппроксимации – использование не просто ортогональных, но *нормированных ортогональных полиномов*: использование вместо метода Грама-Шмидта метода Форсайта, который заключается в установлении простого рекуррентного соотношения между ортогональными нормированными полиномами вида:

$$Q_j(x) = \frac{G_j(x)}{\sqrt{\sum_{i=1}^{n} G_j^2(x_i)}},$$

и

$$Q_0(x) = \frac{1}{\sqrt{n}},$$

$$\sum_{i=1}^{n} Q_j^2(x_i) = 1.$$

Пусть разложение функции $f(x)$ по ортогональным нормированным полиномам имеет вид:

$$\tilde{f}(x) = \omega_0 Q_0(x) + \omega_1 Q_1(x) + \cdots + \omega_r Q_r(x). \tag{14}$$

Для $Q_i(x)$ выполняется рекуррентное соотношения:

$$\lambda Q_j(x) = xQ_{j-1}(x) - \alpha Q_{j-1}(x) - \beta Q_{j-2}(x),$$

где постоянные α и β определяются из уравнений

$$\alpha = \sum x_i Q_{j-1}^2(x_i),$$

$$\beta = \sum x_i Q_{j-1}^2(x_i) Q_{j-2}(x_i).$$

Постоянная λ определяется из условия

$$\sum Q_j^2(x_i) = 1.$$

Применяя МНК-метод к выражению (14), получаем формулу, аналогичную выражению (12):

$$\tilde{\omega} = \left(B^T B\right)^{-1} B^T y = B^T y,$$

поскольку для ортогональных нормированных полиномов: $A^T A = I$. Матрица B есть:

$$B = \begin{pmatrix} Q_0(x_1) & Q_1(x_1) & \cdots & Q_r(x_1) \\ Q_0(x_2) & Q_1(x_2) & \cdots & Q_r(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ Q_0(x_n) & Q_1(x_n) & \cdots & Q_r(x_n) \end{pmatrix}$$

Решение обладает тем свойством, что

$$D[\tilde{\omega}_j] = D\left[\sum y_i Q_j(x_i)\right] = \sigma^2,$$

поскольку

$$D\left[\sum y_i Q_j(x_i)\right] = \sum Q_j^2(x_i) D[y_i] = \sigma^2 \sum Q_j^2(x_i) = \sigma^2.$$

Вычислив $\tilde{\omega}_j$, получаем аппроксимационный полином (14):

$$\tilde{f}(x) = \omega_0 Q_0(x) + \omega_1 Q_1(x) + \cdots + \omega_r Q_r(x) = \tilde{\omega}_0 Q_0(x) + \tilde{\omega}_1 Q_1(x) + \cdots + \tilde{\omega}_r Q_r(x).$$

Остаточная сумма квадратов (невязка или остаточная погрешность) при аппроксимации нормальным ортогональным полиномом степени r есть

$$\epsilon_r^2 = \sum_{i=1}^{n} y_i^2 - \sum_{j=0}^{r} \left(\sum_{i=1}^{n} y_i Q_j(x_i)\right)^2,$$

а остаточная дисперсия

$$\sigma_\epsilon^2 = \frac{\epsilon_r^2}{n - r - 1}.$$

13.4.3 Правила вычисления ортонормальных полиномов Чебышева на дискретном наборе точек

Итак, коэффициенты разложения $\tilde{\omega}_i$ по нормированным ортогональным (или *ортонормальным*) полиномам $Q_k(x)$ определяются значениями $\{y_i\}$ и матрицей B, зависящей только от значений ортонормальных полиномов на наборе точек $x = \{x_i\}$. Таким образом, задача построения нелинейной регрессии сводится к тому, чтобы вычислить значения ортонормального полинома на заданном наборе точек. Существует много разных ортонормальных полиномов. Для примера, рассмотрим полиномы Чебышева, [11].

Особенностью использования полиномов Чебышева является требование того, чтобы все точки $x = \{x_i\}$ были *равноотстоящими*.

Сначала рассмотрим систему ортогональных, но ненормированных полиномов Чебышева. Пусть дана система $n+1$ равноотстоящих точек $x = \{x_i\}(i = 0, 1, 2, \cdots, n)$. С помощью линейного преобразования $t = (x - x_0)/h$ переведем эти точки соответственно в $t = 0, 1, 2, \cdots, n$.

Полиномы $P_{0,n}(t), P_{1,n}(t), \cdots, P_{m,n}(t)(m \leq n)$ соответственно степеней $0, 1, \cdots, m$, ортогональные на множестве точек $\{0, 1, 2, \cdots, n\}$ и отличные от нуля на этом множестве, называются *ортогональными полиномами Чебышева*. Первый индекс в $P_{k,n}(t)$, k – степень полинома, а второй индекс n – число точек, уменьшенное на единицу.

Полиномы Чебышева задаются формулой:

$$P_{k,n}(t) = \sum_{s=0}^{n} \left(-1\right)^s C_k^s C_{k+s}^s \cdot \frac{t^{[s]}}{n^{[s]}},$$

$$(k = 0, 1, 2, \cdots, m)$$

где

$$C_n^k = \frac{n!}{k!(n-k)!},$$

а $t^{[s]} = t(t-1)\cdots(t-s+1)$ и $n^{[s]} = n(n-1)\cdots(n-s+1)$ есть соответствующие т.н. *обобщенные степени*.

Несколько первых ортогональных полиномов Чебышева есть:

$$P_{0,n}(t) = 1,$$

$$P_{1,n}(t) = 1 - 2 \cdot \frac{t}{n}$$

$$(n \geq 1),$$

$$P_{2,n}(t) = 1 - 6 \cdot \frac{t}{n} + 6 \cdot \frac{t(t-1)}{n(n-1)}$$

$$(n \geq 2),$$

$$P_{3,n}(t) = 1 - 12 \cdot \frac{t}{n} + 30 \cdot \frac{t(t-1)}{n(n-1)} - 20 \cdot \frac{t(t-1)(t-2)}{n(n-1)(n-2)}$$

$$(n \geq 3),$$

$$P_{4,n}(t) = 1 - 20 \cdot \frac{t}{n} + 90 \cdot \frac{t(t-1)}{n(n-1)} - 140 \cdot \frac{t(t-1)(t-2)}{n(n-1)(n-2)} + 70 \cdot \frac{t(t-1)(t-2)(t-3)}{n(n-1)(n-2)(n-3)}$$

$$(n \geq 4).$$

Возвращаясь к прежней переменной x, получим систему полиномов, ортогональных на дискретном множестве $x = \{x_i\}$:

$$P_{k,n}\left(\frac{x - x_0}{h}\right)$$

$$(k = 0, 1, \ldots, m; m \leq n)$$

Система полиномов $\{P_{k,n}(t)\}$ не является нормированной. Построим соответствующую нормированную систему и получим ортонормальные полиномы Чебышева. Определим *норму* для $\{P_{k,n}(t)\}$ следующим образом, [11]:

$$\left|\left|P_{k,n}(t)\right|\right|^2 = \sum_{i=0}^{n} P_{k,n}^2(i) = \frac{\left(n + k + 1\right)^{[k+1]}}{\left(2k + 1\right) \cdot n^{[k]}}.$$

Разделив многочлены $P_{k,n}(t)$ на их нормы, получим *нормированную систему ортогональных полиномов Чебышева* или *ортонормальную систему полиномов Чебышева*:

$$\tilde{P}_{k,n}(t) = \frac{P_{k,n}(t)}{\left|\left|P_{k,n}(t)\right|\right|}$$

$$(k = 0, 1, 2, 3 \cdots, m; m \leq n).$$

ПРИМЕР Рассмотрим на конкретном примере, как строить ортонормальные полиномы Чебышева. Пусть задана система точек:

$$x_0 = 1/2; x_1 = 1; x_2 = 3/2; x_3 = 2; x_4 = 5/2; x_5 = 3.$$

Получим систему полиномов до третьей степени включительно, ортонормальную (или *ортонормированную*) на данной системе точек. Отметим, что это возможно сделать, поскольку точки эквидистантны (расстояние между двумя любыми соседними точками есть $h = 1/2$).

Для решения задачи введем замену переменных:

$$t = \frac{x - x_0}{h} = \frac{x - 1/2}{1/2} = 2 \cdot \left(x - 1/2\right),$$

при которой все x_i переходят в целочисленные $t = 0, 1, 2, 3, 4, 5$. В общей формуле

$$P_{k,n}(t) = \sum_{s=0}^{n} \left(-1\right)^s C_k^s C_{k+s}^s \cdot \frac{t^{[s]}}{n^{[s]}},$$

примем $n = 5$, тогда

$$P_{k,5}(t) = \sum_{s=0}^{5} \left(-1\right)^s C_k^s C_{k+s}^s \cdot \frac{t^{[s]}}{5^{[s]}}.$$

где $k = 0, 1, 2, 3$, поскольку по условию ищется аппроксимирующий полином степени $m = 3$. Проведем вычисления для всех указанных k:

$$P_{0,5}(t) = \sum_{s=0}^{5} \left(-1\right)^s C_0^s C_s^s \cdot \frac{t^{[s]}}{5^{[s]}} = 1,$$

$$P_{1,5}(t)=\sum_{s=0}^{5}\Big(-1\Big)^s C_1^s C_{1+s}^s\cdot\frac{t^{[s]}}{5^{[s]}}=1-0.4\cdot t,$$

$$P_{2,5}(t)=\sum_{s=0}^{5}\Big(-1\Big)^s C_2^s C_{2+s}^s\cdot\frac{t^{[s]}}{5^{[s]}}=1-1.2\cdot t+0.3\cdot t(t-1),$$

$$P_{3,5}(t)=\sum_{s=0}^{5}\Big(-1\Big)^s C_3^s C_{3+s}^s\cdot\frac{t^{[s]}}{5^{[s]}}=1-2.4\cdot t+1.5\cdot t(t-1)-0.333\cdot t(t-1)(t-2).$$

Нормы функций $P_{k,5}(t)$, где $k=0,1,2,3$, вычисляем по формуле

$$\Big|\Big|P_{k,n}(t)\Big|\Big|^2=\frac{\Big(n+k+1\Big)^{[k+1]}}{\Big(2k+1\Big)\cdot n^{[k]}}:$$

$$\Big|\Big|P_{0,5}(t)\Big|\Big|=\sqrt{6},$$

$$\Big|\Big|P_{1,5}(t)\Big|\Big|=\sqrt{\frac{7\cdot 6}{3\cdot 5}}=\sqrt{\frac{14}{5}},$$

$$\Big|\Big|P_{2,5}(t)\Big|\Big|=\sqrt{\frac{8\cdot 7\cdot 6}{5\cdot 5\cdot 4}}=\frac{2}{5}\sqrt{21},$$

$$\Big|\Big|P_{3,5}(t)\Big|\Big|=\sqrt{\frac{9\cdot 8\cdot 7\cdot 6}{7\cdot 5\cdot 4\cdot 3}}=\frac{6}{\sqrt{5}}.$$

Теперь разделим полиномы $P_{k,5}(t)$ на их нормы и перейдем от переменной t к исходной переменной x. таким образом, получим искомую ортонормальную систему полиномов Чебышева:

$$\tilde{P}_{0,5}(x)=\frac{1}{\sqrt{6}}=0.408,$$

$$\tilde{P}_{1,5}(x)=\sqrt{\frac{5}{14}}\Big(1-0.8\cdot(x-1/2)\Big)=0.837-0.478\cdot x,$$

$$\tilde{P}_{2,5}(x)=\frac{5}{2\sqrt{21}}\Big(1-2.4\cdot(x-1/2)+0.6\cdot(x-1/2)(x-3/2)\Big)=1.446-1.964\cdot x+0.327\cdot x^2,$$

$$\tilde{P}_{3,5}(x)=\frac{\sqrt{5}}{6}\Big(1-4.8\cdot(x-1/2)+3\cdot(x-1/2)(x-3/2)-0.666\cdot(x-1/2)(x-3/2)(x-5/2)\Big)=$$

$$=2.571-5.452\cdot x+2.235\cdot x^2-0.248\cdot x^3.$$

13.4.4 Нахождение уравнения регрессии с помощью ортонормальных полиномов Чебышева и определение порядка нелинейности с заданной доверительной вероятностью

Если функция $y=f(x)$ задана на множестве узлов $x=\{x_0,x_1,\cdots,x_n\}$ с шагом h, то наилучший (по методу МНК) аппроксимирующий полином ищется в виде

$$\tilde{f}(x)=\sum_{k=0}^{m}\omega_k\cdot P_{k,n}\Big(\frac{x-x_0}{h}\Big),$$

где коэффициенты ω_k называются *коэффициентами Фурье* функции $f(x)$ относительно системы ортогональных полиномов Чебышева $P_{k,n}((x-x_0)/h)$ $(k=0,1,2,\cdots,m)$:

$$\omega_k = \frac{\sum_{i=0}^{n} y_i \cdot P_{k,n}(i)}{\left|\left|P_{k,n}(t)\right|\right|^2}.$$

В частности, если система полиномов не только ортогональна, но и ортонормальна, то

$$\tilde{f}(x) = \sum_{k=0}^{m} \tilde{\omega}_k \cdot \tilde{P}_{k,n}\left(\frac{x-x_0}{h}\right),$$

где коэффициенты $\tilde{\omega}_k$, определяемые выше как элементы матрицы $B^T y$, есть:

$$\tilde{\omega}_k = \sum_{i=0}^{n} y_i \cdot \tilde{P}_{k,n}(i).$$

Остается вопрос о статистическом критерии выбора степени аппроксимирующего полинома m.

Остаточная дисперсия при аппроксимации ортонормальными полиномами Чебышева степени m есть

$$\epsilon_m^2 = \frac{\sum_{i=0}^{n} y_i^2 - \sum_{j=0}^{m}\left(\sum_{i=0}^{n} y_i \tilde{P}_{j,n}(x_i)\right)^2}{(n+1)-m-1}.$$

Так, если

$$\frac{\epsilon_{m+1}^2}{\epsilon_m^2} > 1,$$

то в качестве регрессии принимается полином степени m. Значимость отличия остаточных дисперсий на каждом шаге увеличения степени полинома дается критерием Фишера $F(\gamma; n-m, n-m-1)$ с доверительной вероятностью γ. Так, если

$$\frac{\epsilon_1^2}{\epsilon_2^2} > F\left(\gamma; (n+1)-2, (n+1)-2-1\right),$$

то полином второй степени $(m=2)$ предпочтительнее полинома первой степени (т.е. квадратичная регрессия предпочтительнее линейной). Если

$$\frac{\epsilon_2^2}{\epsilon_3^2} > F\left(\gamma; (n+1)-3, (n+1)-3-1\right),$$

то полином третьей степени $(m=3)$ предпочтительнее полинома второй степени.

Проиллюстрируем сказанное примером.

Пусть в результате наблюдений получены следующие пары $\{x_i, y_i\}$ (см. Таблица (18)).

Поскольку $x_{i+1} - x_i = h = 2.1$, введем замену переменной:

$$t = \frac{x-1.1}{2.1}$$

Таблица 18:

Данные наблюдений, для которых нужно выбрать подходящую регрессионную модель

x_i	1.1	3.2	5.3	7.4	9.5	11.6	13.7	15.8	17.9	20.0
y_i	1.3	4.75	6.8	1.86	-15.6	-51.1	-110.3	-198.6	-321.8	-485.2

Поскольку все точки эквидистантны, построим на множестве точек $\{x_i\}(i = 1, 2, \cdots, 10)$ систему ортогональных полиномов Чебышева ($n = 9$, т.к. нумерация полиномов начинается с нулевого индекса):

$$P_{0,9}(t) = 1,$$

$$P_{1,9}(t) = 1 - 2 \cdot \frac{t}{9}$$

$$P_{2,9}(t) = 1 - 6 \cdot \frac{t}{9} + 6 \cdot \frac{t(t-1)}{9 \cdot 8}$$

$$P_{3,9}(t) = 1 - 12 \cdot \frac{t}{9} + 30 \cdot \frac{t(t-1)}{9 \cdot 8} - 20 \cdot \frac{t(t-1)(t-2)}{9 \cdot 8 \cdot 7}$$

$$P_{4,9}(t) = 1 - 20 \cdot \frac{t}{9} + 90 \cdot \frac{t(t-1)}{9 \cdot 8} - 140 \cdot \frac{t(t-1)(t-2)}{9 \cdot 8 \cdot 7} + 70 \cdot \frac{t(t-1)(t-2)(t-3)}{9 \cdot 8 \cdot 7 \cdot 6}.$$

Для построения ортонормальной системы вычислим соответствующие нормы:

$$\left|\left|P_{0,9}(t)\right|\right| = \sqrt{10},$$

$$\left|\left|P_{1,9}(t)\right|\right| = \sqrt{\frac{11 \cdot 10}{3 \cdot 9}} = \sqrt{\frac{110}{27}},$$

$$\left|\left|P_{2,9}(t)\right|\right| = \sqrt{\frac{12 \cdot 11 \cdot 10}{5 \cdot 9 \cdot 8}} = \sqrt{\frac{11}{3}},$$

$$\left|\left|P_{3,9}(t)\right|\right| = \sqrt{\frac{13 \cdot 12 \cdot 11 \cdot 10}{7 \cdot 9 \cdot 8 \cdot 7}} = \sqrt{\frac{715}{147}},$$

$$\left|\left|P_{4,9}(t)\right|\right| = \sqrt{\frac{14 \cdot 13 \cdot 12 \cdot 11 \cdot 10}{9 \cdot 9 \cdot 8 \cdot 7 \cdot 6}} = \sqrt{\frac{715}{81}}.$$

Окончательно ортонормальная система полиномов (до 4-го порядка) имеет вид:

$$\tilde{P}_{0,9}(t) = 0.316,$$

$$\tilde{P}_{1,9}(t) = 0.495 - 0.110 \cdot t,$$

$$\tilde{P}_{2,9}(t) = 0.522 - 0.392 \cdot t + 0.044 \cdot t^2,$$

$$\tilde{P}_{3,9}(t) = 0.453 - 0.829 \cdot t + 0.243 \cdot t^2 - 0.018 \cdot t^3,$$

$$\tilde{P}_{4,9}(t) = 0.337 - 1.495 \cdot t + 0.956 \cdot t^2 - 0.232 \cdot t^3 + 0.023 \cdot t^4.$$

Вычислим остаточные дисперсии для $m = 1, 2, 3, 4$ по формуле:

$$\epsilon_m^2 = \frac{\sum\limits_{i=0}^{n} y_i^2 - \sum\limits_{j=0}^{m} \left(\sum\limits_{i=0}^{n} y_i \tilde{P}_{j,n}(x_i) \right)^2}{(n+1) - m - 1}.$$

$$\epsilon_1^2 = 7707.451,$$
$$\epsilon_2^2 = 479.190,$$
$$\epsilon_3^2 = 11.580,$$
$$\epsilon_4^2 = 5.358.$$

С увеличением степени аппрокисмирующего полинома остаточная дисперсия падает, т.е. аппроксимация становится все точнее. Сравним, значимы ли различия остаточных дисперсий:

$$\frac{\epsilon_1^2}{\epsilon_2^2} = 16.084 > F(0.95; 8, 7) = 3.73,$$

$$\frac{\epsilon_2^2}{\epsilon_3^2} = 41.381 > F(0.95; 7, 6) = 4.21,$$

но

$$\frac{\epsilon_3^2}{\epsilon_4^2} = 2.161 < F(0.95; 6, 5) = 4.95,$$

что означает, что различие остаточных дисперсий при аппроксимации полиномами 3-го и 4-го порядков незначимо. Отсюда следует, что наблюдательные данные аппроксимируются нелинейной регрессионной моделью в виде полинома 3-го порядка с доверительной вероятностью 0.95.

Найдем этот полином по формуле

$$\tilde{f}(x) = \sum_{k=0}^{3} \tilde{\omega}_k \cdot \tilde{P}_{k,9}\left(\frac{x - 1.1}{2.1} \right),$$

где коэффициенты $\tilde{\omega}_k$ есть:

$$\tilde{\omega}_k = \sum_{i=0}^{9} y_i \cdot \tilde{P}_{k,9}(i).$$

Окончательно получаем

$$\tilde{f}(x) = 0.688 - 0.274 \cdot x + 0.799 \cdot x^2 - 0.100 \cdot x^3.$$

Заметим также, что много задач с решениями по построению различных регрессионных моделей можно найти в сборнике под. ред. А.В. Ефимова и А.С. Поспелова [12] в параграфе «Элементы регрессионного анализа и метод наименьших квадратов».

14 Исследование вида распределения

Пусть $\{x_i\}$, где $i = 1, 2, \cdots n$, есть выборка наблюдений случайной величины X. Пусть ставится задача проверить (с заданной доверительной вероятностью), что функция распределения генеральной совокупности, к которой принадлежит данная выборка, есть $F(x)$. В прикладных задачах чаще всего проверяется, является ли генеральная совокупность распределенной по нормальному закону. Также много задач на проверку соответствия распределению Пуассона. Рассмотрим алгоритм проверки на соответствие функции $F(x)$ общего вида, а далее, в примерах, конкретизируем вид функции $F(x)$.

14.1 Критерий χ^2 («хи-квадрат»)

Алгоритм проверки соответствия случайной выборки заданной функции распределения строится с помощью критерия χ^2, [12].

По выборке наблюдений находят оценки неизвестных параметров предполагаемого закона распределения случайной величины X. Далее, область возможных значений случайной величины X разбивается на r подмножеств $\Delta_1, \Delta_2, \cdots, \Delta_r$, например, r интервалов в случае, когда X – непрерывная случайная величина, или r групп, состоящих из отдельных значений, для дискретной случайной величины X.

Пусть n_k – число элементов выборки, принадлежащих множеству Δ_k, где $k = 1, 2, \cdots, r$. Общее число всех элементов всех выборок есть n, поэтому

$$\sum_{k=1}^{r} n_k = n.$$

Используя предполагаемый закон распределения случайной величины X, находят вероятности p_k того, что значение X принадлежит множеству Δ_k:

$$p_k = P(X \in \Delta_k)$$

$$(k = 1, 2, \cdots, r).$$

Очевидно, $\sum_{k=1}^{r} p_k = 1$.

Полученные результаты можно представить в виде Таблицы (19).

Выборочное значение статистики χ^2-критерия есть:

$$\tilde{\chi}^2 = \sum_{k=1}^{r} \frac{\Big(n_k - n \cdot p_k\Big)^2}{n \cdot p_k}.$$

Пусть задана доверительная вероятность γ. Тогда предложенный закон распределения соответствует генеральной совокупности исследуемой выборки, если выполняется неравенство

$$\tilde{\chi}^2 < \chi^2_\gamma\Big(r - l - 1\Big),$$

где $\chi^2_\gamma(r - l - 1)$ – γ-квантиль (или процентная точка $\alpha = (1 - \gamma)$) распределения χ^2 с $r - l - 1$ степенями свободы, а l – число неизвестных параметров, которые оцениваются по выборке (два параметра μ и σ^2 для сравнения с нормальным распределением, один параметр λ для сравнения с распределением Пуассона и т.д.).

Таблица 19:

Оформление элементов выборки для проверки на соответствие заданной функции распределения

Интервал	Δ_1	Δ_2	$\cdots$	Δ_r	Контрольная сумма элементов
Число наблюдений	n_1	n_2	$\cdots$	n_r	$\sum_{i=1}^{r} n_i = n$
Ожидаемое число наблюдений	np_1	np_2	$\cdots$	np_r	$\sum_{i=1}^{r} n \cdot p_i = n$

Отметим важный момент, что χ^2-критерий использует тот факт, что случайная величина $(n_k - n \cdot p_k)/\sqrt{n \cdot p_k}$, где $k = 1, 2, \cdots, r$, имеет распределение, близкое к стандартному нормальному. Чтобы это утверждение было достаточно точным, рекомендуется, чтобы для всех интервалов выполнялось условие

$$n \cdot p_k \geq 5.$$

Если в некоторых интервалах это условие не выполняется, то эти интервалы следует объединить с соседними до выполнения этого условия.

Рассмотрим пример проверки выборки случайных элементов на соответствие конкретным распределением.

ПРИМЕР Пример проверки выборки на соответствие распределению Пуассона. В первых двух столбцах Таблицы (20) приведены данные об отказах аппаратуры за 10^4 часов работы. Общее число обследованных экземпляров аппаратуры $n = 757$, при этом наблюдался $0 \cdot 427 + 1 \cdot 235 + 2 \cdot 72 + 3 \cdot 21 + 4 \cdot 1 + 5 \cdot 1 = 451$ отказ. Проверить, распределено ли число отказов по закону Пуассона, приняв доверительную вероятность 0.99:

$$p_k = P(X = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}$$

$$(k = 0, 1, 2, \cdots)$$

Оценка параметра λ равна среднему числу отказов:

$$\bar{\lambda} = \frac{451}{757} \approx 0.6.$$

Для $\lambda = 0.6$ вычисляем вероятности p_k и ожидаемое число случаев с k отказами (третий и четвертый столбцы Таблицы (20)).

Для $k = 4, 5$ и 6 значения $n \cdot p_k < 5$, поэтому объединим эти строки со строкой для $k = 3$. В результате полуим значения, указанные в Таблице (21). Так как по выборке

Таблица 20:

Оформление элементов выборки для проверки на соответствие функции распределения Пуассона

Число отказов	Количество случаев, в которых наблюдалось k отказов, n_k	$p_k = \frac{0.6^k}{k!} \cdot e^{-0.6}$	Ожидаемое число случаев с k отказами, $n \cdot p_k$
0	427	0.54881	416
1	235	0.32929	249
2	72	0.09879	75
3	21	0.01976	15
4	1	0.00296	2
5	1	0.00036	0
≥ 6	0	0.00004	0
Сумма	757		

Таблица 21:

Оформление элементов выборки для проверки на соответствие функции распределения Пуассона после малочисленных обьединения интервалов

k	n_k	$n \cdot p_k$	$\frac{(n_k - n \cdot p_k)^2}{n \cdot p_k}$
0	427	416	0.291
1	235	249	0.787
2	72	75	0.120
≥ 3	23	17	2.118
			$\tilde{\chi}^2 = 3.316$

Таблица 22:

Выборка, проверяемая на соответствие нормальному закону распределения $n = 55$

20.3	15.4	17.2	19.2	23.3	18.1	21.9
15.3	16.8	13.2	20.4	16.5	19.7	20.5
14.3	20.1	16.8	14.7	20.8	19.5	15.3
19.3	17.8	16.2	15.7	22.8	21.9	12.5
10.1	21.1	18.3	14.7	14.5	18.1	18.4
13.9	19.1	18.5	20.2	23.8	16.7	20.4
19.5	17.2	19.6	17.8	21.3	17.5	19.4
17.8	13.5	17.8	11.8	18.6	19.1	

оценивается один параметр λ, то $l = 1$ и число степеней свободы равно $4 - 1 - 1 = 2$. По таблице χ^2-распределения находим $\chi^2_{0.99}(2) = 9.21$, следовательно, принимается предположение о распределении Пуассона.

ПРИМЕР Пример проверки выборки на соответствие нормальному распределению. Дана выборка из 55 наблюдений (см. Таблица (22)). Размах выборки $J_n(x) = x^*_n - x^*_1 = 23.8 - 10.1 = 13.7$. Длина интервала группировки $b = 13.7/7 \approx 2$. В качестве первого интервала удобно взять интервал $[10, 12)$.

Результаты группировки сведены в Таблицу (23). В четвертом столбце Таблицы (23) приводятся вероятности, вычисленные по формуле:

$$p_k = P(X \in \Delta_k) = \Phi\left(\frac{\beta_k - \bar{x}}{s}\right) - \Phi\left(\frac{\alpha_k - \bar{x}}{s}\right)$$

$$(k = 1, 2, 3, 4, 5, 6, 7)$$

Здесь α_k и β_k – соответственно нижняя и верхняя границы интервалов, а значения функции Лапласа-Гаусса берутся из соответствующей статистической таблицы.

Поскольку после объединения осталось $r = 5$ интервалов, а по выборке оценены два параметра ($\bar{x}$ и s), т.е. $l = 2$, то число степеней свободы равно $5 - 2 - 1 = 2$. Задаваясь статистической вероятностью 0.90, по статистической таблице χ^2-распределения находим $\chi^2_{0.90}(2) = 4.61$. Выборочное значение статистики критерия есть $\tilde{\chi}^2 = 0.928 < 4.61$, следовательно, предположение о нормальном распределении верно.

Таблица 23:

Выборка, проверяемая на соответствие нормальному закону распределения $n = 55$ (во втором и третьем столбце приведены результаты группировки по k интервалам, в четвертом столбце приведены вычисленные вероятности (см. текст). В пятом столбце приводятся ожидаемые частоты, а в шестом – значения ожидаемых частот после объединения первых двух и последних двух интервалов).

k	Δ_k	Наблюдаемая частота n_k	Вероятность попадания в интерв. Δ_k (p_k)	Ожидаемая частота $n \cdot p_k$	$n \cdot p_k$	$n_k - n \cdot p_k$	$\frac{(n_k - n \cdot p_k)^2}{n \cdot p_k}$
1	$(-\infty, 12)$	2	0.0228	1.254			
2	$[12, 14)$	4	0.0731	4.020	5.274	0.725	0.010
3	$[14, 16)$	8	0.1686	9.273	9.273	-1.273	0.175
4	$[16, 18)$	12	0.2576	14.168	14.168	-2.168	0.332
5	$[18, 20)$	16	0.2484	13.662	13.662	-2.338	0.400
6	$[20, 22)$	10	0.1519	8.354	12.663	0.366	0.011
7	$[22, +\infty)$	3	0.0778	4.279			
	Сумма	55	1.0001	55	55		0.928

15 Непараметрические критерии сравнения распределений

Для сравнения двух выборок, законы распределения которых неизвестны или они сильно отличаются от хорошо известных законов распределения, а также при анализе малых выборок (с числом элементов $n < 10$) используют *непараметрические статистические методы*. Основная идея этих методов – это сравнение параметров положения и параметров масштаба двух выборок (т.е. количественный анализ того, как сильно смещены средние двух выборок друг относительно друга и как сильно распределения выборок искажены друг относительно друга).

Ранговые критерии – одни из самых эффективных методов непараметрической статистики (эффективность лучших из них составляет до 95% от мощности t-критерия Стьюдента и сопоставима с последним для случая больших выборок, [7]; далее эффективность везде указывается относительно t-критерия). Они основываются на использовании рангов, приписываемых значениям случайных величин в общей упорядоченной по возрастанию выборке (т.е. в упорядоченном ряду чисел $x_1 \leq x_2 \leq \ldots \leq x_n$ значению x_i приписывается ранг R_i). При этом одинаковым величинам присваивается усредненный ранг. Таким образом, анализируются не сами значения выборок, но их ранги.

Рассмотрим несколько ранговых критериев: быстрый ранговый критерий, критерий Ван дер Вардена и критерий Манна-Уитни-Вилкоксона, частным случаем которого является аппроксимация Имана.

Быстрый ранговый критерий

Быстрый (грубый) ранговый критерий – самый простой с точки зрения вычислительной сложности. Эффективность метода составляет более 86% для любых распределений, отличных от нормального, и 96% для нормальных распределений.

Элементы двух выборок объемами n и m соответственно, записываются как единая выборка. Далее, для проверки, смещены ли средние, объединенная выборка записывается в виде вариационного ряда: $x_1 \le x_2 \le x_i \le ... \le x_{n+m}$ где $i = R_i$ – порядковый номер элемента считается равным его рангу. Для проверки, есть ли смешение масштабов, объединенная выборка записывает так же в виде вариационного ряда $x_1 \le x_2 \le x_i \le ... \le x_{n+m}$, а затем элементы перестанавливаются следующим образом: $x_1, x_n, x_{n-1}, x_2, x_3, x_{n+m-2}, x_{n+m-3}, x_4, x_5, ...$

В обоих случаях статистика критерия $d^* = d/s_d$ – величина, имеющая стандартное нормальное распределение

$$d^* \sim N(0,1),$$

где

$$s_d = \sqrt{(\sum R_{(1)} + \sum R_{(2)}) \cdot (1/n + 1/m)/6},$$

$$d = \sum R_{(1)}/n - \sum R_{(2)}/m,$$

а $\sum R_{(1)}$ и $\sum R_{(2)}$ есть суммы рангов первой и второй выборки соответственно, в объединенной выборке.

Далее, если вычисленная статистика $|d^*| > u_{(\gamma+1)/2}$, где $u_{(\gamma+1)/2}$ есть табличное значение квантили нормального распределения, то для двух выборок признается отличие в среднем значении (или, соответственно, в параметре масштаба). Следовательно, выборки признаются различными. Если это равенство не выполнено (как для первого, так и для второго случая), то выборки признаются одинаковыми, хотя рекомендуется дополнительное исследование.

Критерий Ван дер Вардена

Критерий Ван-дер-Вардена основан на алгоритме перехода к статистике X-критерия и различен для выборок средних и малых объемов.

Для малых выборок статистика Ванн-дер-Вардена имеет вид

$$X_m = \sum_{j=1}^{m_2} u_{\gamma_j},$$

где u_{γ_j} есть γ_j-квантиль стандартного нормального распределения (суммирование можно вести относительно m_1, результат не меняется). Величину u_γ можно определить по приближенной формуле

$$u_{\gamma_j} \approx 4.91 \left[\gamma_j^{0.14} - (1-\gamma_j)^{0.14}\right],$$

где $\gamma_j = R_l/(m_1 + m_2 + 1)$, $j = \overline{1, m_2}$.

Для больших выборок статистику Ван-дер-Вардена можно аппроксимировать нормальным распределением со средним $m = 0$ и дисперсией

$$D(X) = \sqrt{\frac{n_1 n_2}{(n_1+n_2)(n_1+n_2-1)} \sum_{p=1}^{n_1+n_2} u_{\gamma_p}^2},$$

где $\gamma_p = R_p/(n_1+n_2+1)$, $p = \overline{1,n}$.

Метод обладает высокой эффективностью (его мощность равна мощности t-критерия Стьюдента) только для больших выборок.

Критерий Манна-Уитни-Вилкоксона

Критерий Манна-Уитни-Вилкоксона основан на U-статистике Манна-Уитни и R-статистике Вилкоксона. Его эффективность 95%. U-статистика определяется как

$$U = \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} h_{ij}, \qquad \text{где} \qquad h_{ij} = \begin{cases} 1, & \text{если } x_i < y_j, \\ 0, & \text{если } x_i > y_j. \end{cases}$$

Здесь k_1, k_2 – объемы двух выборок в рассматриваемой группе; $i = \overline{1,k_1}, j = \overline{1,k_2}$. Для вычисления U необходимо подсчитать количество элементов первой выборки, не превосходящих по своему значению случайные величины из второй выборки. При этом не важно, относительно какой из выборок ведётся суммирование. В случае малых выборок гипотеза сдвига отклоняется, если найденная величина U не входит в числовой интервал, определяемый критическими значениями U_1 и U_2.

Для выборки большего объема лучшую оценку дает R-статистика:

$$R = n_1 n_2 + \frac{n_2(n_2+1)}{2} - U,$$

которая аппроксимируется W-распределением:

$$W = \frac{R - n_2(n_1+n_2+1)/2}{g}.$$

Величина g в знаменателе:

$$g = \left[\frac{n_1 n_2 (n_1+n_2+1)}{12}\left(1 - \frac{\sum\limits_{s=1}^{q} t_s(t_s^2-1)}{(n_1+n_2)(n_1+n_2+1)(n_1+n_2-1)}\right)\right]^{0.5},$$

где s – количество групп, значения элементов в которых одинаковы, t_s – количество элементов в каждой группе, q – общее число групп. При этом элементы, численно равные друг другу, но принадлежащие разным выборкам, не учитываются. Величина g учитывает совпадающие элементы в выборках, благодаря чему, полученное распределение может быть аппроксимировано нормальным.

Аппроксимация Имана

Одним из наиболее точных наряду с рассмотренными непараметрическими методами является метод аппроксимации Имана. Его эффективность 95%. Соответствующая J-статистика строится на основе W-статистики Вилкоксона:

$$J = \frac{W}{2}\left[1 + \left(\frac{n_1+n_2-2}{n_1+n_2-1-W^2}\right)^{0.5}\right].$$

Далее полученное значение сравнивается с критическим: $J_{\alpha'} = (z_{\alpha'} + t_{\alpha'})/2$. Здесь $z_{\alpha'}$ и $t_{\alpha'}$ есть α'-квантили нормального распределения и распределения Стьюдента с $r = n_1+n_2-2$ степенями свободы соответственно.

ПРИМЕР Рассмотрим две однородные выборки (т.е. принадлежащие одной генеральной совокупности и, следовательно, статистически одинаковые), Таблица (24). Как можно видеть из Таблицы (25), использованные для анализа критерии однозначно подтверждают одинаковость распределения случайных величин в тестовых выборках, что указывает на устойчивость работы критериев.

Таблица 24: Однородные выборки

$n_1 = 31$	$m_1 = 29$	$n_2 = 24$	$m_2 = 13$	$n_3 = 12$	$m_3 = 7$
27.97	24.34	2	5	1	6
27.71	16.71	6	1	6	6
11.37	13.55	2	2	6	5
11.64	24.4	0	4	4	5
23.29	10.46	5	6	1	1
11.56	24.96	5	3	6	2
23.27	29.43	4	2	5	
12.14	21.19	5	6	3	
27.02	10.05	0	1	6	
13.19	17.49	3	4	2	
26.14	15.55	4	3	5	
15.10	28.85	3	4		
10.44	23.89	1			
16.89	13.82	6			
15.21	26.24	2			
27.83	21.86	3			
28.9 0	29.48	2			
12.33	20.70	3			
24.95	25.91	5			
14.05	28.15	5			
15.97	9.58	4			
23.65	19.46	6			
9.53	27.38	1			
29.91	28.66				
26.87	18.33				
20.07	19.43				
28.27	25.50				
11.59	26.99				
24.16					
15.57					

Благодарности

Выражаю благодарность проф. М.В. Сажину и проф. В.Е. Жарову за полезные обсуждения в процессе работы.

Таблица 25: Результаты анализа однородных выборок

Непараметрический критерий сдвига	$n_1 = 31$, $m_1 = 29$	$n_2 = 24$, $m_2 = 13$	Критическое значение	$n_3 = 12$, $m_3 = 7$	Критическое значение
Быстрый ранговый	1.04	0.03	1.96	0.08	1.96
Ван-дер-Вардена	0.97	0.06	1.96	0.02	3.62
Манна-Уитни-Вилкоксона	1.04	0.67	1.96	31	18 <...< 63
аппроксимация Имана	1.04*	0.67	2.00 (1.98*)	-	-

Список литературы

[1] Б.М. Щиголев «Математическая обработка наблюдений» (1969).

[2] В.Б. Монсик, А.А. Скрынников «Вероятность и статистика» (2011).

[3] Е.С. Кочетков, А.В. Осокин «Случайные события» (2000)

[4] L. Wasserman «All of Statistics. A Concise Course in Statistical Inference» (2004)

[5] Д. Худсон «Статистика для физиков» (1970).

[6] Т.А. Агекян «Основы теории ошибок для астрономов и физиков» (1972).

[7] А.И. Кобзарь «Прикладная математическая статистика» (2006 и последующие переиздания).

[8] Б.Л. ван дер Варден «Математическая статистика» (1960).

[9] В.А. Ильин, Э.Г.Позняк «Линейная алгебра» (1999).

[10] Дж. Деммель «Вычислительная линейная алгебра. Теория и приложения» (2001).

[11] Б.П. Демидович, И.А. Марон, Э.З. Шувалов «Численные методы анализа. Приближение функций, дифференциальные и интегральные уравнения» (1967).

[12] «Сборник задач по математике для ВТУЗов» под. ред. А.В. Ефимова и А.С. Поспелова (2003).